INFORMAÇÃO DIGITAL E SUAS DIVERSAS ABORDAGENS PELA ÓTICA DE UM CIENTISTA DA INFORMAÇÃO

LUANA FARIAS SALES
CARLA MARIA MARTELLOTE VIOLA
ORGANIZADORAS

ibict

Informação digital e suas diversas abordagens pela ótica de um cientista da informação

Esta publicação tem a cooperação da UNESCO no âmbito do projeto "Ampliação e Modernização das Ações do IBICT relacionadas às Atividades de Coleta, Armazenamento, Sistematização, Análise, Disseminação e Preservação de Dados e Informações Relativos à Ciência, Tecnologia e Inovação" (Prodoc 914BRZ2005). As indicações de nomes e a apresentação do material ao longo deste livro não implicam a manifestação de qualquer opinião por parte da UNESCO a respeito da condição jurídica de qualquer país, território, cidade, região ou de suas autoridades, tampouco da delimitação de suas fronteiras ou limites. As ideias e opiniões expressas nesta publicação são as dos autores e não refletem obrigatoriamente as da UNESCO nem comprometem a Organização.

COLEÇÃO PPGCI
50 ANOS

Informação digital e suas diversas abordagens pela ótica de um cientista da informação

Luana Farias Sales
Carla Maria Martellote Viola
organizadoras

ibict
Instituto Brasileiro de Informação em Ciência e Tecnologia

Rio de Janeiro
2021

Capa: Fernanda Estevam
Ilustração: GK Vector (br.freepik.com)
Projeto Gráfico: Paulo César Castro
Normalização e catalogação: Selo Nyota
Diagramação: Fernanda Estevam

Ficha Catalográfica: Priscila Fevrier - CRB 7-6678

I43

Informação digital e suas diversas abordagens pela ótica de um cientista da informação/ Luana Farias Sales; Carla Maria Martellote Viola (orgs.). – Rio de Janeiro: Ibict, 2021.

346p. -- (Coleção PPGCI 50 anos)

Inclui Bibliografia.
Disponível em: https://ridi.ibict.br/
ISBN 978-65-89167-13-6 (digital)
DOI: 10.22477/9786589167136

1. Ciência da informação. 2. Sistemas de informação. 3. Gestão de dados de pesquisa. 4. Preservação digital. I. Sales, Luana Farias. II. Viola, Carla Maria Martellote. III. Título.

CDD 020

Projeto editorial em colaboração com o Programa de Educação Tutorial (PET) da Escola de Comunicação (ECO-UFRJ): Paulo César Castro (tutor) / aluno(a)s: Carolina Torres, Dandara Campello, João Maurício Maturana, Juliana Sorrenti, Kethury Santos, Lianne Henriques, Mariana da Paz, Ludmila Rancan, Moniqui Frazão, Robertha Braga, Sabrina Oliveira e Sara Maluf.

Programa de Pós-Graduação em Ciência da Informação (PPGCI), desenvolvido pelo Instituto Brasileiro de Informação em Ciência e Tecnologia, do Ministério da Ciência e Tecnologia e Inovação (IBICT/MCTI) em convênio com a Escola de Comunicação da Universidade Federal do Rio de Janeiro (ECO/UFRJ).

Rua Lauro Muller, 455 - 4º andar
Botafogo - Rio de Janeiro - RJ
http://www.ppgci.ufrj.br

A pesquisa que resulta nesta publicação obteve o fomento de
CNPq
Capes

& com o apoio de
Unesco
IBICT
UNIRIO
UFRJ
Grupo de BRIET

Para os atuais e futuros pesquisadores em Ciência da Informação, em especial os discentes do PPGCI IBICT/UFRJ

"The basis of science is the empirical method, which uses the senses to build up a picture of the world; but science tells us that our senses have evolved to help us get by, not to show us the world as it is. Science is only a systematic examination of our impressions, and in the end all each of us has left are our own sensations. […] The end-result of the empirical method, then, is that each individual is left alone with their own experiences. We can escapethis solitude, Balfour suggested, only if we accept that there is a divine mind"

(GRAY, Cross-correspondences in The Immortalization Commission: The Strange Quest to Cheat Death, 2011, p. 69-70).

Sumário

Luis Fernando Sayão: uma travessia produtiva e vitoriosa da Física à Ciência da Informação

Lena Vania Ribeiro Pinheiro[1]

Luis Fernando Sayão, assim como muitos pesquisadores da Ciência da Informação no mundo, é formado em Física, mas tem dedicado sua vida acadêmica e profissional a este novo e fascinante campo do conhecimento. Aliás, é próprio da área a presença de profissionais das mais diversas formações, embora predominem os voltados à documentação, informação e questões correlatas. Alguns físicos marcaram a História da Ciência da Informação, especialmente os que também tinham doutorado em História da Ciência, tal como Derek de Solla Price, além de John Michael Ziman e Jack Meadows, astrônomo. e matemático, entre outros importantes cientistas e autores de livros clássicos em comunicação científica.

Luis Fernando Sayão é tecnologista concursado, desde 1980, na Comissão Nacional de Energia Nuclear (CNEN), lotado no Centro de Informações Nucleares (CIN), onde começou como bolsista, em uma atividade da Ciência da Informação, isto é, indexação de obras da Física. Assim sendo, o seu início foi nos denominados atualmente sistemas de organização do conhecimento, ou melhor, na representação da informação para sua recuperação,

Portanto, os caminhos de Luis Fernando Sayão seguiram o traçado dos primeiros documentalistas, aqueles que foram trabalhar em centros de informação ou bibliotecas especializadas, surgidos do desenvolvimento das ciências e do seu desdobramento em novas áreas, o que exigia conhecimento especializado, daí a chegada dos chamados documentalistas. Esta circunstância faz de Sayão um pesquisador que mescla a tradição, como a dos documentalistas, ao que podemos chamar modernidade ou contemporaneidade, por mergulhar em temas contemporâneos e ainda pouco pesquisados, por exemplo, ciência invisível, cauda longa da ciência, resultados negativos da ciência. No exercício do ensino e pesquisa em Ciência da

1 IBICT.

Informação, Luis Fernando Sayão se faz presente no IBICT na qualidade de egresso da pós-graduação, professor e pesquisador, em longa e rica trajetória.

E, no amplo território epistêmico da Ciência da Informação, quais são os temas ou questões motivadoras? Na sua declaração, no resumo do currículo Lattes, Luis Fernando Sayão aponta as suas áreas de interesse: bibliotecas/repositórios digitais, publicações eletrônicas, interoperabilidade de sistemas de informação para apoio à pesquisa, gestão de dados de pesquisa e preservação digital, sobre as quais tem incidido sua produção maior, marcada pelas tecnologias da informação, no âmbito da Ciência Aberta.

O primeiro passo, na verdade salto para Ciência da Informação, foi a saída do mestrado no Centro Brasileiro de Pesquisas Físicas (CBPF) para o de Ciência da Informação, no convênio IBICT/UFRJ. Consolidou a sua decisão anos depois, como primeiro doutor em Ciência da Informação, IBICT/UFRJ, no ano de 1994. Em outras ocasiões exerceu à docência como professor convidado e atualmente é professor credenciado no Programa de Pós-Graduação em Ciência da Informação - PPGCI, do IBICT e UFRJ. Essas passagens tão frequentes, dão a impressão de permanência, sem interrupções. Na verdade, não foi somente a frequência dessas atividades, mas também as fortes relações interinstitucionais entre IBICT e CIN, no desenvolvimento de projetos, sistemas e serviços de informação, além do número de pesquisadores do CIN que se titularam mestres e doutores pela pós-graduação no Programa IBICT/UFRJ.

Esta coletânea reúne as suas publicações mais citadas, que correspondem a quase totalidade de suas áreas de interesse, em um período de 20 anos, a mais antiga de 1996, e a mais recente, do ano de 2016, em um total de 13 documentos, sendo 12 artigos e um texto inédito. Estes documentos, por sua vez, estão distribuídos na coletânea em quatro (4) blocos., assim denominados pelos organizadores: Bloco 1: Método de pesquisa, um artigo; Bloco 2: Sistemas e bibliotecas digitais, quatro artigos; Bloco 3: Dados digitais, três artigos; e Bloco 4: Preservação, três artigos. Os artigos foram publicados nas seguintes revistas: Ciência da Informação (4 artigos); Ponto de Acesso (3); Revista USP (1); RECIIS (1); Informação & Sociedade: estudos (1); Encontros Bibli (1); texto em Seminário, São Paulo (1) e uma pesquisa inédita, que encerra a coletânea.

Quanto a autorias únicas e coletivas, predominam seus artigos como único autor, no total de oito (8), mantendo a tendência nas Ciências Sociais e Humanas embora, de um modo geral, em anos mais recentes a área venha apresentando número maior de artigos de autoria coletiva, mas não ainda superior à autoria única, conforme pesquisa no Brasil. Em seguida, aparecem cinco (5) artigos de dupla autoria, sendo três como autor principal e dois em colaboração, nos quais é o autor secundário.

Além de sua produção mais citada incluída nesta coletânea, Luis Fernando Sayão percorre caminhos interdisciplinares, especialmente em Ciência da Computação, Arquivologia e Museologia. Sua atuação nessas duas últimas áreas ocorre não somente em publicações, mas como integrante de comissões para estudos e resoluções relacionadas à aplicação *das tecnologias, normas* e padrões.

Ações interdisciplinares Implicam em formação ou espírito humanista, o que ocorreu fora da educação formal, como o amor pelas artes plásticas, em especial a pintura, que Luis Fernando Sayão exercita como prazer pessoal, mas ´com reflexos nas ilustrações de suas pesquisas, em gráficos coloridos e esteticamente bonitos. Um bom exemplo são suas delicadas ilustrações para visita guiada na Biblioteca Virtual Anísio Teixeira.

Luis Fernando Sayão completa a sua travessia não somente entre a Física e a Ciência da Informação, mas rompendo fronteiras entre culturas, em movimento essencial no mundo de hoje.

Rio de Janeiro, 31 agosto de 2021.

Prefácio

Luana Farias Sales e Carla Maria Martellote Viola

Desde o seu reconhecimento, em 1948, o campo da ciência da informação tem se mostrado instrumental na análise de problemas relativos à compressão, armazenamento e transmissão de dados e informações. Permitir a análise dos limites fundamentais da comunicação e compressão de dados lançou luz sobre o projeto prático de sistemas de comunicação por décadas. Os últimos anos testemunharam um renascimento no uso de métodos teóricos da informação para resolver problemas além da compressão de informações, comunicações de informações e rede, como as novas formas de cooperação entre sistemas de informação, integração e interoperabilidade no acesso a recursos informacionais eletrônicos, acesso aberto a dados de pesquisa, repositórios de dados de pesquisa, curadoria digital, gestão de metadados, reuso de dados, preservação digital, detecção, análise, compressão e aquisição de dados.

Nesse contexto, apresentamos esta obra que compreende, de acordo com consulta ao Google Acadêmico em 30 de março de 2021, uma coletânea dos artigos mais citados de acordo com Google Acadêmico[1], do professor Dr. Luís Fernando Sayão, escritos durante sua trajetória pela Ciência da Informação e por suas diversas passagens pelo PPGCI-IBICT/UFRJ, como aluno, consultor, pesquisador e professor.

O livro está dividido em 4 blocos, o primeiro bloco composto pelo capítulo 1, o segundo bloco, pelos capítulos 2 ao 6, o terceiro pelos capítulos 7 ao 9 e o último bloco, pelos capítulos 10 ao 13, totalizando 13 capítulos, 12 deles escritos entre 1996 e 2016, que versam sobre as questões relacionadas à Tecnologia da Informação aplicada à Ciência da Informação, bem como, aos aspectos ligados à informação digital. Somado a esses doze capítulos, o professor Sayão nos brinda ainda com um artigo inédito no final do livro.

Seguindo o método de classificação por categorias, organizamos os capítulos desta obra em quatro blocos: Método de Pesquisa (Capítulo 1), Sistemas, Bibliote-

1 Google Acadêmico. Luis Fernando Sayão. Disponível em: https://scholar.google.com.br/citations?user=nKCGcxwAAAAJ&hl=pt-BR&oi=sra

cas Digitais e Interoperabilidade (Capítulos 2–6), Preservação de Objetos Digitais (Capítulos 7-9) e Dados na Pesquisa Científica (Capítulos 10-13).

O Capítulo 1 analisa a importância dos modelos teóricos em Ciência da Informação enquanto recursos metodológicos e instrumentos de abstração e examina, na literatura sobre modelos e modelagem, a natureza, características básicas, funções e principais tipos de modelos. Este capítulo é citado em diversas áreas do conhecimento, sendo usado para embasar a adoção de modelos de abstração na condução metodológica de diferentes pesquisas.

Nos Capítulos 2 ao 6, reunimos os artigos que tratam de sistemas de informação e bibliotecas digitais. O Capítulo 2 discute os novos mecanismos de cooperação entre sistemas de informação em ciência e tecnologia surgidos a partir da emergência do arquivamento de publicações digitais em acesso aberto e do surgimento dos protocolos para interoperabilidade entre sistemas.

Neste mesmo viés, o Capítulo 3 descreve as opções tecnológicas e metodológicas para propiciar acesso e interoperabilidade de recursos informacionais eletrônicos, disponíveis na Internet, no âmbito do projeto da Biblioteca Digital Brasileira em Ciência e Tecnologia, desenvolvido pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT).

Em uma perspectiva mais conceitual, o Capítulo 4 aborda o surgimento das bibliotecas digitais num contexto que sobrepõe, por um lado, a integração e o uso das tecnologias de informação e de comunicação, das redes de computadores e, por outro, a disponibilidade crescente de conteúdos digitais em escala planetária e a possibilidade de digitalização, a um custo economicamente viável, de conteúdos em mídias convencionais. A partir dessa abordagem, o autor discute o papel das bibliotecas digitais, seu conceito e sua adoção como um novo serviço a ser oferecido pelas bibliotecas tradicionais.

Continuando a abordagem histórico-conceitual do Capítulo 4, o Capítulo 5 segue em uma narrativa epistemológica que traça a trajetória das bibliotecas digitais, desde quando elas ainda se configuravam como uma utopia, até o seu surgimento como algo real, com todos os seus desafios e perspectivas em diversos segmentos da sociedade.

O Capítulo 6 se vale das abordagens dos capítulos 2 e 3, isto é, interoperabilidade de sistemas e concomitantemente do objeto conceitual descrito nos capítulos 4 e 5, ou seja: as bibliotecas digitais. Com essa reunião, o autor apresenta, juntamente com o seu colega Prof. Carlos Henrique Marcondes, os principais problemas de interoperabilidade ocorridos no âmbito das bibliotecas digitais, bem como as soluções possíveis.

O terceiro bloco reúne diversos aspectos da preservação de objetos digitais. Inicialmente, o Capítulo 7 apresenta a importância dos repositórios digitais confiá-

veis para a preservação de periódicos eletrônicos científicos e a implementação da ideia dos arquivos digitais confiáveis pelas comunidades envolvidas.

Em seguida, o Capítulo 8 revela os principais conceitos, padrões e tecnologias envolvidos no desenvolvimento de esquemas de metadados de preservação e apresenta o dicionário de dados PREMIS[2], que teve como base de desenvolvimento a infraestrutura conceitual definida pela norma OAIS. Interessante nesse capítulo, é ver uma abordagem inédita para os metadados, que até então eram vistos, especialmente no âmbito da Biblioteconomia brasileira, apenas como elementos descritivos e pontos de acesso na catalogação de publicações.

O Capítulo 9, por sua vez, aborda a preservação para além das publicações, colocando os objetos digitais que compõem os acervos culturais digitais como elemento central na discussão, que é direcionada para a identificação das possibilidades de reuso, que expandam potencialidades informacionais e comunicacionais em razão do aumento das demandas por serviços on-line oferecidos por instituições de patrimônio cultural.

No último bloco reunimos os capítulos 10,11,12 e 13, que revelam algumas visões do autor sobre a temática "dados na pesquisa científica". O Capítulo 10 traça um paralelo entre as formas de incorporação de conhecimento nas bases de dados internacionais e o conceito de memória coletiva dentro do âmbito da produção científica mundial para analisar as barreiras de acesso à ciência impostas pelas linguagens de indexação. Assim, apresenta as bases de dados como um repositório de conhecimento da ciência moderna.

O Capítulo 11, no âmbito da perspectiva da ciência aberta, apresenta a importância da abertura dos dados científicos, os impactos nos atuais sistemas de informação para a pesquisa e propõe elementos para a composição de um modelo de gestão de dados de pesquisa para o país.

No Capítulo 12, o autor analisa brevemente a importância dos dados de pesquisa, a necessidade de preservação desses dados, bem como a noção de curadoria digital, que se desenvolve no âmbito da gestão de dados de pesquisa, indo desde o seu planejamento até seus impactos na formulação de novos documentos e na comunicação científica.

Por fim, no último capítulo, o professor apresenta uma nova abordagem para a Gestão de Dados de Pesquisa, focando a gestão como um serviço de informação novo e necessário.

2 O Dicionário PREMIS é um padrão internacional de metadados, desenvolvido por uma equipe internacional de especialistas, para apoiar a preservação de objetos digitais e garantir sua usabilidade em longo prazo.

O Objetivo desta obra foi reunir os diversos textos utilizados, especialmente em cursos e disciplinas ministradas no PPGCI-IBICT/UFRJ, facilitando sua localização e acesso.

Esperamos sinceramente que a expressão das ideias e percepções do professor Luís Fernando Sayão, como autor singular, ou em parceria, forneça estímulo e inspiração, e que seja usada como base para pesquisas futuras, não apenas no nosso Programa de Pós-Graduação, mas também em outros programas de pós-graduação.

PARTE 1

A CIÊNCIA DA INFORMAÇÃO COMO MÉTODO DE PESQUISA

Modelos teóricos em ciência da informação: abstração e método científico

Luís Fernando Sayão

1 Introdução

Na busca de novos esclarecimentos e conhecimentos, de novos fenômenos e eventos, o ser humano não os identifica somente pelas sensações ou pelas manifestações imediatas, mas recorre à reflexão e ao conhecimento acumulado, através da formulação de hipóteses e da estruturação de modelos (almeida, 1981). Dessa forma, a abstração constitui uma ferramenta poderosa no exercício eterno de aquisição de conhecimento, uma vez que, para se compreender a imensa variedades de formas, estruturas, comportamentos e fenômenos residentes no nosso universo, é necessário selecionar aqueles de maior relevância para o problema objeto de investigação e elaborar para eles descrições adequadas. Constroem-se, assim, esquemas abstratos da realidade, nos quais as coisas são reduzidas a seus perfis mais convenientes. O conhecimento racional é, dessa forma, um sistema de símbolos e conceitos abstratos, caracterizado pela estrutura seqüencial e linear tão típica de nosso pensamento e de nossa fala (capra, 1983).

Nesse sentido, a evolução da humanidade no seu aspecto mais abrangente - a evolução das ciências, artes, filosofia, tecnologia - pode ser encarada como uma trajetória rumo à aquisição progressiva da capacidade individual de abstração. De um ser intimamente ligado à natureza, ao mundo real, concreto e objetivo, o homem tornou-se ao longo do tempo um ente independente, isolado e com cada vez maior capacidade de introspecção (setzer, 1986).

Os cientistas, hoje em dia, apercebem-se do fato de que todas as suas teorias são criações da mente humana; são propriedades do nosso mapa conceitual da realidade, e não pertencentes ao domínio da realidade. Esse esquema conceitual é necessariamente limitado e aproximado como, de resto, o são todas as teorias científicas.

Segundo Capra (1983), "o que torna a ciência tão bem-sucedida é a descoberta de que podemos utilizar aproximações. Se nos satisfizermos com uma 'compreensão'

aproximada da natureza podemos descrever grupos selecionados de fenômenos, negligenciando outros que se mostrem menos relevantes. Assim podemos explicar muitos fenômenos em termos de poucos e, conseqüentemente, compreender diferentes aspectos da natureza de forma aproximada, sem precisar entender tudo ao mesmo tempo. Esse é o método científico: todas as teorias e modelos científicos são aproximações da verdadeira natureza das coisas; o erro envolvido na aproximação é, não raro, suficientemente pequeno para tornar significativa essa aproximação".

Dessa forma, um modelo é uma criação cultural, um "mentefato", destinada a representar uma realidade, ou alguns dos seus aspectos, a fim de torná-los descritíveis qualitativa e quantitativamente e, algumas vezes, observáveis. A existência de modelos jaz na impossibilidade cultural de descrever os objetos com perfeição, esgotando as possibilidades de sua observação. Não sendo transparente para o homem, o mundo se lhe apresenta como um permanente desafio à sua descrição. Essa limitação filosófica de percepção é que permite e exige o aparecimento de modelos (almeida, 1981).

Dentre os vários aspectos, os modelos apresentam uma analogia, sempre que possível, mas nem sempre desejável, com o objeto real. Por analogia entende-se a representação de uma mesma função em diversos materiais e por meio de princípios diversos. Ela pode ser construída por meio de formalismos matemático, fenomenológico ou conceitual. É mais simplificada, permite testar hipóteses, tirar conclusões, caminhar no sentido da generalização e da particularização, através de processos de indução, e tem sempre uma vida provisória.

Cada modelo expressa e justifica um método de abordagem de uma realidade física, ao mesmo tempo em que cada método subentende um modelo, nem que seja um modelo meramente operacional.

Os modelos apresentam também uma dimensão heurística, na medida em que, criado para explicar e fazer compreender alguns aspectos de uma realidade, são factível de evolução e de assegurar a percepção de outros aspectos não imaginados antes de sua elaboração.

Por outro lado, uma mesma realidade física pode possuir mais de um e diferentes modelos, como acontece com o núcleo atômico. Cada modelo se destina a explicar faixas características de fenômenos nucleares, podendo, além de suas limitações, chegar a explicações complementares ou contraditórias com outros modelos.

2 A natureza dos modelos

Os modelos, em uma generalização arriscada, buscam a formalização do universo através de meios de expressões controláveis pelo ser humano; derivam da necessidade humana de entender a realidade aparentemente complexa do universo

envolvente. São, portanto, representações simplificadas e inteligíveis do mundo, que permitem vislumbrar características essenciais de um domínio ou campo de estudo. A necessidade de idealização é, portanto, uma reação tradicional do homem à aparente complexidade da realidade em que está submerso. A mente tenta decompor o mundo real em uma série de sistemas simplificados e atingir assim em um único ato "uma visão das características essenciais de um domínio" (APOSTEL, 1991). Esta simplificação exige criatividade, tanto sensorial quanto intelectual, o que, evidentemente, implica admitir-se que, na construção de modelos, algumas características da realidade, que não se referem diretamente aos objetivos buscados, são desprezadas ou abandonadas, em função da maior inteligibilidade ou facilidade de compreensão (CHORLEY; HAGGETT, 1975).

Enquanto representação de algum aspecto da realidade, um modelo assume a natureza ambígua de ser igual e desigual à realidade que ele modela. Ele possui a sua própria forma e estrutura, independentemente do original que representa; as afinidades e divergências entre o modelo e a realidade devem ser "expressáveis" e expressadas. Dessa forma, um modelo também exige um modo de expressão que pode ser, só para exemplificar alguns, gráfico, procedural, discursivo (BURT; KINNUCAN, 1990).

Um modelo serve a muitos propósitos, mas serve fundamentalmente para comunicar alguma coisa sobre o objeto da modelagem de forma a gerar um entendimento mais completo sobre a realidade; a ação de modelar, por sua vez, impõe a quem modela uma visão clara e sem ambigüidades de quem ou do que está sendo modelado, além de exigir uma correta seleção dos elementos do universo do discurso que comporão a visão a ser representada.

Como observam Haggett & Chorley (1975), "um modelo é uma estruturação simplificada da realidade que apresenta supostamente características ou relações sob forma generalizada. Os modelos são aproximações altamente subjetivas, no sentido de não incluírem todas as observações e mensurações e medições associadas, mas, como tais, são valiosas por ocultarem detalhes secundários e permitirem o aparecimento dos aspectos fundamentais da realidade. Esta seletividade significa que os modelos têm graus variáveis de probabilidade de aplicação e um alcance limitado de condições sobre as quais se aplicam. Os modelos de maior sucesso possuem alta probabilidade de aplicação e extensa gama de condições sob as quais aparecem apropriados. Com efeito, o valor de um modelo é muitas vezes diretamente relacionado ao seu nível de abstração. Capra observa que todas as "leis da natureza" que os modelos estabelecem são transitórias e destinadas a serem substituídas por leis mais precisas à medida que os modelos são aperfeiçoados. Esse estado provisório é atestado pelas "constantes fundamentais" (por exemplo, a velocidade da luz),

ou seja, "quantidades cujos valores numéricos não são explicados pela teoria, mas que nela têm de ser inseridos após terem sido determinados empiricamente". inda usando os problemas de modelagem da física, verifica-se que as teorias quântica, de campos e da relatividade não podem explicar algumas grandezas que são consideradas na visão clássica constantes fundamentais da natureza. Na visão moderna, seu papel de "constantes fundamentais" é tido como algo provisório e que reflete a limitação das teorias de que dispomos (capra, 1983). Embora só recentemente os físicos tenham adquirido um razoável conhecimento sobre as forças nucleares, não foi ainda possível usar esse conhecimento para construir uma teoria nuclear ampla. Existe, entretanto, um grande número de modelos ou de teorias rudimentares de validade restrita, cada um deles explicando uma pequeno espectro das propriedades nucleares (masterman, 1970).

Para Skilling (1964), os modelos podem ser hipóteses, hip teses não testadas ou insuficientemente testadas, teorias, sínteses de dados, funções, relações ou equações. Podem ser, até, idéias estruturadas, conectando argumentos que apresentam algum poder explanatório. São, assim, estruturações que representam a realidade, apresentando supostas características ou relações de forma generalizada.

3 Características dos modelos

Herbert Stachowiak (1972) apresenta três características básicas dos modelos:

a) característica de mapeamento - modelos sempre modelam alguma coisa, ou seja, são representações de "originais" (ou "protótipos"), naturais ou artificiais, que, por sua vez, também podem ser modelados.
b) característica de redução - modelos geralmente não mapeam todos os atributos do original que eles representam, mas unicamente aqueles que são relevantes para quem modela.
c) característica de pragmatismo - modelos não são em si pertencentes à mesma classe que seus originais. Eles sempre cumprem suas funções de substituição orientados unicamente para objetivos dependentes de operações mentais ou factuais, dentro de uma faixa limitada de tempo.

De acordo com Hagget & Chorley (1975), a característica mais importante dos modelos é que sua construção implica uma atitude altamente seletiva em relação às informações, na qual não s as interferências como os sinais menos importantes são eliminados para permitir que se observe algo da intimidade das coisas. Desta forma, os modelos podem ser considerados como aproximações seletivas que, pela eliminação de detalhes acidentais, permitem o aparecimento de alguns aspectos

fundamentais relevantes ou interessantes do mundo real sob alguma forma generalizada. possibilidade de ser inexato e desigual em relação ao seu original é que, em última análise, permite ao modelo revelar o que se deseja.

Outra característica importante dos modelos é que eles são estruturados, no sentido de que os aspectos importantes selecionados da realidade são explorados em termos de suas relações com outros modelos e aspectos da realidade; seguem as características gerais das estruturas conforme enunciados por Piaget(1972), estabelecem que as estruturas constituem-se uma totalidade, com leis próprias independentes das características particulares dos seus elementos e que consistem de um sistema de operações de transformação cujo conjunto de combinações internas nunca geram produtos fora da estrutura. "ciência tirou grande proveito desta busca de padrões, na qual os fenômenos são considerados em termos de uma espécie de relação orgânica" (CHORLEY; HAGGETT, 1975); isto acontece principalmente quando se pensa em termos dos referenciais estabelecidos por Von Bertalanffy (1962), através da teoria geral dos sistemas, que propunha visualizar o mundo e o universo em termo de um grande conjunto interconectado, dentro do qual se poderia separar subsistemas para análise.

Esta característica dos modelos implica imediatamente a sua natureza sugestiva, no sentido de que um bom modelo traz, em si, na sua própria estrutura, sugestões para a sua própria extensão e generalização. Isto significa, primeiramente, que toda a estrutura do modelo tem maiores implicações do que um estudo de suas partes individuais e, segundo, que pelo modelo, por meio de operações e transformações proporcionadas por suas leis estruturais, podem ser feitas previsões do mundo real. Dessa forma, os modelos são instrumentos especulativos cujas implicações mais positivas conduzem a hip teses e especulações novas no campo primário da investigação (BLACK, 1962; HESSE, 1954). Ainda em relação à natureza estrutural dos modelos, é interessante notar que, segundo Kaplan (1964), o que é denominado "modelo" pelos lógicos é chamado de "estrutura" pelos economistas.

Como bem resumem Mendonça de Souza & Dodebei (1993), "por serem os modelos diferentes do mundo real, são então analogias que permitem reformular o conhecimento sobre alguns aspectos do mundo real em uma forma mais familiar, simplificada e acessível, observável e facilmente formulada ou controlável, da qual se pode tirar conclusões que, por sua vez, possam ser aplicadas no mundo real. reaplicação é um pré-requisito dos modelos nas ciências empíricas".

4 Funções dos modelos

De acordo com Apostel (1991), "os modelos são necessários por constituírem uma ponte entre os níveis da observação e o te rico e tratam da simplificação,

redução, concretização, experimentação, ação, extensão, globalização, explicação e formação da teoria". Dentro dessa perspectiva, uma das suas funções principais é a explanatória e redutora de complexidade, no sentido em que permite que uma determinada classe de fenômenos possa ser visualizada e compreendida, o que de outra forma não seria possível devido à sua magnitude e complexidade.

Chorley & Haggett (1975) consideram ainda a função aquisitiva, que diz respeito à estrutura proporcionada pelos modelos, através da qual a informação pode ser definida, coletada e ordenada. Além dessa função organizacional, considera-se uma função que permite a otimização da extração de informações a partir do modelo - a fertilidade. Os modelos também desempenham uma função lógica que ajuda a explicar como ocorre determinado fenômeno; alinha-se também a função normativa que permite a comparação de fenômenos com outros mais familiares, além da função sistemática da construção de modelos, segundo a qual a realidade é vista em termos de sistemas interligados. Esta função conduz a uma outra, a função construtiva dos modelos que acentuam o papel destes na construção de teorias e leis. "Finalmente, há a função de parentesco dos modelos, promovendo a comunicação das ideias científicas" (CHORLEY; HAGGETT, 1975). Segundo Kaplan (1964), esta comunicação "não é uma questão meramente de sociologia da ciência, mas intrínseca à sua lógica; como na arte, a ideia não representa nada até que tenhamos encontrado a expressão».

5 Tipos de modelos

Como observam Mendonça de Sousa & Dodebei (1993), os modelos são tipologizados de várias formas, em função das próprias ideologias inerentes a cada autor, área de conhecimento ou ainda segundo objetivos específicos; isto significa que "considerando-se forma e expressão, os modelos podem ser agrupados ou classificados em uma série interminável de tipos". Entretanto, o termo "modelo" tem sido usado em uma variedade tão ampla de contextos que é difícil definir, sem ambiguidades, até mesmo os tipos mais gerais.

Haggett & Chorley (1975) propõem uma classificação que pode ser interessante para os objetivos desse trabalho. Segundo esses autores, os modelos podem ser descritivos e normativos. O primeiro grupo trata de certa descrição estilística da realidade, e o segundo, do que se pode esperar que ocorram sob certas condições estabelecidas. Os modelos descritivos podem ser predominantemente estáticos - concentrando-se nos aspectos de equilíbrio estrutural - ou dinâmicos, concentrando-se, neste caso, nos processos e funções através do tempo. Quando o elemento tempo é particularmente salientado, resultam os modelos históricos ou temporais. Os modelos descritivos podem tratar da organização das informações empíricas e

assim serem denominados modelos de dados, classificatórios (taxionômicos) ou de fim experimental.

Ainda segundo Haggett & Chorley (1975), os modelos também podem ser classificados segundo a natureza de sua constituição. Em uma primeira visão, podem ser visualizados com construções s lidas, físicas ou experimentais e, em segundo lugar, como modelos te ricos, simbólicos, conceituais ou mentais. Nos primeiros, as propriedades importantes do mundo real podem ser representadas de duas formas: modelos icônicos - as propriedades do mundo real são representadas pelas mesmas propriedades com uma mudança apenas de escala; modelos análogos - as propriedades do mundo real são representadas por propriedades diferentes. Os modelos te ricos, simbólicos, conceituais ou mentais tratam de afirmações simbólicas ou formais de tipo verbal ou matemático; os modelos matemáticos podem ainda ser classificados, segundo o grau de probabilidade associada com sua forma de previsão, em determinísticos e estocásticos.

Por fim, temos os paradigmas. "Um modelo que se revela correto e útil em uma infinidade de aplicações, em circunstâncias distintas e sobre dados diferentes, que apresenta, ao mesmo tempo, um amplo poder explanatório, pode ser definido como um paradigma". Os paradigmas podem ser considerados como modelos estáveis da atividade científica, sendo, em certo sentido, modelos em escala ampla. Diferem, entretanto, destes no que diz respeito às suas fronteiras de validade. Os paradigmas raramente são formulados tão especificamente, trata-se de modelos de busca do mundo real. Neste sentido, os paradigmas podem ser entendidos como "supermodelos" dentro dos quais os modelos são colocados em escala mais reduzidas (CHORLEY; HAGGETT, 1975).

6 Modelos em sistemas de informação

Um modelo é antes de mais nada uma representação de um recorte da realidade, que, de acordo com a sua função utilitária e por meio do seu modo de expressão, sua estrutura e suas igualdades e desigualdades em relação ao seu original, tenta comunicar algo sobre o real. Nesse sentido, um modelo de informação é uma representação de um ser humano enquanto usuário e/ou parte de um sistema de informação e das suas relações de aquisição, organização e manipulação de informação.

Burt & Kinnucan (1990) referem-se à modelagem de informação como o exercício de identificação de componentes de modelos e seus elos, a explicitação dos modos de expressão, bem como o delineamento dos paradigmas e seus efeitos sobre os tipos de modelos que estão sendo construídos, pois os paradigmas refletem ao mesmo tempo os propósitos e as fronteiras de um modelo.

Em comunicação e ciência da informação, o modelo de maior sucesso e ampla utilização foi a teoria da comunicação dos matemáticos americanos Shannon & Weaver (1949), que propuseram um modelo matemático para explicar a comunicação entre dois pólos, denominados "emissor" e "receptor". "Tal modelo, criticado, adaptado, modificado, ainda hoje está sendo amplamente utilizado, na medida em que, de modo preciso, simples e preditivo, propicia uma boa ideia de como se dá a comunicação humana. É, em essência, um modelo matemático, da mesma forma que as leis de Zipf, Bradford, Ortega, 80/20 e outras amplamente utilizadas na bibliometria, mas é, também, na sua concepção geral, um modelo sistêmico interligando o emissor ao receptor" (souza; dodebei, 1993).

A ciência da informação, pela sua própria natureza ampla e interdisciplinar, para mapear toda a sua realidade, teve obrigatoriamente de tomar, como seus, paradigmas e modelos de outras áreas, tais como informática, inteligência artificial, lingüística, economia, marketing. Kuhn16 se refere às ciências cujos cientistas não são guiados por um único paradigma de "ciências preparadigmáticas"; nesse estado se encontram as ciências comportamentais, as sociais e a ciência da informação. Entretanto, Masterman (1970) caracteriza a ciência da informação como uma ciência multiparadigmática, dentro de uma escala em que se pode identificar as ciências "normais" e as de duplo paradigma, que são ciências normais em estado de crise, em que dois paradigmas estão em competição (ellis, 1992).

7 Tipos de modelos de informação

A área de modelos e modelagem de informação caracteriza-se mais por não possuir fronteiras claras dos seus domínios internos e externos, do que por possuir um corpo coerente e consistente de trabalhos. Não obstante, esta é uma área importante e, provavelmente, sua importância será ainda maior, considerando que as pessoas e as organizações têm exigências cada vez mais sofisticadas em relação aos sistemas de informação.

Ao considerarmos todo o domínio de possibilidades dos modelos de informação, teremos, com base na proposta de Burt & Kinnucan (1990), uma configuração contínua. Em um dos extremos desse contínuo está o ser humano com sua realidade pessoal, interna e presumidamente idiossincrática, ou seja, com o que chamamos conhecimento; no extremo oposto, está o sistema de informação com a sua realidade dependente dos seus próprios limites internos. Entre esses dois extremos, está localizado o campo de representações, o espaço onde se encontram representações de uma ou de outra realidade. Essas representações experimentam criar pontes ou elucidar algumas ou todas as estranhezas que se supõe existam entre essas realidades.

Os modelos que estão mais próximos do que seja a representação do usuário humano e do que se passa em sua cabeça em relação ao sistema são denominados modelos cognitivos; ao passo que os modelos que se identificam com o sistema e tentam descrever o que se passa em seu interior são mais conhecidos como modelos de dados. Na região intermediária do contínuo, é o lugar geométrico dos modelos que interpretam os usuários, o sistema e a interação entre eles; os modelos que se enquadram nesta categoria são coletivamente denominados de modelos conceituais. Dentro do escopo dos modelos cognitivos e conceituais, localiza-se um número significativo de subgrupos importantes. Não é possível, entretanto, estabelecer limites claros entre alguns desses modelos. Não obstante, Burt & Kinnucan (1990) enfatizam que deve ficar clara "a distinção que se faz entre a visão individual da realidade - isto é, modelo cognitivo - e a visão que alguma outra pessoa tem de como um grupo de indivíduos devem estar vendo alguns aspectos de um sistema de Informação - isto é, modelo conceitual".

Os autores que tentaram caracterizar esta área como um todo concordam que ela é ampla e fragmentada. Algumas regiões do contínuo, especialmente aquelas que possuem longa tradição de pesquisa, estão relativamente desenvolvidas, ao passo que outras estão praticamente intocadas. Vamos analisar alguns dos tipos de modelos mais importantes no contexto desse trabalho.

Modelos Cognitivos - Daniels, no seu artigo de revisão intitulado "*Cognitive models in information retrieval - an evaluative review*" (DANIELS, 1986), cujo propósito é sugerir como os modelos de usuário podem ser usados para otimizar o desempenho e a aceitabilidade dos sistemas de recuperação, lança um olhar crítico sobre os modelos segundo a perspectiva cognitiva e discute o seu papel na biblioteconomia e ciência da informação. Ela confirma que existe um vasto espectro de significados para os conceitos expressos por "modelo mental" ou "modelo cognitivo", mas, "de uma forma geral, modelos cognitivos podem ser considerados como imagens que os componentes de um sistema, sejam eles pessoas ou máquinas, têm de si próprios, de cada um dos outros componentes e da realidade". Isto se relaciona fortemente com o fato consensual de que, para haver comunicação entre duas partes, é necessário que cada parte tenha incorporado um modelo da outra, que muito certamente não corresponde ao modelo que cada um tem de si próprio. Allen (1991), nesta mesma linha, estabelece que os modelos cognitivos referem- se ao conhecimento que os usuários têm sobre um sistema de informação e, em alguns casos, ao conhecimento que os sistemas de informação têm sobre os usuários.

Com o crescimento da área de pesquisa conhecida como "ciências cognitivas", interligando disciplinas tais como psicologia, lingüística, inteligência artificial, filosofia, educação e ciência da informação, o interesse pelos modelos cognitivos no

âmbito da ciência da informação tem sido cada vez maior, pois os problemas de representação, informação, comunicação e conhecimento são fundamentais para todas essas disciplinas; além do mais, as pesquisas cognitivas em ciência da informação extraem os recursos metodológicos e os quadros conceituais dessas mesmas disciplinas, criando um corpo de conhecimento extraordinariamente multidisciplinar (GAINES, 1991; VICKERY, 1986). Allen (1991), na sua análise denominada "*Cognitive research in information science: implication for design*", indica-nos que o ponto de inflexão nas pesquisas cognitivas em ciência da informação foi o *International Workshop on the Cognitive Viewpoint*, que aconteceu em Ghent, 1977.

Desde que a matéria-prima dos processos cognitivos são "objetos mentais", tais como conceitos, idéias e conhecimento, muitos dos trabalhos nestas áreas consideram como as pessoas organizam conhecimento, como os conceitos são formados na mente humana, como as pessoas agrupam objetos em suas mentes, ou seja, como as pessoas os categorizam, quais são as teorias que as pessoas têm sobre como o mundo funciona. No contexto dos sistemas de informação, é bastante interessante para o presente trabalho os estudos de Humphrey (1985) sobre como os papéis assumidos por uma pessoa na sua organização influencia sua percepção e comportamento. Ele descobriu que a visão individual que cada pessoa tem sobre a sua posição na estrutura organizacional e seu fator motivacional de participação têm influência na sua avaliação e seleção da informação disponível.

Sobre a perspectiva cognitiva em biblioteconomia e ciência da informação, Daniels (1986) afirma que, "nos últimos anos, na área de recuperação de informação, se consolidou o consenso de que o sistema homem-computador deve ser visto como um sistema adaptativo cognitivo. perspectiva cognitiva implica que o processamento de informação é sempre intermediado por algum tipo de modelo da realidade" e, além do mais, um sistema cognitivo pode ser considerado um sistema adaptativo cujo funcionamento, planejamento e mudanças de ações estão baseados no conhecimento de si próprio e do seu contexto.

O elenco de pesquisas que adotam a perspectiva cognitiva em recuperação da informação incluem:

a) Representação de usuários e seus problemas - modelam situações problemáticas dos usuários em face dos sistemas de informação, tais como de indivíduos cujo modelo interno de conhecimento e contextualização não é suficiente para atingimento de seus objetivos; ou ainda o "estado anômalo do conhecimento" preconizado por Belkin28, que tenta estabelecer, diante da impossibilidade de o usuário identificar o de que ele precisa, quadros (framework) através dos quais os motivos da busca de informação por par-

te de um usuário possam ser bem explicitados e seus resultados usados na recuperação de informação.

b) Representação de estratégias de busca - examina, por exemplo, os aspectos cognitivos do processo de transferência de informação do usuário para o especialista em informação (intermediário). A interação entre usuário e intermediários em sistemas de informação consiste, em grande parte, na construção de modelos cognitivos apropriados das várias facetas do usuário.
c) Representação de documentos e informação.
d) Modelos conceituais – Borgman (1986), que escreveu sobre a interação homem-máquina segundo uma perspectiva psicológica, sugere que os modelos mentais, no contexto de interfaces, referem-se ao modelo do usuário segundo a perspectiva do sistema, enquanto modelos conceituais são aqueles que são apresentados ao usuário pelo projetista do sistema.

Conforme foi sugerido pelo relatório NSI/SPARC (1975), um sistema de informação pode ser visualizado em três níveis: interno, conceitual e externo. O nível conceitual concentra-se no "significado" (conceitos) da informação. A "tarefa de desenvolver um esquema conceitual é chamado de modelagem de informação. Seu objetivo primordial é desenvolver uma descrição estável e coerente do significado dos dados, ou seja, um esquema conceitual. Assim sendo, modelagem de informação difere da modelagem de dados, conforme desenvolvida na década de 70, que tratava principalmente da descrição de estrutura de dados (relacional, redes, hierárquicos) com vista ao acesso e armazenamento de dados", conforme analisa Lyytinen (1987). Nessa linha, ele propõe duas instâncias para modelagem de informação. A primeira delas chama-se "mapeamento da realidade" (*reality mapping*), que é essencialmente uma técnica descritiva para representar alguma coisa que é claramente compreendida e que apresenta um comportamento sem ambiguidades. Os enfoques nos quais é baseada essa visão são relativamente comum na literatura e, basicamente, supõem um processo de mapeamento do "mundo real" em modelos formais, ou seja, em esquemas conceituais. De acordo com esta visão, "um sistema de informação é um sistema formal completamente previsível que espelha o comportamento determinístico de um universo do discurso". A necessidade de certeza em todos os níveis nesse enfoque traz uma limitação severa à sua aplicabilidade. Lyytinen (1987) prefere o segundo paradigma, "desenvolvimento de linguagem formal" (*formal language development*), cujo enfoque está sobre a representação, estrutura conteúdo e uso da mensagem lingüística, uma vez que ela pode lidar com maior precisão com a natureza essencialmente ambígua da maioria das configurações da realidade.

Visão do usuário/modelagem de usuário - Esses modelos apresentam interpretações de um sistema de informação real ou te rico a partir de parâmetros extraídos ou postulados de um usuário ou de um grupo de usuários que possuem características específicas que o construtor do modelo julga serem relevantes para o uso do sistema de informação. Daniels (1986) agrupa esses modelos em duas grandes classes: "modelos quantitativos empíricos" e "modelos cognitivos analíticos". Os primeiros são formalizações abstratas de uma classe geral de usuários definidas em termos de parâmetros de projetos de interfaces para o usuário. Tais modelos são geralmente desenhados tomando como base pessoas médias no desempenho de várias atividades em várias ambientações. Em contraste com esses modelos, os modelos cognitivos analíticos "buscam modelar aspectos do comportamento cognitivo do usuário sob o enfoque qualitativo, o que inclui: o conhecimento do usuário, seus objetivos, planos, convicções, experiência, tipo de interação preferida etc".

De uma forma geral, modelos que incorporam visões do usuário tendem a considerar diferenças entre grupos específicos de usuários ou, de uma forma mais simples, de entre tipos de usuários. Muitos desses grupos são criados, pré-concebidamente, segundo a perspectiva de desempenho timo dos sistemas. A dicotomia mais comum é que enquadram os usuários em "iniciantes" e "*seniors*", onde a habilidade modelada é a experiência no uso do computador, e não o domínio ou conhecimento da aplicação (burt; kinnuncan, 1990; daniels, 1986; lyyttinen, 1975).

De acordo com Rich (1979), os modelos de usuários podem ser classificados segundo três dimensões principais:

a) modelo de um usuário simples, típico ou "canônico" *versus* uma coleção de modelos de usuários individuais;
b) modelos construídos pelo usuário ou especificados pelo projetista do sistema versus modelos pressuposto pelo computador com base no comportamento do usuário.
c) modelos de usuário de característica de longo prazo, tais como áreas de interesse e experiência *versus* modelos de curta validade. Spark Jones (daniels, 1986) sugere uma quarta dimensão, que em alguns casos pode se confundir com a terceira.
d) modelos dinâmicos, ou seja, modelos mutáveis dependendo do contexto, *versus* modelos estáticos, que representam as características permanentes do usuário.

Uma linha importante para os sistemas de informação diz respeito aos métodos de representação do conhecimento dos usuários e de que forma eles podem

ser agrupados segundo esse princípio. Hammond & Barnard (DANIELS, 1986), por exemplo, identificaram nove tipos de conhecimento necessários a uma interação usuário versus sistema de informação. Uma segunda linha considera a maneira como alguém estrutura seus conhecimentos. Durding (BURT; KINNUCAN, 1990), já há algum tempo, demonstrou que as pessoas usam estruturas distintas, tais como redes e estruturas hierárquicas, para organizar conceitos quando estas estruturas são próprias e inerentes aos itens que estão sendo processados. Diagramas e outros tipos de representação gráfica também são úteis na estruturação de conhecimento. Os usuários de bases de dados organizadas hierarquicamente fazem buscas mais eficientemente quando estão de posse de um "mapa" com a estrutura de árvore da base de dados. Provavelmente o diagrama possibilita ao usuário chance de conceber a sua própria representação hierárquica da base de dados (BURT; KINNUCAN, 1990; DANIELS, 1986; ALLEN, 1991).

Modelos Semânticos de Dados - Modelos de dados representam a área mais próxima do sistema e de sua realidade interna. O modelo hierárquico, o modelo de redes, bem como o modelo relacional de Codd (1979), são tradicionais representantes de modelos de dados; tipicamente essa categoria de modelos enfatiza os aspectos sintáticos e estruturais dos dados sem, entretanto, considerar o significado dos dados ou o relacionamento próprio e lógico entre eles. Com o advento de sistemas de informações mais sofisticados e de maior abrangência, tornou-se absolutamente necessário o desenvolvimento de modelos que viessem facilitar o entendimento do usuário em relação ao sistema e ao mesmo tempo evitassem o envolvimento dele com a estrutura física dos dados dentro do computador. As pesquisas nessa área se concentram predominantemente no desenvolvimento de modelos que espelhem com maior fidelidade a "complexidade semântica do mundo real da informação". Com efeito, nos últimos anos, os pesquisadores da área de banco de dados voltam suas inteligências no sentido de incorporar aspectos comportamentais (ou dinâmicos) de dados nos formalismos de modelagem; este trabalho tem sido fortemente influenciado pelo paradigma da programação orientada por objetos. Modelos que caminham nesta direção são denominados pela literatura de "modelos semânticos de dados" (HULL; KING, 1987; CHEN, 1976).

Tsicchritzis & Lochovsky (1982) consideram que o papel próprio dos modelos de dados é servir como um meio de comunicação dirigido às pessoas em geral, de forma que a estrutura imposta pelo modelo não poderia discordar da estrutura natural da realidade tal como ela é percebida pelo usuário humano. Dentro dessa perspectiva, eles discutem a proposta de dois tipos de modelos de dados: o primeiro deles lidaria com o mapeamento das informações do mundo real em conceitos básicos humanos - domínio infológico (*infological realm*); em uma segunda instân-

cia teríamos o domínio datalógico (*datalogical realm*), que mapearia estes conceitos básicos humanos em representações em computador. Neste caso, os modelos de dados tradicionais estariam categorizados como "datalógicos", ao passo que os modelos de dados semânticos seriam, ao menos, o primeiro passo em direção aos modelos "infológicos".

Os modelos conceituais de dados conhecidos foram criados como ferramentas de representação que funcionam em ambientes específicos, e, assim sendo, a literatura não reconhece nenhum modelo generalizado. Hull & King (1987) concluem que, apesar da dificuldade de definições precisas, a literatura aponta uma trajetória evolucionária para a área de modelos semânticos de dados. s pesquisas nesta área estão relacionadas principalmente na extensão do modelo relacional, de forma a enriquecê-lo com abstrações semânticas provenientes da pesquisa em lingüística.

As abstrações semânticas são formas de especificar relacionamentos entre conceitos lingüísticos que trabalham as diferenças sutis de significado. Burt & Kinnucan (1990) enunciam quatro dessas abstrações como as mais comumentes usadas nos modelos semânticos: generalização, agregação, classificação e associação.

a) Generalização - Esta abstração ocorre quando objetos ou entidades são agrupados em um relacionamento hierárquico no qual os objetos do nível mais baixo são vistos como subtipos daqueles de nível mais alto. Por exemplo, os objetos "biografias" e "novelas" podem ser visualizados como exemplares específicos do objeto "livro". Este tipo de agregação é conhecido como uma relação "é- um" (em inglês: "is-a" ou "isa") (hull; king, 1987).
b) Agregação - Esta abstração ocorre quando objetos são agrupados em um relacionamento de composição, onde cada objeto contribui para a formação de visualizações específicas de um objeto maior. Por exemplo: os objetos "página", "capa", "encadernação" e "tinta" podem ser agrupados para formar uma visão do objeto "livro"; enquanto o "título", "autor" e "editor" podem ser agrupados para formar uma outra visão. Este relacionamento é também conhecido como relacionamento "é-parte-de" (em inglês: "is-part- of").
c) Classificação - Esta abstração ocorre quando objetos são agrupados por serem exemplos particulares de um tipo mais geral. Por exemplo: "Sagarana" e "O estorvo" são exemplos de "novelas". Este relacionamento é também conhecido como relacionamento "é-exemplo-de" (em inglês: "is-instance-of").
d) Associação - Esta abstração ocorre quando objetos são agrupados segundo a sua virtude em satisfazer algum critério. Este relacionamento é também conhecido como relacionamento "é-membro-de" (em inglês: "is-member-of").

Hull & King (1987) fazem, ainda, uma distinção entre modelos semânticos que "incorporam aspectos estruturais dos objetos" e modelos orientados a objeto que "incorporam aspectos comportamentais dos objetos". O enfoque da orientação a objeto "simplesmente desloca a ênfase do relacionamento entre os componentes do modelo para o comportamento dos componentes individuais ou componentes grupais, onde quanto mais graus de liberdade estão presente no comportamento do objeto mais fortemente ele será caracterizado".

Conforme visto, os modelos que mais e mais desempenham um papel importante como recurso metodológico para todas as áreas cujo interesse são os fenômenos relacionados à informação, como informática e ciência da informação. Acerca do papel das ferramentas de modelagem para o desenvolvimento das atividades profissionais dos cientistas da informação, Burt & Kinnucan (1990) afirmam que "os cientistas da informação (...) podem encontrar, nas técnicas de modelagem, um mecanismo útil para capturar e comunicar seus conhecimentos sobre fontes de informação e sobre padrões de comportamento de quem busca informação. Os modelos resultantes podem ser amplamente desenvolvidos mediante seleção e composição de conceitos e técnicas de modelagem provenientes de várias disciplinas (informática, psicologia, física, lingüística e outras).

Por fim, mais pesquisas sobre como as pessoas usam, selecionam e se posicionam diante da informação são extremamente necessárias para a concepção e projetos de sistemas de informação que preencham com mais completeza as necessidades dos usuários desses sistemas.

Referências

ALLEN, Bryce L. *Cognitive research in information science*. **Annual Review of Information Science and Technology**, v. 26, p. 3-37, 1991.

ALMEIDA , Elizabeth; TAUHATA, Luiz. **Física nuclear**. Rio de Janeiro: Guanabara Dois, 1981. 413p.

ANSI/X3/SP RC. Study group on data base management system. **FDT-Bulletin**, v. 7, n. 2, 1975.

APOSTEL, L. *Toward the formal study of models in the non-formal sciences*. In: FREUDENTH L, H. The concept and the role of the model in mathematics and natural and socials sciences. Amsterdam: Dordtreccht, 1991. p. 1-37.

BELKIN, Nicholas J., ROBERTSON, Sthephen E. *Information Science and the phenomenon of information.* **Journal of the American Society for Information Science**, p. 197-2O4, Jul./ ug. 1976.

BELKIN, Nicholas J. *Anomalous states of knowledges as a basis for information retrieval*. **Canadian Journal of Information** Science, v. 5, p. 133- 140, 1980.

BERTALANFFY, Ludwig. *General system theory: a critical review.* **General Systems**, v.7, p.1- 20, 1962.
BLACK, M. Models and methaphors. New York: Ithaca, 1962. 267 p.
BORGMAN, Christine L. The user's mental model of an information retrieval system: an experiment on a prototype online catalog. **International Journal of man-machine studies**. v. 24, p. 47-64, 1986.
BURT, Patricia; KINNUC N, Mark. *Information models and modelling techniques for information systems.* Annual Review of Information Science and Technology, p.175-208, 1990.
CAPRA , Fritjof. **O tao da física.** São Paulo : Cultrix, 1983. 160p.
CHEN, Peter P. *The entity-relational model - toward a unified view of data.* **ACM Transactions on Database Systems**, v. 1, n. 1, p. 9-36, Mar. 1976.
CHORLEY, Richard; HAGGETT, Peter. Modelos, paradigmas e a nova geografia. In: CHORLEY, Richard; HAGGETT, Peter. **Modelos sócios- econômicos em geografia**. Rio de janeiro: Livros Técnicos e Científicos/USP, 1975. p.1-22.
CODD, E. F. *Extending the relational model to capture more meaning.* **ACM Transactions on Database Systems**, v. 4, n. 4, p. 397-434, Dec. 1979.
DANIELS, P.J. *Cognitive models in information retrieval: an evaluative review.* **Journal of Documentation**, v. 42, n. 4, p. 272-304, Dec. 1986.
EISBERG, Robert; RESNICK, Robert. Quantum physics of atoms, molecules, solids, nuclei, and particles. New York: John Wiley, 1974. 713p.
ELLIS, David. *The physical and cognitive paradigms in information retrieval research.* **Journal of Documentation**, v. 48, n. 1, p. 389-392, Mar. 1992.
FARRADANE, J. *The nature of information.* **Journal of Information Science**, v. 1, n. 1, p. 13-17, 1979.
GAINES, Brian R. *Modeling and forecasting the information sciences***. Information Sciences**, v. 57-58, p. 3-22, 1991.
GILBERT, Nigel G. Cognitive and social models of the user. In: **INTERACT: HUMAN**-computer **INTERACTION,** 87., 1987, Amsterdan. Proceeding of the Second IFIP Conference. North-Holland : Elsevier, 1987. p.165-169.
HESSE, M. *Models in physics.* **British Journal of the Philosophy of Science,** v. 4, p. 198-214, 1953-1954.
HULL, Richard; KING, Roger. *Semantic database modelling: survey, aplications, and research issues*. **ACM Computing Surveys**, v. 19, n. 3, p. 2O1-26O, Sep. 1987.
HUMPHREY, Ronald. How work roles influence perception: structural cognitive processes and organizational behavior. **American Sociological Review**, v. 5O, n. 2, p. 242-252, 1985.
KAPLAN, A. **The conduct of inquiry.** San Francisco: Chandler, 1964. 428 p.

KUHN, T.S. **The structure of scientific revolutions**. Chicago: University of Chicago, 1970.
LYYTINEN, Kalle. Two views of information modelling. **Information & Management**, v. 12, p. 9-19, 1987
MASTERM N, M. *The nature of a paradigm*. In: LAKATOS, I.,MUSGRAVE, A. **Criticsim and the growth of knowledge**. Cambridge: Cambridge University, 197O. p. 59-91.
PIAGET, Jean. *The concept of structure.* In: **SCIENTIFIC thought**: concepts, methods and procedures. Paris: Unesco, 1972. p. 35-56.
RICH, Elaine. *User modelling via esterotypes.* **Cognitive Sciences**, v. 3, p. 329-352, 1979.
RICH, Elaine. *Users are individuals: individualizing user models.* **International Journal of Man-Machine Studies,** v. 18, n. 3, p. 199-214, Mar. 1983.
SETZER, Waldemar. **Projeto lógico e projeto físico de base de dados**. Belo Horizonte: UFMG, 1986. 284p.
SHANNON, Claude; WEAVER, Warren. **The mathematical theory of communication**. Urbana: University of Illinois Press, 1949.
SKILLING, H. *An operational view*. **American Scientist**. v.52, p. 388- 396, 1964.
SMITH, Linda C.; WARNER, Amy J. *A taxonomy of representations in information retrieval system design*. **Journal of Information Science**, v. 8, p. 113-121, 1984.
SOUZA, MENDONÇA A. DE; DODEBEI, V. L. **Modelos e sistemas em ciência da informação.** Rio de Janeiro, 1993. 20 p. (Seminário apresentado à disciplina Linguagem e Ciência da Informação III. Curso de Doutorado em Ciência da Informação).
STACHOWIAC, Herbert. Models. In: SCIENTIFIC thought: concepts, methods and procedures. Paris : Unesco, 1972, p. 145-166.
TSICHRITZIS, D., LOCHOVSKY, F.H. **Data models**. New Jersey : Prentice-Hall, 1982. 381 p.
VICKERY, B. C. *Knowledge representation: a brief review*. **Journal of Documentation**, v. 42, n. 3, p. 145-159, Sep. 1986.
WIESER, Wolfgang. **Organismos, estruturas, máquinas: para uma estrutura do organismo.** São Paulo: Cultrix, 1972, 124 p.

PARTE 2

SISTEMAS DE INFORMAÇÃO, BIBLIOTECAS DIGITAIS E INTEROPERABILIDADE

Documentos digitais e novas formas de cooperação entre sistemas de informação em C&T[1]

Carlos Henrique Marcondes[2] e Luís Fernando Sayão[3]

1 Introdução

É inquestionável o papel central que desempenham hoje as tecnologias de informática, computação e comunicação nas práticas de informação (marcondes, 1999). Quando se fala em informação para ciência e tecnologia, este papel é mais acentuado ainda. Isto porque a ciência institucionalizada está assentada em mecanismos de comunicação rápida dos resultados de pesquisa, que por sua vez estão hoje baseados fortemente nas tecnologias de informação. No ciclo de comunicação científica, as bibliotecas têm um papel fundamental. A elas cabem, neste ciclo, os papéis de coleta, registro, estocagem e disseminação de informações. A evolução das tecnologias de informação, no entanto, vem alterando substancialmente este papel e junto com isto o próprio conceito de biblioteca.

Um aspecto problemático da cultura de nosso tempo relacionado à questão informacional é o assim chamado fenômeno da explosão informacional, a grande quantidade de informações produzidas e disponibilizadas por diferentes atividades sociais, dificultando sua identificação, acesso e utilização. Desde a emergência deste fenômeno, a partir da metade do século xx, a saída encontrada pelas bibliotecas foi a cooperação: a publicação das fichas do catálogo da *Library of Congress* (eua) data de 1905, o projeto marc é do fim da década de 60 (robredo, 1994). Desde a invenção do computador na década de 50, as tecnologias de informação passaram a ser usadas pelas bibliotecas para prover acesso não só a documentos dos seus próprios acervos, mas também aos armazenados em acervos de outras bibliotecas (marcondes, 2001). Catálogos coletivos e bases de dados automatizadas, me-

1 Trabalho apresentado na Reunión de Especialistas en Información Cientifica Digital, promovida pela Bireme/ OPS/ OMS e Unesco, São Paulo, 26-27 de março de 2002.

2 Doutor em Ciência da Informação. Universidade Federal Fluminense. marcon@vm.uff.br.

3 Doutor em Ciência da Informação. Comissão Nacional de Energia Nuclear. luis.sayao@cnen.gov.br.

canismos de empréstimo e provisão de cópias digitalizadas de documentos entre bibliotecas são exemplos do uso de tecnologias de informação para prover acesso a documentos (ainda em papel), disponíveis em acervos para além dos muros de cada biblioteca.

O surgimento da Internet a partir dos anos 90 vem mudando de maneira radical o papel das bibliotecas no ciclo intermediação e acesso ao documento. As possibilidades abertas pela Internet com seus mecanismos de publicação direta na rede tornam o acesso a um documento digital uma mera questão de conhecer sua URL. No entanto, esta facilidade de acesso tem como contrapartida a grande dificuldade de encontrar informação relevante, às atividades de *information discovery*. Encontrar a informação relevante é fundamental para que a mesma possa ser utilizada. O uso dos mecanismos de busca gerais é uma solução parcial para o problema. Shneiderman, Byrd e Croft (1997), discutindo a eficácia dos mecanismos de busca de uso geral disponíveis na Internet, diz que: "Embora mecanismos de busca como as *Infoseek*, *AltaVista*, *Lycos*, *WebCrawler* e *Open Text* sejam largamente usados, existe um consenso público e geral e entre profissionais acerca das grandes dificuldades em se buscar informações".

Um relatório do NEC Research Institute publicado pelo boletim Edupage, em português, datado de 3 de abril de 1998, relata que, na ocasião, os melhores mecanismos de busca cobriam não mais de 30% de todas as páginas Web. Esta afirmação reforça a dimensão do problema localização/identificação colocado pela Internet, devido principalmente ao seguinte fator: O estudo mais recente de Bergman (2001) sobre a "*Deep Web*"X "*Surface Web*" enfatiza a gravidade da questão.

- grande quantidade de informações e seu número cresce de forma acelerada; acentua-se a chamada "explosão informacional", existem bilhões de páginas; publica-se de tudo, sobretudo, de forma caótica, na Internet;
- a Internet não é como uma biblioteca: não existe ordem, a informação é disponibilizada de maneira caótica;
- a informação disponibilizada é sobre uma infinidade de temas, sob os mais diferentes enfoques; isto é, um problema adicional para sua recuperação; não é à toa que os profissionais de informação, principalmente aqueles que trabalham com informação especializada para ciência e tecnologia, desenvolveram técnicas como vocabulários controlados e linguagens artificiais, técnicas de indexação pós-coordenada, conectivos booleanos para coordenar conceitos temáticos simples, formando conceitos temáticos complexos etc.;
- a informação é disponibilizada em diferentes idiomas, o que agrava o problema do controle do vocabulário a ser usado na recuperação;

- os mecanismos de busca gerais indexam a Internet periodicamente, de forma automática, cegamente, sem compreender o tema de uma página, simplesmente extraindo palavras do texto HTML da página e armazenando estas palavras isoladas junto ao endereço da página, em uma base de dados para consulta. Além disso, a indexação é feita por páginas HTML isoladas, não considerando que diversas páginas estão inter-relacionadas, formando um site;
- os programas-robôs dos mecanismos de busca só "enxergam" páginas HTML estáticas quando fazem sua rotina de indexação, deixando de considerar grande quantidade de informações sob a forma de registros contidos em bases de dados disponíveis na Internet – a chamada *deep web* – (BERGMAN, 2001), que fica assim "invisível". Estes registros são acessados somente por meio das interfaces destas bases de dados, que pressupõem uma interação entre um usuário humano e a base de dados e, portanto, ficam inacessíveis aos programas-robôs. Segundo Bergman, a *surface web* é estimada em cerca de 2,5 bilhões de páginas, enquanto a *deep web* seria cerca de 500 vezes maior;
- tem sido extremamente difícil estabelecer filtros de qualidade para as informações encontradas na Internet.

O surgimento da Internet vem tendo um impacto significativo também nas formas de comunicação científica e, consequentemente, nos sistemas de informação em C&T. Diferentes processos sociais, econômicos e tecnológicos convergem para configurar a situação atual das formas de comunicação científica. Desde o surgimento do primeiro arquivo eletrônico de *preprints*, ou *eprints*, o ArXiv, no *Los Alamos National Laboratory*, criado em 1991 pelo físico Paul Ginsparg (GINSPARG, 1996), que a própria comunidade científica internacional oferece uma alternativa prática para a publicação de seus trabalhos e sua disponibilização gratuita como alternativa aos periódicos científicos controlados pelos grandes editores internacionais. Este debate envolvendo pesquisadores e grandes editores vêm sendo coberto pela revista *Nature*[4].

Pesquisadores passaram a criar arquivos eletrônicos de *preprints* e *posprints* como alternativa para publicação direta de seus trabalhos em texto completo, os assim chamados *open archives*. Este movimento vem crescendo desde então e como resultado deste debate surgem iniciativas como a NCSTRL – *Networked Computer Science Technical Report Library*[5], a NDLTD – *Networked Digital Library of Theses*

4 Disponível em: https://www.nature.com/. Acesso em: 1 dez. 2002.

5 Disponível em: http://www.ncstrl.or. Acesso em: 1 dez. 2002.

and Dissertations[6] – e mais recentemente o PUBMED *Central*[7] – *Public Library of Science*[8] – e *Open Archives Initiative*[9].

Os padrões de tecnologia da informação utilizados ou derivados da *Open Archives Initiative* têm um impacto potencial muito grande sobre os sistemas de informação em C&T, afetando substancialmente a maneira como bibliotecas e centros de documentação desempenham funções tradicionais, como seleção, aquisição, registro/tratamento técnico/indexação/classificação e disseminação. Colocam também a questão da cooperação entre sistemas de informação com vistas ao acesso à informação em C&T em um novo patamar. Este trabalho se propõe a discutir este novo quadro que surge com a emergência de documentos digitais em C&T, publicados diretamente na rede e armazenados em bibliotecas digitais e repositórios como os *Open Archives*, e como estes padrões e tecnologias podem ser usados pelos países em desenvolvimento para ampliar o alcance e o nível de cooperação de seus sistemas de informação.

2 Documentos e arquivos digitais em ciência e tecnologia

A ciência hoje é uma atividade altamente institucionalizada. O conhecimento passa a ter um papel fundamental como insumo produtivo, em um processo que tem suas origens no Renascimento e se acentua fortemente na Revolução Industrial (GONZÁLEZ DE GÓMEZ, 1987). Sua gestão, a utilização e principalmente o acesso se tornam cada vez mais atividades econômicas críticas (MARCONDES, 2001).

Nesta ciência tão institucionalizada, não existe praticamente lugar para o gênio isolado, capaz de dar conta de uma descoberta científica do início ao fim. A ciência atual é fundamentalmente um trabalho coletivo, em que pesquisadores e grupos de pesquisa trabalham sobre resultados já obtidos por seus pares, e tem como objetivo acrescentar um tijolo a mais em um vasto edifício.

Daí o papel fundamental que desempenha neste contexto a comunicação científica. A ciência não pode avançar sem mecanismos eficientes de comunicação científica que integrem em um ciclo a produção de conhecimento, o registro dos resultados, a coleta e estocagem destes registros, a disseminação dos resultados e o reuso, tanto em atividades produtivas quanto como fonte para gerar novos conhecimentos. Até muito recentemente, o periódico científico tradicional era o veículo por excelência para a comunicação científica, mas estamos no meio de uma profunda mudança desta situação (DAY, 1999).

6 Disponível em: http://www.ndltd.org. Acesso em: 22 ago. 2021.
7 Disponível em: http://www.pubmedcentral.nih.gov/. Acesso em: 22 ago. 2021.
8 Disponível em: http://www.publiclibraryofscience.org/. Acesso em: 22 ago. 2021.
9 Disponível em: http://www.openarchives.org. Acesso em: 22 ago. 2021.

Os mecanismos de comunicação científica combinavam canais desde os informais, como conversas pessoais, cartas, telefonemas, *preprints* distribuídos aos pares, bastante ágeis e atualizados em termos de velocidade, mas sem qualquer filtro de qualidade, com canais cada vez mais formais, como as comunicações e trabalhos em congressos e os artigos de periódico, em que os trabalhos para serem aceitos são submetidos à avaliação por pares. No percurso dos canais informais para os formais, a maior rapidez na comunicação é trocada pela maior lentidão da publicação dos resultados de pesquisas e pela maior qualidade graças à revisão por pares (ziman, 1979; le coadic, 1996).

Desde a publicação do *Journal des Savans* e das *Philosophical Transactions* da *Royal Society* , em 1665 (day, 1999), que o periódico acadêmico vem tendo um papel destacado como veículo por excelência de comunicação científica. A partir do fim do século xix, com a instituição do sistema de revisão por pares, este papel se institucionalizou, juntamente com todo aparato acadêmico e científico, atingindo a configuração institucional atual. No mundo dos documentos impressos, os periódicos científicos tinham papel destacado como coroamento de um sistema de comunicação científica institucionalizado e reconhecido pela comunidade acadêmica, que fazia um compromisso entre velocidade e filtros de qualidade. Dentre os papéis por eles cumpridos, destacam-se os seguintes: disseminação ampla dos resultados de pesquisa; controle de qualidade, através do mecanismo de revisão por pares; homologação de prioridade nas descobertas científicas; reconhecimento dos autores; criação de um arquivo público de conhecimentos com cópias armazenadas em bibliotecas de centros de documentação (day, 1999).

Paralelamente à criação e à institucionalização destes mecanismos de comunicação científica e de filtros de qualidade, vive-se um processo de parcelamento e fragmentação da ciência em áreas cada vez mais especializadas. A isto se soma um processo de identificação e questionamento de novos problemas de pesquisa que demandam o aporte de diferentes áreas do conhecimento como as ciências cognitivas, a ecologia, a moderna visão integrada das ciências da vida etc., levando à identificação de pontos de contato entre diferentes ramos da ciência.

A consequência deste processo para a comunicação científica é a fragmentação e o surgimento de novos periódicos e eventos científicos, dedicados a novas áreas ou a novos problemas de caráter interdisciplinar. Devido à fragmentação descrita, estes veículos de comunicação científica têm audiências cada vez mais restritas e especializadas, levando à diminuição das edições e ao consequente encarecimento do custo final dos periódicos científicos e edição de anais de eventos para seus usuários.

Entretanto, a produção destes veículos de comunicação científica tornou-se um negócio que movimenta vultosas somas e é dominado por grandes empresas,

os grandes editores científicos. Os custos das assinaturas de periódicos cada vez mais proibitivos para bibliotecas e centros de documentação, como intermediárias no ciclo de comunicação científica, ameaçam quebrar este ciclo. Este processo vem tendo consequências sérias para o desenvolvimento da ciência e especialmente sérias para países em desenvolvimento, com recursos escassos a serem investidos na ciência e em apoio bibliográfico para a mesma, como o Brasil.

O surgimento da Internet e dos mecanismos de publicação direta na rede tem sido visto pela comunidade acadêmica como uma possível alternativa. O maior retorno que a comunidade acadêmica almeja, publicando os resultados de suas pesquisas, é que estes possam servir de bases a outras pesquisas, sendo citados por outros trabalhos. A citação é a medida clássica do prestígio e do valor de uma contribuição para a ciência em geral. Com base nos mecanismos de citação, são criados indicadores consagrados para cientometria e bibliometria como o *Science Citation Index*, produzido pelo ISI - *Institute of Scientific Information*[10]. Estudos recentes confirmam que as publicações eletrônicas são muito mais citadas que as publicações em papel: *"The mean number of citations to offline articles is 2,74, and the mean number of citations of online articles is 7.03, an increase of 157%"* (LAWRENCE, 2001, p. 521).

À medida que os periódicos científicos se tornam crescentemente mais caros, vem diminuindo sua audiência, resultando disso menor impacto das comunicações nele veiculadas. Segundo este ponto de vista, a cobrança e as restrições ao acesso por parte dos grandes editores científicos internacionais impediriam o livre fluxo dos resultados da pesquisa e o próprio avanço desta, em prol de interesses comerciais restritos. As consequências deste processo para o desenvolvimento da ciência começam a ser percebidas por um número crescente de cientistas e pesquisadores que se lançaram na busca de alternativas. Stevan Harnard, um dos ideólogos deste movimento, assim formula esta questão:

> *Unlike the authors of books and magazine articles, who write for royalty or fees, the authors of refereed journal articles write only for 'research impact'. To be cited and built on in the research of others, their findings have to be accessible to their potential users. From the authors' viewpoint, toll-gating access to their findings is as counterproductive as toll-gating access to commercial advertisements* (HANARD, 2001, p. 1024).

E ainda:

10 Disponível em: http://www.isinet.com. Acesso em: 1 dez. 2002.

> *Researchers never benefited from the fact that people had to pay access tolls to read their papers (as subscriptions, and for the online version, site--licences or pay-per-view). On the contrary, those access barriers represent impact barriers for researchers, whose careers and standing depend largely on the visibility and uptake of their research* (HANARD, 2001, p. 1024).

Este debate envolvendo pesquisadores e grande editores está sendo coberto pela revista *Nature*[11].

A comunidade científica vê as publicações eletrônicas na rede como um meio de aumentar sua visibilidade, acelerar o avanço da ciência e disseminar amplamente os resultados das pesquisas, vistas como patrimônio da humanidade (HARNARD, 2001). Ela soube, com muita propriedade, acercar-se das novas possibilidades abertas pelas tecnologias de informação para criar mecanismos alternativos de publicação de resultados das atividades de pesquisa.

Observa-se, nos últimos anos, uma forte tendência, surgida no seio da comunidade científica mundial, para a criação de arquivos eletrônicos informais e autogeridos, voltados para este fim. Estes arquivos são conhecidos como *eprints archives* e configuram claramente uma transição do modelo de comunicação tradicional, baseado em publicações periódicas formalmente estabelecidas, para um novo e surpreendente paradigma. Eles exemplificam o mais democrático e eficiente modelo para disseminação de resultados de pesquisa. Uma referência importante em português sobre o histórico, os desenvolvimentos recentes e a dimensão da *Open Archives Initiative*, inclusive com uma entrevista a Paul Ginsparg, pode ser encontrada em Sena (2000).

Além das condicionantes tecnológicas oferecidas pela Internet, que facilitaram extraordinariamente a publicação de documentos eletrônicos, há um consenso absoluto de que outros fatores foram decisivos na emergência dos arquivos de *eprints*, principalmente os seguintes: a) a lentidão do ciclo de edição das revistas comparado à rapidez da geração de novos conhecimentos de algumas áreas; b) a renúncia ao direito sobre a obra imposto pelas revistas, que impede a ampla disseminação pelo autor de suas ideias; c) a perspectiva extremamente rígida e conservadora dos esquemas de *peer review* adotados pelas revistas, que não raro são impeditivos ao surgimento de ideias inovadoras, como é ilustrado no artigo de Martin (1993) sobre uma hipótese pouco ortodoxa sobre o surgimento da Aids, cuja publicação foi recusada pelos principais periódicos do *establishment* médico; d) o alto custo da subscrição dos periódicos, seja em papel ou em meio eletrônico. Os arquivos aber-

11 Disponível em: www.nature.com. Acesso em: 22 ago. 2021.

tos não são, como possa aparecer a princípio, uma proposta anárquica, que elimina os critérios de qualidade da ciência; muitos deles utilizam o esquema de *peer-review*, ou uma separação entre textos avaliados e não-avaliados; muitos destes textos são cópias livres de artigos já publicados ou a serem publicados em periódicos convencionais (HANARD, 2001).

Muito recentemente, em 1999, em um passo seguinte à criação dos arquivos *eprints*, a comunidade científica internacional se mobilizou para torná-los interoperáveis, isto é, passíveis de serem consultados simultaneamente . Esta interoperabilidade foi alcançada mediante adoção de um conjunto de especificações técnicas e princípios organizacionais bastante simples, porém potencialmente poderosos e de grande alcance, no objetivo de integração desses arquivos. Esta iniciativa é conhecida como OAI – *Open Archive Initiative*, http://www.openarchives.org/ (SOMPEL; LAGOZE, 2000) e tem como objetivos básicos apoiar o desenvolvimento de arquivos de *eprints* e criar uma arquitetura tecnológica padronizada que sustente a interoperabilidade entre eles. No bojo da *Open Archives Initiative*, foram desenvolvidas tecnologias, padrões e metodologias para publicação, disponibilização, metadados e intercâmbio automático de metadados entre bibliotecas digitais. A dimensão da iniciativa dos *Open Archives* em nível mundial e sua cobertura regional e temática podem ser mais bem avaliadas consultando-se a lista dos arquivos eletrônicos existentes[12].

Um repositório eletrônico aberto apresenta características específicas. Seu site apresenta facilidades que permitem a um autor submeter diretamente seus trabalhos, armazená-los em forma digital permanentemente, editá-los, substituí-los e receber críticas e contribuições; ao submeter um trabalho para armazená-lo e disponibilizá-lo no arquivo eletrônico, um autor também o descreve, em um formulário de catalogação, de onde serão extraídos os metadados como autor, título, idioma, assunto etc. que permitirão recuperar o documento; os metadados são, portanto, obtidos como um subproduto da submissão de um documento. O site permite também a consulta e o acesso direto aos trabalhos eletrônicos nele armazenado. Um servidor de *eprints* compatível com o *Open Archives Initiative Protocol for Metadata Harvesting* – OAI PMH – permitirá a exposição de metadados dos trabalhos nele armazenados para coleta automática (*harvesting*) e reuso por provedores de serviços de informação, que com eles podem criar diferentes serviços de valor agregado.

A publicação na rede de textos completos de interesse para C&T começa a ser uma realidade também no Brasil. Iniciativas pioneiras como o SCIELO[13] (PACKER,

12 Disponível em: http://www.osti.gov/eprints/ppnbrowse.html. Acesso em: 1 dez. 2002.

13 Disponível em: http://www.scielo.br. Acesso em: 22 ago. 2021.

1998), um portal que abrange dezenas de periódicos, associado a uma metodologia para publicar e prover acesso a periódicos eletrônicos em texto completo, os diferentes periódicos brasileiros já publicados na Web em texto completo, a proposta de desenvolvimento de um ambiente Web para edição de anais de congresso pelo CIN/CNEN, as recentes iniciativas do arquivo de *eprints* do IMPA (CHATAIGNIER, 2001), as publicações digitais da biblioteca digital do LAMBDA – Laboratório de Automação de Museus, Bibliotecas Digitais e Arquivos da PUC-Rio[14], o Banco de Teses e Dissertações do Programa de Pós-graduação em Engenharia de Produção da UFSC[15] e do repositório de teses da USP[16], (MASIERO, 2001) mostram que esta é uma tendência irreversível aqui também.

Recentemente (MARCONDES; SAYÃO, 2001) o Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), por meio do projeto da Biblioteca Digital Brasileira em C&T (BDB), passou a fomentar o desenvolvimento de recursos informacionais brasileiros de interesse para C&T em texto completo, como teses e dissertações, artigos de periódicos, trabalhos em congressos, arquivos eletrônicos de *preprints*, integrando e provendo interoperabilidade entre estes recursos mediante acesso unificado aos mesmos, via única interface Web. O projeto prevê o uso do OAI PMH como um dos mecanismos para prover interoperabilidade entre os diversos recursos informacionais contendo documentos digitais de interesse para C&T, coletando seus metadados para uma base comum de onde será feito o acesso.

3 Cooperação entre sistemas de informação na era dos documentos e arquivos digitais

A produção da ciência e a comunicação dos resultados de pesquisa na nossa sociedade se dão de forma dispersa, segundo dois eixos: dispersa espacialmente – pesquisadores em diversos locais geográficos produzem e comunicam seus resultados de pesquisa por mecanismos de comunicação científica diferentes; dispersa temporalmente – pesquisadores produzem e comunicam seus resultados em momentos diversos.

Neste contexto, os sistemas de informação desempenham um papel fundamental na economia da informação-conhecimento, agregando valor ao servirem como pontos focais, que concentram a informação científica, produzida, por natureza, de forma dispersa. A eles os usuários recorrem para encontrar as informações de que necessitam, com um alto grau de probabilidade de encontrá-las.

14 Disponível em: http://www.ele.puc-rio.br/posgraduacao/laboratorios/lambda/. Acesso em: 22 ago. 2021.

15 Disponível em:http://teses.eps.ufsc.br/. Acesso em: 1 dez. 2002.

16 Disponível em: http://www.teses.usp.br. Acesso em: 22 ago. 2021.

Tradicionalmente a informação de interesse para a pesquisa científica é em grande parte composta pela chama da documentação não-convencional, também chamada de "literatura cinzenta", documentos que não são encontrados no circuito editorial convencional, como relatórios de pesquisa, trabalhos apresentados em eventos, *preprints*, teses e dissertações, que noticiam com grande atualidade os resultados de pesquisa.

O papel dos sistemas de informação sempre foi o de se contrapor a esta dispersão, provendo um ponto de concentração para a comunicação dos resultados das pesquisas. Os sistemas especializados em ICT organizaram-se para prover acesso principalmente a este tipo de documentação, a "literatura cinzenta", como o NTIS, o *Dissertation Abstracts*, o *Chemical Abstracts*, INIS, AGRIS, ERIC e muitos outros. Estes sistemas trabalham dentro do paradigma organizacional/metodológico/ tecnológico da informação referencial como etapa para o acesso a cópias de documentos em papel.

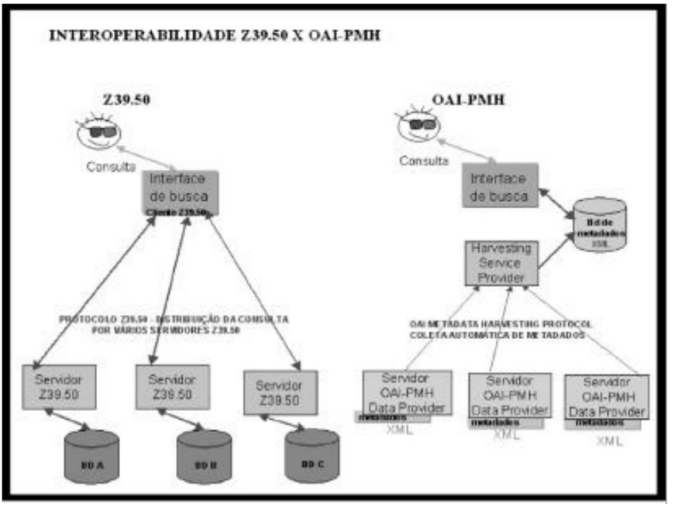
Coletar esta "literatura cinzenta" sempre foi caro e extremamente trabalhoso para os sistemas de ICT, devido à dispersão da literatura científica. O surgimento das publicações eletrônicas começa a mudar radicalmente este quadro. Hoje em dia não é mais suficiente para garantir o máximo de visibilidade de seu acervo que bibliotecas digitais simplesmente disponibilizem seus dados na Internet. A quantidade de informações disponível na rede é tão grande que identificar, localizar, descobrir a existência e acessar informações relevantes torna-se um problema crítico, demandando um tempo proibitivo aos usuários.

Às bibliotecas digitais é novamente colocada, como já foi há tempos para as bibliotecas convencionais, a questão de cooperarem, agora sob novas bases organizacionais e tecnológicas, para garantir o máximo de visibilidade a seus acervos. Atingir esta visibilidade não significa mais necessariamente que alguém buscando informações terá de acessar o site da biblioteca digital ou arquivo eletrônico para ter acesso aos documentos digitais nele depositados. A possibilidade que seus acervos possam ser consultados simultaneamente, sem que um usuário acesse cada site individualmente, a chamada *interoperabilidade*, tem sido perseguida como um mecanismo que viabilize esta possibilidade. Atingir a interoperabilidade entre repositórios de *eprints* ou bibliotecas digitais, distintos e heterogêneos, possibilitando que possam ser consultados simultaneamente, envolve um aporte intenso em termos de tecnologias, protocolos e padronização.

O OAI PMH é um protocolo que provê interoperabilidade não imediata (ou seja, não é, portanto, um protocolo para busca online) entre repositórios de *eprints*, bibliotecas digitais ou qualquer servidor na rede que queira *expor*, ou seja, tornar visíveis metadados de documentos nele armazenados para um programa externo

que queira coletá-los. Os participantes da *Open Archives Initiative* rejeitaram opções como, por exemplo, o protocolo de recuperação de informações Z39.50, que distribui uma busca imediata e simultânea por vários servidores que hospedam catálogos de bibliotecas capazes de resolvê-la, em favor de uma solução mais simples e menos onerosa em termos de recursos computacionais consumidos (TROLL, 2001). As diferenças na interoperabilidade com o uso do protocolo Z39.50 x o OAI-PMH podem ser bem visualizadas na figura 1, a seguir.

Figura 1 - Interoperabilidade Z39.50 X OAI-PMH.



Fonte: Os autores

O OAI PMH define a troca de solicitações e de metadados entre o servidor de *eprints* e um programa robô externo. Dentro da concepção OAI, existem as instituições chamadas provedoras de dados (*Data Providers*), que são bancos de documentos eletrônicos que oferecem facilidades para publicação e armazenamento de documentos eletrônicos e sua disponibilização em um servidor conectado à Internet, e as instituições provedoras de serviços (*Service Providers*), que coletam metadados de um ou mais provedores de serviço e com estes metadados prestam serviços de valor agregado. Exemplos destes serviços seriam o acesso unificado a acervos de diferentes provedores de dados, por meio de um portal Web único ou a constituição de bases de dados qualificadas sobre temas específicos, ou um periódico eletrônico com textos avaliados e submetidos a um esquema de *peer-review*, desenvolvido a partir dos metadados coletados de diversos provedores de dados.

A troca de mensagens entre o servidor do provedor de dados e o programa robô externo do provedor de serviços para a transferência de metadados é unidirecional – o provedor de serviços faz solicitações ao provedor de dados, que responde enviando metadados. As solicitações do provedor de serviço são feitas via protocolo HTTP[17], usando comandos CGI codificados por meio dos métodos GET ou POST. As solicitações são respondidas pelo provedor de dados com o envio de dados das respostas ou metadados dos documentos armazenados, codificados em XML[18]. O OAI PMH estabelece o *Dublin Core Metadata Element Set*[19] como conjunto mínimo de metadados a ser suportado pelos provedores de dados em resposta a uma solicitação de um provedor de serviços. No entanto, o provedor de serviços pode, a seu critério, oferecer outros formatos de metadados, mais amplos e complexos, como o MARC.

Os metadados, formando um registro de cada documento armazenado no provedor de dados, têm um identificador único, formado pelo identificador do provedor de dados mais um identificador do registro. Cada registro apresenta também um selo temporal, denominado *date stamp*, que indica a data da criação ou última alteração do documento associado a este registro; o *date stamp* é a chave que permite a coleta automática dos metadados do provedor de dados a partir de uma determinada data, possibilitando, portanto, a sincronização entre os registros do provedor de dados e de um provedor de serviços que forneça um serviço de acesso simultâneo a metadados de documentos armazenados em diversos provedores de serviço.

O protocolo prevê ao todo seis "verbos" que um programa robô de provedor de serviços pode enviar a um provedor de dados para coletar metadados de documentos aí armazenados: *identify* – obtém dados administrativos sobre o provedor de dados, sua política de publicação de documentos, seu escopo etc.; *ListSets* – lista as classificações sob as quais os documentos são organizados no provedor de dados; *ListMetadataFormats* – lista os formatos de metadados por meio dos quais os metadados dos documentos armazenados no provedor de dados podem ser apresentados; *ListIdentifiers* – lista os identificadores de registros armazenados no provedor de dados, podendo opcionalmente limitar estes registros a partir de uma data, ou pertencentes a um *set*; *ListRecords* – lista os metadados dos registros armazenados no provedor de dados segundo um formato de metadados, especificando todos que pertencem a um *set* ou todos a partir de uma data; e *GetRecords* – obtém os

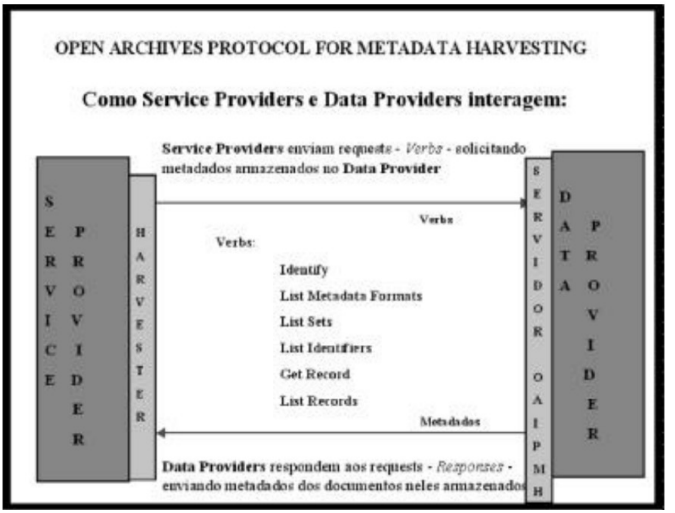
17 Disponível em: http://www.ietf.org/rfc/rfc2616.txt. Acesso em: 22 ago. 2021.

18 Disponível em: http://www.w3.org/XML/. Acesso em: 22 ago. 2021.

19 Disponível em: https://www.dublincore.org/specifications/dublin-core/dces/. Acesso em: 22 ago. 2021.

metadados dos registros armazenados segundo um formato de metadados, dado um identificador de registro. A interação entre *Service Providers e Data Providers* usando o OAI-PMH pode ser visualizada na figura 2, a seguir.

Figura 2 - Interação Service Providers e Data Providers.



Fonte: os autores

A interação entre provedores de dados e provedores de serviços usando o OAI PMH pode ser vista através do OAI *Repository Explorer*[20], uma interface interativa para testar a compatibilidade de repositórios de documentos eletrônicos com OAI PMH.

4 Novos serviços cooperativos viabilizados pelo OAI PMH

Os países em desenvolvimento têm dificuldades históricas, não só no acesso aos resultados de pesquisas, mas também em dar visibilidade à ciência por eles produzida, não obstante ser uma ciência de grande importância para uma parcela imensa da população mundial (exemplo mais eloquente disto são as pesquisa da Fiocruz que levaram à produção de várias substâncias que fazem parte do coquetel *antiaids* que podem ser um alívio para a tragédia africana). Isto acontece principalmente por conta da série de impeditivos encontrados para a inserção do produto dessa ciência nos sistemas de informação e nas bases de dados internacionais. Esses impeditivos são de natureza distinta, tais como idioma (tudo que não é publicado em inglês é considerado hieróglifo), publicação em periódicos não coletada pelas

20 Disponível em: http://oai.clarin-pl.eu/. Acesso em: 22 ago. 2021.

bases de dados internacionais, o laboratório que sedia a pesquisa está fora do circuito da *big science*, o não-enquadramento ao *mainstream* da ciência mundial e assim por diante (SAYÃO, 1996). Assim sendo, é de grande importância a criação de sistemas regionais, como o LILACS, que além de prestar serviços de informação aos seus usuários em todo o mundo, estabelecem meios de disseminação adequados à sua audiência e contribuem para incluir a ciência do Terceiro Mundo nos fluxos internacionais.

Informação em ciência e tecnologia livre na Internet, associada ao conjunto de metodologias colocadas à disposição da comunidade acadêmica pela *Open Archive Initiative* (OAI), abrem grandes possibilidades para os sistemas de informação que se dispuserem a avaliar com espírito criativo as oportunidades oferecidas por estas metodologias. Uma série de novos serviços baseados em reuso de metadados pode ser concebida, incluindo redes cooperativas e sistemas de informação regionais. O conjunto de metodologias OAI PMH – protocolo de coleta automática de metadados – é um protocolo de fácil implementação. Outro dado promissor é o que indica ser possível montar repositórios digitais *Open Archives* em servidores PC-Intel, rodando sistema operacional Unix/Linux com *softwares* que podem ser obtidos livremente da Web[21]. Estes repositórios serão compatíveis com o número crescente de repositórios digitais espalhados por todo mundo, nas mais diversas áreas, permitindo alto grau de interoperabilidade entre eles.

Nesta direção, as instituições desses países podem se aproveitar das metodologias OAI para expor os seus metadados aos repositórios digitais internacionais e assim dar mais visibilidade aos seus conteúdos; podem também reaproveitar os metadados disponíveis em escala planetária pelos repositórios digitais OAI para coleta automática, agregando valores e criando versões novas para os serviços e produtos de informação já tradicionais, como busca retrospectiva e DSI e, sobretudo, criando novos conceitos de serviços próprios para o ambiente de rede, como é o caso das bibliotecas servindo como publicadoras Web dos trabalhos de sua comunidade de usuários e dos esquemas de acompanhamento das novidades publicadas na Internet em áreas específicas e da descoberta de recursos informacionais.

A proposta de maior alcance potencial da OAI é a que diz respeito à integração de repositórios digitais livres por meio de serviços independentes. Esta proposta indica um caminho viável para a construção de uma estrutura global unificada para a literatura acadêmica, que inclui não somente *prepints*, mas também periódicos, relatórios, anais de congresso e outros tipos de literatura acadêmica. Em um outro eixo, esta estrutura global cria também facilidades para o surgimento de sis-

21 Disponível em: http://www.eprints.org. Acesso em: 22 ago. 2021.

temas de informação organicamente mais complexos, como são as redes cooperativas locais, regionais ou internacionais. O conjunto de metodologias e protocolos OAI torna o ônus administrativo-financeiro destas redes muito menor, a tarefa de criar os catálogos coletivos mais simples e, sobretudo, deixa os centros cooperantes com um grau maior de independência, principalmente para cooperar com outros serviços/sistemas aderentes ou não ao protocolo OAI. Por exemplo: poderíamos ter uma rede de teses eletrônicas das universidades paulistas que estivesse integrada ao Consórcio Brasileiro de Teses Eletrônicas, que, por sua vez, fosse cooperante, ao mesmo tempo, de uma rede de teses em língua portuguesa, formada pelos países que falam este idioma e também de uma rede de teses eletrônicas de países latino-americanos.

No entanto, para se chegar a esse nível de integração e articulação entre redes e sistemas de ICT, é necessária a criação de níveis distintos de interoperabilidade (MILLER, 2000), que se sobreponham à interoperabilidade tecnológica estabelecida pelo conjunto de padrões e protocolos preconizados pelo OAI e viabilizados facilmente pelo uso do *software Eprint* .

A interoperabilidade tem muitas faces: é ela que permite que sistemas de ICT distintos e heterogêneos possam aproveitar e agregar valor à informação criada por outro, gerando novos serviços e novas visões para a mesma informação. Não é, entretanto, uma questão meramente técnica e tecnológica, ela também depende da gestão, articulação e cooperação mútua entre sistemas de ICT no plano político. Esta faceta da interoperabilidade, aqui chamada de "interoperabilidade política", depende fundamentalmente da criação de organizações detentoras de canais e fóruns adequados, nos quais a discussão e o consenso possam se estabelecer e as decisões possam ser tomadas endossadas pelo grau de representatividade des- sas organizações. Este é o caso da DLF (*Digital Library Federation*), da PILA (*Publisher International Linking Association*) e do próprio OAI (*Open Archives Initiative*), para citar somente três iniciativas internacionais importantes neste momento.

Outra faceta importante do mesmo problema é a "interoperabilidade semântica". Ela requer o uso generalizado de instrumentos comuns de descrição temática. Um conjunto mínimo de metadados padronizados, uma linguagem de descrição temática de cobertura ampla, tal como a Tabela de Áreas do Conhecimento Capes/CNPq e a criação de servidores de autoridades cooperativos. A interoperabilidade semântica tem como reflexo imediato a melhoria na qualidade da recuperação das informações e a otimização da consulta a sistemas interligados. Esses resultados são bastante perceptíveis pelo usuário final.

A Internet nos traz outra forma de interoperabilidade que é a linkagem ou enlaces entre sistemas. A linkagem permite a navegação via *hiperlinks* entre as vá-

rias manifestações do trabalho acadêmico de um indivíduo, normalmente dispersas em vários sistemas, seja como autor, orientador ou membro de banca de teses ou dissertações eletrônicas, seja como autor de artigos de periódico, de trabalhos em congressos ou acessando seu currículo em um sistema de currículos. Mas é de fundamental importância a adoção de padrões e metodologias que garantam a persistência dos endereços eletrônicos dos recursos informacionais, como o PURL - *Persistent* URL[22] e o DOI – *Digital Object Identifier*[23], no sentido de preservar o investimento na linkagem entre sistemas.

A seguir são descritos alguns produtos e serviços de informação que podem ser implementados usando-se as metodologias OAI.

- **Formação de bases de dados**

A alternativa de gerar bases de dados a partir de informações disponíveis na Internet e do reuso de seus metadados se torna cada vez mais uma opção viável que começa a ser adotada por sistemas internacionais de grande porte como é o caso do *Energy Technology Data Exchange* (ETDE), da Agência Internacional de Energia. Isto acontece principalmente pelo alto custo da coleta manual e do tratamento técnico – catalogação e indexação –, fatores que têm inviabilizado a sustentação econômica do modelo tradicional de formação de bases de dados que, diga-se de passagem, sempre foi deficitário e precisava de fortes subsídios externos. A coleta automática de metadados – *harvesting* –, baseada no protocolo OAI, pode facilitar tremendamente a formação de bases de dados, diminuir drasticamente os custos de coleta, catalogação e indexação e também o custo de gestão destes processos. E o mais importante: as bases de dados organizadas dessa forma têm agregado um valor importantíssimo que é o acesso ao objeto digital referenciado, seja ele um texto, um vídeo ou uma peça de museu. Pode-se construir bases temáticas, qualificadas, por tipologia de documento (teses, relatórios), orientadas por problemas (epidemia de dengue, crise energética), orientadas por projetos ou missão (prospecção de petróleo em águas profundas, despoluição da Baía de Guanabara), com metadados e informações disponíveis por todo mundo.

- **Estabelecimento de redes cooperativas nacionais, regionais e internacionais**

O protocolo OAI pode facilitar sobremaneira o estabelecimento de redes de informação. Um exemplo importante é a rede de teses eletrônicas da *Virginia*

22 Disponível em: http://www.purl.org/. Acesso em: 22 ago. 2021.

23 Disponível em: http://www.doi.org. Acesso em: 22 ago. 2021.

Tech. University – EUA –, a NDLTD que tem alcance mundial. No Brasil, dentro do âmbito da BDB, tem-se o Consórcio Brasileiro de Teses Eletrônicas, projeto em andamento que, inicialmente, reúne as teses e dissertações eletrônicas da USP, UFSC, PUC-Rio, ENSP (Fiocruz) e conta com o apoio do CNPq. O Consórcio tem a perspectiva de estender-se por Portugal, por meio de convênio com a Universidade do Minho. O uso do Protocolo OAI pode ainda tornar-se uma alternativa a mais para as redes já estabelecidas, criando novas opções de contribuição e coleta menos onerosa em termos financeiros e gerenciais para os membros dessas redes. Existe uma série de *software* livres, do tipo *front-end*, que podem tornar uma base de dados de metadados aderente ao protocolo OAI sem interferências com os processos internos da rede. Muitos destes *software* podem ser encontrados em http://www.openarchives.org/tools/tools.html. É também possível estabelecer formas ainda mais simples de coleta automática baseadas, por exemplo, na transferência de arquivos-textos ou arquivos HTML, adequadamente formatados, via FTP ou via mecanismos simples de coleta automática, que podem implementar formas simples de crítica de dados.

- **Edição de periódicos eletrônicos**

O *software Eprint* (distribuído livremente) desenvolvido pela Universidade de Southampton – UK, para disseminar a implantação de repositórios digitais compatíveis com OAI, oferece facilidades interessantes para quem deseja editar revistas diretamente na rede. A característica mais marcante para esta aplicação é que o *software* implementa um ambiente de submissão eletrônica bastante sofisticado, que permite avaliação por parte de revisores e comentários vindos dos leitores, além de instrumentos para a gerência de publicação dos artigos em uma biblioteca digital. Está crescendo rapidamente o número de revistas eletrônicas publicadas com esta tecnologia, e isto indica mais um serviço que será descrito a seguir.

- **Catálogo coletivo de revistas eletrônicas OAI**

Já é possível implementar catálogos coletivos de revistas eletrônicas que tenham seus metadados coletados com o uso do Protocolo OAI, que permitam busca simultânea em todas as revistas e a recuperação dos artigos em texto completo. Esses catálogos podem ser enriquecidos com *links*, classificados por áreas do CNPq, por exemplo, para o universo em expansão das revistas eletrônicas livres na Internet, tornando-se uma opção extremamente barata às assinaturas dos periódicos convencionais, sejam eles impressos ou eletrônicos.

- **Ambiente para submissão eletrônica de contribuições a congressos**

O Centro de Informações Nucleares está no momento desenvolvendo, também dentro do escopo da BDB, o projeto Edição Eletrônica de Anais de Eventos em C&T, que também fará uso das facilidades de submissão eletrônica e de críticas oferecidas pelo *software Eprint*. O projeto prevê a criação de uma biblioteca digital de anais de congressos, além de mecanismos para edição de anais impressos por meio convencional ou/em meio digital.

- **Biblioteca como publicadora Web**

Nesta nova configuração, a biblioteca e as demais unidades de informação podem assumir um papel de grande relevância que é o de publicador *Web* da produção acadêmica de sua comunidade de usuários. O investimento para isso é pequeno: é necessário equipá-la com servidores compatíveis com o Protocolo OAI, que podem ser microcomputadores PC-Intel ligados à Internet, rodando o conjunto de *software* que formam o *Eprints*. É ainda necessário o estabelecimento de políticas de submissão e metodologias de gestão dos arquivos digitais. Um dado importante é que as unidades de informação que possuem estes servidores serão interoperáveis entre si, podendo trocar dados via processos de *harvesting* automático ou via processos tradicionais. Isto traz a perspectiva de outros serviços descritos a seguir.

- **Busca unificada**

Um bom exemplo é o Arc – *Cross Archive Searching Service*[24] – primeiro serviço a proporcionar acesso integrado a diversos arquivos eletrônicos – 30 em março de 2001 – a partir de uma única interface. Outro exemplo é o site My.OAI[25].

- **Descoberta de recursos via perfis de DSI**

É possível implementar via processos de *harvesting*, a descoberta de recursos informacionais em fontes específicas, previamente selecionadas. As informações capturadas podem ser filtradas segundo perfis de DSI, funcionalidade já implementada pelo *software Eprints*, e enviadas via e-mail para indivíduos, comunidades, departamentos, programas etc., dependendo do nível de capilaridade que se queira alcançar.

24 Disponível em: http://arc.cs.odu.edu/. Acesso em: 1 dez. 2002.

25 Disponível em: http:// www.myoai.com/. Acesso em: 1 dez. 2002.

Outros exemplos de serviços baseados em "*harvesting*" de metadados usando o OAI PMH, implementados pela DLF – *Digital Library Federation*[26] – muitos já em caráter operacional, estão relatados em Waters (2001).

5 Conclusões

Só aos poucos a comunidade envolvida com publicações eletrônicas e bibliotecas digitais vem conhecendo e avaliando o grande potencial de interoperabilidade do OAI PMH. Nestas comunidades, o protocolo vem tendo grande aceitação, devido provavelmente à sua grande simplicidade conceitual e tecnológica. Já existem propostas de extensões para este protocolo, endereçando as interfaces entre módulos que comporiam uma biblioteca digital completa (SULEMAN *et al.*, 2001). Maiores detalhes técnicos sobre o OAI PMH podem ser encontrados em Lynch (2001) e Warner (2001).

A divisão de trabalho entre provedores de dados e provedores de serviços prevista pelo OAI PMH, possibilitando a criação de novos serviços de valor agregado, abre grandes possibilidades. Nada impede que uma biblioteca digital seja simultaneamente provedor de dados e de serviços. Na era dos documentos em papel, cada cópia deste documento, armazenada em uma biblioteca, era descrita e indexada, produzindo registros que, organizados em fichários manuais ou catálogos automatizados, eram mecanismos para viabilizar a recuperação destes documentos. Portanto, a criação de metadados era separada da produção do documento e repetida em diferentes bibliotecas. A criação de metadados simultaneamente à submissão de um documento digital para publicação em um arquivo eletrônico, isto é, o encapsulamento do documento e seus metadados, realiza na prática a proposta antiga na área de biblioteconomia, da catalogação na fonte. Esta se torna viável pelo uso de um conjunto de metadados – o Dublin Core – tão simples "a ponto de poder ser utilizado pelo próprio autor" (WEIBEL, 1995). A biblioteca fica dispensada de realizar o tratamento técnico deste documento, geralmente tão oneroso. O documento tratado na fonte se torna imediatamente insumo para diferentes serviços, possibilitando seu reuso. A criação de documentos digitais formatados em XML (*Extensible Markup Language*, 2000) avança ainda mais nesta direção: conteúdo e metadados vão compor um todo único, já que a linguagem XML permite a marcação das partes que compõem um documento digital, permitindo que um programa possa identificar que porção do conteúdo do documento constitui o título, identifica o autor, o assunto, as hipóteses, a conclusão etc. (BERNERS-LEE; HENDLER; LASSILA, 2001).

A possibilidade de coleta automática de metadados viabilizada pelo OAI PMH é a chave para uma nova prática de cooperação entre bibliotecas, desonerando os

26 Disponível em: http://www.diglib.org . Acesso em: 22 ago. 2021.

cooperantes do pesado ônus administrativo de gerenciar o envio de lotes, correções, atualizações, operações estas que permitem, em um esquema de cooperação tão conhecido dos sistemas de informação brasileiros, manter um catálogo coletivo. *Software* para tornar provedores de dados e de serviços compatíveis com o OAI PMH está disponível gratuitamente na Internet, no site da O*pen Archives Initiative.*

No contexto das bibliotecas especializadas, atividades como seleção e aquisição podem tornar-se quase automáticas, realizando periodicamente a coleta automática de metadados dos arquivos eletrônicos suscetíveis de atender à determinada comunidade. Metadados dos documentos digitais coletados já vêm tratados. Uma bi blioteca digital não precisa mais ter a posse (LANCASTER, 1994) dos documentos que compõem o seu acervo, mas simplesmente manter metadados dos mesmos apontando para o texto completo, armazenado em um outro servidor. O reuso de metadados que também foi sempre perseguido pelas bibliotecas, via sistemas e bancos de catalogação cooperativa, torna-se uma possibilidade real.

O antigo ciclo de comunicação científica, incluindo pesquisador, editor, serviços de informação e biblioteca nas etapas de produção de conhecimentos, registro deste conhecimento, publicação, seleção, aquisição, descrição, armazenamento e disseminação, fica profundamente alterado pelo aporte das tecnologias de informação. Estas etapas se integram, estreitam-se. Produzir textos digitais já implica, praticamente, publicá-los, descrevê-los e disponibilizá-los para disseminação imediatamente.

A existência de documentos livres com os resultados de pesquisas de ponta em diversas áreas de C&T disponíveis na Internet configura uma oportunidade altamente significativa e até então inédita para a ciência dos países em desenvolvimento como o Brasil, conforme analisa Chan e Kirsop (2001). Constitui um mecanismo potencial de democratização no acesso aos resultados de pesquisas e do conhecimento em geral. Esta oportunidade e suas potencialidades não podem passar despercebidas pela comunidade acadêmica brasileira, nem pelos gestores e planejadores de C&T, nem pelos gestores dos sistemas de ICT.

Referências

BERGMAN, Michael K. The deep web: surface hidden value. **Journal of Electronic Publishing**, v. 7, n. 1, Aug. 2001. Disponível em: http://www.press.umich.edu/jep/07-01/bergman.html. Acesso em: 6 ago. 2021.

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. The semantic web. **Scientific American** , New York, n. 5, May 2001. Disponível em: https://www.sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf. Acesso em: 6 ago. 2021.

CHAN, Leslie; KIRSOP, Barbara. Open archiving opportunities for developing countries: towards equitable distribution of global knowledge. **Ariadne**, v.30, Dec. 2001. Disponível em: http://www.ariadne.ac.uk/issue/30/oai-chan/. Acesso em: 6 ago. 2021.

CHATAIGNIER, Maria Cecília Pragana; SILVA, Margareth Prevot. Biblioteca digital: a experiência do IMPA. **Ciência da Informação**, Brasília, v. 30, n. 3, p. 7-12, set./dez. 2001. Disponível em: http://revista.ibict.br/ciinf/article/view/907. Acesso em: 6 ago. 2021.

DAY, Michael. The scholarly journal in trasition and the PubMed Central proposal. **Ariadne**, v. 21, Sept. 1999. Disponível em: http://www.ariadne.ac.uk/issue21/pubmed/. Acesso em: 6 ago. 2021.

EXTENSIBLE markup language. **World Wide Web Consortium**, 2000. Disponível em: http://www.w3.org/XML/. Acesso em: 6 ago. 2021.

GINSPARG, P. Winners and losers in the global research village. In: CONFERENCE ON ELECTRONIC PUBLISHING IN SCIENCE, 1996, Paris. **Proceedings**... Disponível em: https://www.tandfonline.com/doi/pdf/10.1300/J123v30n03_13. Acesso em: 6 ago. 2021.

GONZÁLEZ DE GÓMEZ, Maria Nélida. O papel do conhecimento e da informação nas formações políticas ocidentais. **Ciência da Informação**, Brasília, v.16, n. 2, p. 157-167. jul./dez. 1987. Disponível em: http://revista.ibict.br/ciinf/article/view/259/259. Acesso em: 6 ago. 2021.

HARNARD, Stevan. The self-archiving initiative: nature web debates. **Nature**, v**.**410, 1024–1025, 2001. Disponível em: https://www.nature.com/articles/nature28061.pdf. Acesso em: 6 ago. 2021.

LANCASTER, F. W. Ameaça ou oportunidade? O futuro dos serviços bibliotecários à luz das inovações tecnológicas. **Revista de Biblioteconomia da UFMG**, v. 23, n. 1, p. 7-27, jan./jul. 1994.

LAWRENCE, Steve. **Free online availability substantially increases a paper's impact**: nature web debates. Disponível em: https://www.nature.com/articles/35079151.pdf. Acesso em: 6 ago. 2021.

LE COADIC, Yves-Francois. **A ciência da informação**. Brasília: Briquet de Lemos, 1996.

LYNCH, Clyfford. Metadata harvesting and the open archives initiative. **ARL Bimonthly Report**, n. 217, p.1-9, Aug. 2001. Disponível em: https://www.cni.org/wp-content/uploads/2001/08/Metadata-Harvesting-and-the-Open-Archives-Initiative.pdf. Acesso em: 6 ago. 2021.

MASIERO, Paulo Cesar *et al.* A biblioteca digital de teses e dissertações da Universidade de São Paulo. **Ciência da Informação**, Brasília, v. 30, n. 3, p. 34-41,

set./dez. 2001. Disponível em: http://revista.ibict.br/ciinf/article/view/910/947. Acesso em: 6 ago. 2021.
MARCONDES, Carlos Henrique. Tecnologias de informação e impacto na formação do profissional de informação. **Transinformação**, Campinas, v. 11, n. 3, p. 189-194, 1999.
MARCONDES, Carlos Henrique. Representação e economia da informação. **Ciência da Informação**, Brasília, v. 30, n. 1, p. 61-70, jan./abr. 2001. Disponível em: http://revista.ibict.br/ciinf/article/view/939/976. Acesso em: 6 ago. 2021.
MARCONDES, Carlos Henrique; SAYÃO, Luis Fernando. Integração e interoperabilidade no acesso a recursos informacionais em C&T: a proposta da Biblioteca Digital Brasileira. **Ciência da Informação**, Brasília, v. 30, n. 3, p. 24-33, set./dez. 2001. Disponível em: http://revista.ibict.br/ciinf/article/view/909/946. Acesso em: 6 ago. 2021.
MILLER, Paul. Interoperability. What is it and why should I want it? **Ariadne**, v. 24, Jun. 2000. Disponível em: http://www.ariadne.ac.uk/issue/24/interoperability/. Acesso em: 6 ago. 2021.
NEC RESEARCH INSTITUTE. **Finding a needle in the web**. Boletim Edupage, Apr. 1998.
PACKER, Abel *et al*. SciELO: uma metodologia para publicação eletrônica. **Ciência da Informação**, Brasília, v. 27, n. 2, 1998. Disponível em: https://www.scielo.br/j/ci/a/XhRCDr87m5VTswK5WtNdYzL/?format=pdf&lang=pt. Acesso em: 6 ago. 2021.
ROBREDO, Jaime; CUNHA, Murilo Bastos. **Documentação de hoje e de amanhã**: uma abordagem automatizada da biblioteconomia e dos sistemas de informação. 2. ed. São Paulo: Global, 1994.
SAYÃO, Luis Fernando. Bases de dados: a metáfora da memória científica. **Ciência da Informação**, Brasília, v. 25, n .3, 1996. Disponível em: http://revista.ibict.br/ciinf/article/view/629/633. Acesso em: 6 ago. 2021.
SENA, Natália Kneipp. Open archives: caminho alternativo para a comunicação científica. **Ciência da Informação**, Brasília, v. 29, n. 3, p. 71-78, set./dez. 2000. Disponível em: https://www.scielo.br/j/ci/a/gcmzNYH3R8FbKHwMRdGh7gJ/?lang=pt. Acesso em: 6 ago. 2021.
SHNEIDERMAN, Ben; BYRD, Don; CROFT, W. Bruce. Clarifying search: a user-interface framework for text searches. **Dlib Magazine**, Jan. 1997. Disponível em: https://www.dlib.org/dlib/january97/retrieval/01shneiderman.html. Acesso em: 6 ago. 2021.
SOMPEL, Herbert van de; LAGOZE, Carl. The Santa Fe Convention of the Open Archives Initiative. **Dlib Magazine**, v. 6, n. 2, Feb. 2000. Disponível em: http://

www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html. Acesso em: 6 ago. 2021.
SULEMAN, Hussein *et al.* Networked digital library of theses and dissertations: bringing the gap for global access - part 1: mission and progress. **D-Lib Magazine**, v. 7, n. 9, Sept. 2001. Disponível em: https://www.dlib.org/dlib/september01/suleman/09suleman-pt1.html. Acesso em: 6 ago. 2021.
TROLL, Denise; MOEN, Bill. **Report to the DLF on the Z39.50 Implementers' Group.** Acesso em Disponível em: https://old.diglib.org/architectures/zig0012.htm. Acesso em: 6 ago. 2021.
WARNER, Simeon. Exposing and harvesting metadata using the OAI Metadata Harvesting Protocol. **Tutorial, High Energy Physics Libraries Webzine**, v. 4, June 2001. Disponível em: https://webzine.web.cern.ch/4/papers/3/. Acesso em: 6 ago. 2021.
WATERS, Donald J. The metadata harvesting initiative of the Mellon Foudation. **ARL Bimonthly Report**, n. 17, Aug. 2001. Disponível em: www.alr.org/newsltr/waters.html. Acesso em: 15 fev. 2002.
WEIBEL, Stuart. Metadata: the foundations of resource description. **D-Lib Magazine**, July 1995. Disponível em: https://www.dlib.org/dlib/July95/07weibel.html. Acesso em: 6 ago. 2021.
ZIMAN, John. **Conhecimento público**. Belo Horizonte: Itatiaia/São Paulo: USP, 1979.

Integração e interoperabilidade no acesso a recursos informacionais eletrônicos em C&T: a proposta da Biblioteca Digital Brasileira

Carlos Henrique Marcondes[1] e Luís Fernando Sayão[2]

1 Introdução

"Um laboratório sem uma biblioteca é como se fosse um animal descorticado: as atividades motoras continuam a funcionar, mas falta a coordenação da memória e da vontade"
(ZIMAN, 1979, p. 115).

Há muito tempo está constatada a relação entre desenvolvimento econômico e social e o estágio de desenvolvimento da ciência e tecnologia de um país. À ciência e tecnologia está reservado um papel fundamental na luta pelo desenvolvimento da sociedade brasileira. Informação é insumo fundamental para o desenvolvimento da ciência. É em torno do problema da otimização dos fluxos e da transferência de informação científica que surge, na segunda metade do século XX, a ciência da informação (PINHEIRO; LOUREIRO, 1995).

A convergência e o uso integrado das tecnologias de comunicação, de computação e de conteúdos em formato digital, cujo paradigma é a Internet, tem contribuído nos anos recentes para criar um novo ambiente de acesso, disseminação, cooperação e promoção do conhecimento em uma escala global. Estamos em meio a este processo, cujas consequências ainda não podemos avaliar completamente. Novos suportes de conhecimento, que não guardam similares com os materiais impressos em papel, estão sendo inventados a cada dia.

Do ponto de vista da informação como subsídio às atividades acadêmicas e em C&T, a Internet vem proporcionar facilidades que extrapolam o conceito

1 Doutor em ciência da informação. Universidade Federal Fluminense. marcon@vm.uff.br.
2 Doutor em ciência da informação. Comissão Nacional de Energia Nuclear. luis.sayao@cnen.gov.br.

tradicional de informação bibliográfica baseada em documentos, como artigos de periódico, trabalhos em congressos, teses etc. Novos recursos informacionais estão à disposição da comunidade de pesquisa além desses tradicionais, agora em versão eletrônica, como documentos multimídia, listas de discussão, fóruns eletrônicos, conferências em linha, imagens (de satélites, de microscópios, em tempo real), modelos animados, bancos de *preprints* eletrônicos, os *e-prints* etc. Estes recursos tanto servem de subsídio à pesquisa quanto de canais de comunicação e publicação dos resultados e de garantia de primado e originalidade intelectuais dos mesmos.

Mais que somente recursos informacionais, os novos recursos disponíveis via Internet são acima de tudo novas ferramentas cognitivas, no sentido emprestado a elas por Pierre Lévy (1993), capazes de abrir novas possibilidades cognitivas e intelectuais que extrapolam em muito aquelas oferecidas por documentos em papel, de leitura linear. Para muitos autores, a Internet representa, neste sentido, uma mudança de paradigma comparável à invenção da imprensa por Gutemberg. Esta mudança de paradigma se faz sentir também no aspecto da comunicação científica. A Internet é um mecanismo de comunicação de alcance mundial, instantâneo, interativo e multidirecional: qualquer um pode publicar nela, o que foi publicado é imediatamente acessível, o autor pode receber um retorno e avaliação imediatos sobre o que publicou, de qualquer lugar. Um autor acadêmico almeja a máxima divulgação para seus trabalhos, para que os resultados de sua pesquisa tenham o maior impacto possível sobre as pesquisas de seus pares e sobre outras publicações. Estudos recentes confirmam que as publicações eletrônicas são muito mais citadas que as publicações em papel: "*The mean number of citations to offline articles is 2,74, and the mean number of citations of online articles is 7.03, an increase of 157%*" (lawrence, 2001). Desenvolver mecanismos de publicação eletrônica para a comunidade acadêmica brasileira, aumentando sua visibilidade, torna-se, portanto, uma questão essencial para o desenvolvimento e maturidade da pesquisa científica brasileira.

Dessa forma, os paradigmas de comunicação científica, tendo por base o periódico científico em papel, com seu esquema de revisão por pares e o monopólio das grandes editoras científicas, vêm sofrendo grande impacto com o surgimento da Internet e grande questionamento por parte da comunidade científica de todo o mundo. Desde o surgimento do primeiro arquivo eletrônico de *preprints*, ou *eprints*, o ArXiv, no *Los Alamos National Laboratory*, criado em 1991 pelo físico Paul Ginsparg (ginsparg, 1996), que a própria comunidade científica internacional oferece uma alternativa prática para a publicação de seus trabalhos. Uma lista que permite dimensionar a amplitude dos arquivos eletrônicos por todo o mundo

pode ser encontrada em http://www.osti.gov/eprints/ppnbrowse.html. A alternativa dos *e-prints* vem também se articulando na chamada *Open Archives Initiative*[3].

Estas transformações têm exercido profunda influência sobre a concepção e funcionamento dos sistemas de informação automatizados, especialmente aqueles voltados para as atividades de pesquisa. O rompimento de barreiras tecnológicas importantes, experimentadas na última década, permitiram o surgimento de um novo patamar para esses sistemas: antes orientados basicamente para recuperação de referências bibliográficas em bases de dados isoladas e textos em papel, voltam-se hoje para a recuperação distribuída de objetos digitais – textos completos, imagens em movimento, som etc. –, estabelecendo como palavras de ordem a publicação na Internet e a interoperabilidade entre fontes de informação heterogêneas e globalmente distribuídas.

Com o projeto da Biblioteca Digital Brasileira, o IBICT quer abrir a possibilidade, fomentar e fornecer meios para que a comunidade brasileira de C&T possa publicar seus trabalhos de forma rotineira, diretamente na rede, aumentando com isso sua visibilidade nacional e internacional, otimizando o fluxo da comunicação científica e reduzindo o ciclo de geração de novos conhecimentos.

Por outro lado, somente a disponibilidade de textos brasileiros em C&T Online não teria grande impacto sobre a comunicação científica e a ciência no país sem a existência de serviços de informação que viabilizem o acesso de forma fácil a estes conteúdos. O país também tem acumulado experiências bastante significativas, embora isoladas, na criação de bibliotecas digitais e repositórios de informações na rede. À medida que experiências brasileiras neste sentido se multiplicam, como o Prossiga, o Scielo, o repositório de teses da USP, o arquivo de *e-prints* do Impa[4] etc., disponibilizando de forma crescente recursos informacionais em texto completo na Web, fica patente, para as organizações brasileiras que trabalham com sistemas de informação para C&T, a importância da questão da interoperabilidade entre bibliotecas digitais e outros recursos informacionais digitais: como consultar, de uma única vez, todas estas fontes de forma integrada e transparente, com o mínimo de esforço, com a máxima rapidez, e obter resultados consolidados?

São assim dois os objetivos fundamentais do projeto BDB: a) fomentar a publicação de recursos informacionais de interesse para C&T na rede, propiciando à comunidade científica brasileira meios para publicar diretamente na Web, dando maior visibilidade à produção brasileira em C&T, tanto nacionalmente, quanto internacionalmente; b) viabilizar o acesso rápido e integrado a estes recursos, facili-

3 Disponível em: https://www.openarchives.org/. Acesso em: 22 ago. 2021.

4 Disponível em: https://impa.br/biblioteca/servidores-de-pre-publicacoes/. Acesso em: 23 ago. 2021.

tando a descoberta na Internet de recursos informacionais brasileiros de interesse para a ciência e tecnologia, de forma integrada e, dessa forma, encurtando o ciclo de comunicação científica entre pares nas comunidades brasileiras de C&T.

2 O problema da interoperabilidade

Um aspecto problemático da cultura de nosso tempo é o assim chamado fenômeno da explosão informacional, a grande quantidade de informações produzidas e disponibilizadas por diferentes atividades sociais, dificultando sua identificação, acesso e utilização. Na emergência da sociedade da informação, o valor desta como insumo para qualquer atividade, seja ela uma decisão econômica, um processo cultural ou de ensino/aprendizagem, uma pesquisa científica ou tecnológica, está relacionado diretamente ao seu potencial de orientar de forma econômica o dispêndio de energia para a realização desta atividade. Para que possa realizar todo este potencial, a informação relevante para um dado problema deve estar disponível no tempo certo. De nada adianta a informação existir, se quem dela necessita não sabe da sua existência ou se ela não puder ser encontrada.

Esta situação assume proporções alarmantes com o surgimento da Internet. Uma notícia divulgada no Boletim Edupage em português, de 05/04/98, publicado pela Rede Nacional de Ensino e Pesquisa (RNP), levanta o problema da busca de informações na Internet e comenta os resultados de um estudo sobre o desempenho dos assim chamados "mecanismos de busca":

"ACHANDO UMA AGULHA (OU 7.079 PÁGINAS EM UMA AGULHA) NA WEB

Um estudo realizado pelo *NEC Research Institute* afirma que a Internet explodiu para mais de 320 milhões de páginas na Web, uma estimativa que não inclui milhões de páginas com acesso protegidas por senhas ou "muros de pesquisa" que bloqueiam acesso a *browsers* ou mecanismos de busca. O estudo indica que a pesquisa do mecanismo de busca *HotBot* tem o índice mais abrangente da Web, mas, ainda assim, cobre apenas 34% das páginas indexáveis. A cobertura de alguns dos outros mecanismos de busca inclui: AltaVista (28%); Northern Light (20%); Excite (14%); Lycos (3%)."

Uma novidade em termos de mecanismos de busca que parece alentadora são os projetos CLEVER e GOOGLE[5], com suas propostas de ordenamento e priorização (*ranking*) dos resultados de uma busca, tendo por base os sites mais referenciados por links a partir de outros (CLEVER, 1999).

5 Disponível em: http://www.google.com. Acesso em: 23 ago. 2021.

A enorme quantidade de informação armazenada e disponibilizada via Internet torna cada vez mais crítico o problema da identificação de informação relevante, assim chamada *information discovery*. Diferentes estratégias para fazer frente à explosão informacional trazida pela Internet podem hoje ser divisadas, como os mecanismos de busca gerais (AltaVista, Excite, Lycos, Infoseek, Yahoo e outros), os localizadores de informações especializados, como o GILS[6] ou portais temáticos como o SIGNPOST[7] americano, o OMNI[8] e o SOSIG[9] ingleses, o PROSSIGA – Comunicação e Informação para a Pesquisa[10] – ou LIS – Localizador de informações em Saúde[11] – no Brasil.

Ambas as alternativas, os mecanismos de busca gerais e os portais temáticos oferecem soluções parciais para a localização de informações na Internet, principalmente as de interesse para C&T. As deficiências dos mecanismos de busca são já bastante conhecidas e discutidas na literatura (SHNEIDERMAN, 1997). Entre as principais, pode-se citar as seguintes: baixa qualidade da indexação, por ser feita automaticamente, que resulta em grande quantidade de informações recuperadas, a maioria sem relevância (em termos de recuperação de informação, oferecem alta revocação, mas baixa precisão); cobertura parcial da Internet; as ferramentas de busca não são especializadas; indexam páginas HTML isoladas, e não recursos; além disso, grande quantidade de informações disponíveis na Internet estão sob a forma de registros contidos em bases de dados, que ficam assim "escondidas"; estes registros são acessados somente por meio das interfaces destas bases de dados, o que pressupõe uma interação entre um usuário humano com a base de dados e, portanto, ficam inacessíveis aos programas robôs.

Por sua vez, as bibliotecas digitais e os portais temáticos isolados resolvem somente em parte o problema do acesso a recursos informacionais de interesse para C&T publicados na rede: continuam limitadas ao "seu" acervo. Descobrir, avaliar, tratar e indexar estes recursos por profissionais de informação é caro e lento. Estes recursos estão sendo criados em número crescente, armazenados em diferentes servidores isolados, operados por interfaces de busca diferentes, o que obriga um usuário a uma dispendiosa busca, site a site, para encontrar informações relevantes.

Do ponto de vista de um usuário acadêmico ou pesquisador, o interessante e confortável seria poder submeter sua necessidade de informação e interagir com

6 Disponível em: http://www.usgs.gov/gils/. Acesso em: 1 dez. 2001.
7 Disponível em: http://www.signpost.org. Acesso em: 1 dez. 2001.
8 Disponível em: http://www.omni.ac.uk. Acesso em: 1 dez. 2001.
9 Disponível em: http://www.sosig.ac.uk. Acesso em: 1 dez. 2001.
10 Disponível em: http://www.prossiga.br. Acesso em: 1 dez. 2001.
11 Disponível em: http://www.bireme.br. Acesso em: 1 dez. 2001.

uma única interface e ter retornadas informações de diferentes fontes, de forma consolidada. Este é um tema que, sob diferentes denominações, está sendo cada vez mais discutido: *digital libraries federation e distributed archives* (LIU *et al.*, 2001), *confederated digital libraries* (LEINER, 1998), *distributed subject gateways* (IMESH, 1999), *networked digital library* (DAVIS, 1995), *multiple information sources* (PAEPCKE, 2000) etc.

A questão começa a ser levantada na *An Intenational research agenda for digital libraries*, de 1998 – uma agenda de pesquisa conjunta da NSF (EUA) e União Européia – em três grupos de trabalho temáticos: *global resource discovery*, *interoperability* e *metadata*. Hoje é endereçada diretamente por diferentes iniciativas de pesquisa, como a *Joint NSF* – JISC *International Digital Libraries Research Programme*, como o consórcio Imesh[12], o *Scout Project*, e por iniciativas práticas como *OpenArchives Initiative, Arc - Cross Archive Searching Service*, NCSTRL (Univ. Cornell, EUA), NDLTD (*Virginia Tech, University*, EUA), *Digital Library Federation* (consórcio de bibliotecas digitais americanas), ROADS (UKOLN JISC, Inglaterra). As diferentes denominações sob as quais o tema aparece na literatura convergem para os conceitos de integração e interoperabilidade entre bibliotecas digitais, que consistiria na possibilidade de um usuário realizar buscas a recursos informacionais heterogêneos, armazenados em diferentes servidores na rede, utilizando-se de uma interface única sem tomar conhecimento de onde nem como estes recursos estão armazenados.

Hoje, no cenário mundial, identificam-se várias alternativas de interoperabilidade e acesso integrado a recursos informacionais heterogêneos publicados na rede. Estas podem ser agrupadas basicamente em duas alternativas, embora ainda não tenha se fixado uma nomenclatura amplamente aceita: buscas distribuídas a diferentes servidores e busca em uma base de metadados centralizada. Em ambas as alternativas, o usuário interage com uma única interface Web, de onde é submetida a busca.

Na primeira alternativa, a interface de busca distribui a consulta (*broadcast search*) a diferentes sites, segundo um protocolo padrão, identificados pela interface como capazes de fornecer respostas satisfatórias, e os resultados são consolidados e integrados. Exemplo típico desta alternativa é o conhecido protocolo Z39.50, usado para proporcionar interoperabilidade entre catálogos automatizados de bibliotecas. Esta alternativa apresenta as seguintes vantagens: novos provedores de dados podem ser acrescentados, desde que sejam aderentes ao padrão ou padrões utilizados, com reconfiguração mínima. Como desvantagens, pode-se apontar que provedores de dados precisam rodar *software* servidor do protocolo padrão para

12 Disponível em: http://www.desire.org/html/subjectgateways/community/imesh/. Acesso em: 1 dez. 2001.

serem consultáveis. Alguns destes *softwares* consomem muitos recursos por parte dos provedores de dados, como no caso do Z39.50 (TROLL; MOEN, 2001). Em alguns casos, são necessários servidores especializados, como os servidores de índices, que roteiam as consultas para os servidores capazes de atendê-las. Alguns dos padrões tecnológicos utilizados são os seguintes: Z39.50 (ISO/NISO), Whois++, LDAP, CIP, SDLIP, DIENST. Esta alternativa é utilizada nos seguintes sistemas: NCSTRL (*University of. Cornell*, EUA), NDLTD – *Networked Digital Library of Theses and Dissertations - federated search* (Powel, 1998), *California Digital Library*[13], *Berkeley Environmental Digital Library*, EUA, ROADS, ISAAC/SCOUT *Project* (*University of Stanford*, EUA).

Na segunda alternativa, metadados referentes a documentos eletrônicos são coletados periodicamente, alimentando uma base comum de metadados sobre a qual são realizadas as buscas. Este esquema é bastante conhecido da colaboração/cooperação entre as instituições participantes para manutenção do Catálogo Coletivo/base de metadados centralizada. Dentro desta alternativa, variam os esquemas de centralização destes metadados. O esquema do envio de metadados por parte das instituições cooperantes é mais tradicional e largamente conhecido pela comunidade de informação, inclusive a brasileira, em sistemas/bases de dados como LILACS/ BIREME, SITE/IBICT, INIS/CIN. Este esquema apresenta as seguintes vantagens: novos provedores de dados podem ser acrescentados, desde que sejam aderentes ao padrão ou padrões utilizados; melhor desempenho em função da consulta ser na base de metadados local do provedor de serviços. Como desvantagens este esquema apresenta: manutenção pelo provedor de dados da base comum de metadados; grande ônus administrativo e gerencial por parte do provedor de serviços para sincronizar o envio dos dados por parte dos provedores de dados e processá-los para incluí-los na base comum de metadados; necessidade de sincronização entre os dados armazenados nos provedores de dados e os metadados coletados pelo provedor de serviços.

O esquema de coleta automática de metadados (*harvesting*) é mais recente: metadados de diversos provedores de informação tornam-se "visíveis" através de protocolos padronizados e são coletados automaticamente de forma periódica e armazenado em um *data warehousing*, ou base centralizada de metadados, onde são efetuadas as buscas de forma integrada. Vantagens deste esquema são as seguintes: novos provedores de dados podem ser acrescentados, desde que sejam aderentes ao padrão utilizado; melhor desempenho em função de a consulta ser na base de metadados local do provedor de serviços. Desvantagens: manutenção pelo provedor de dados da base comum de metadados; necessidade de sincronização entre

13 Disponível em: http://www.cdlib.org/. Acesso em: 23 ago. 2021.

os dados armazenados nos provedores de dados e os metadados coletados pelo provedor de serviços (LIU *et al.*, 2001). Padrões utilizados: *Open Archives Harvest Protocol*, *Open Archives Metadata Set,* Dublin Core, XML. Exemplos são as experiências do sistema MARIAN (*Virginia Technical University*), do portal da NDLTD (SULEMAN, 2001), da *Open Archives Initiative*, que, através do protocolo *OAI harvest protocol*, permite colheita (*harvesting*) de metadados, do Arc – *Cross Archive Searching Service*[14], primeiro serviço a proporcionar acesso integrado a diversos arquivos eletrônicos.

3 Modelo de interoperabilidade da BDB

O projeto de implantação da BDB no país pressupõe forte ação de integração, liderada pelo IBICT, dos mais importantes provedores de conteúdos e de serviços de informação para C&T do país. Esta integração se dará em torno das questões prioritárias para a BDB, que são a publicação e a disponibilidade de textos completos e outros objetos digitais na Internet e a interoperabilidade entre os diversos sistemas/serviços de informação participantes através de um portal único de acesso, preservando-se a independência e peculiaridades de cada sistema/serviço participante. Para conseguir cooperação dos eventuais provedores de dados, o conjunto de metadados, a configuração, os padrões e procedimentos por parte dos provedores de dados para garantir interoperabilidade com a BDB deverão ser os mais simples e menos onerosos para os provedores de dados, garantindo sua máxima independência.

Apesar do avanço acelerado das tecnologias Web e das tecnologias de informação e comunicação, o projeto da BDB prevê a utilização de tecnologias consolidadas, cujo grau de estabilidade e confiabilidade tenham sido comprovados no país e no exterior, e que, prioritariamente, tenham sido aplicados nas principais experiências internacionais análogas à da proposta da BDB. Além do mais, essas tecnologias devem ser passíveis de serem implantadas, mantidas e, quando necessário, alteradas pelo corpo técnico do IBICT. Sempre que possível, serão adotadas tecnologias abertas, não proprietárias, que permitam garantir o grau de interoperabilidade desejável entre os diversos partícipes da BDB e que possam facilmente ser repassadas aos parceiros do Projeto. Esta opção tem como objetivo a disseminação no país de um corpo de protocolos e padrões que possam ser adotados pelos futuros integrantes da BDB e por outros sistemas que queiram aderir a sistemas/redes internacionais.

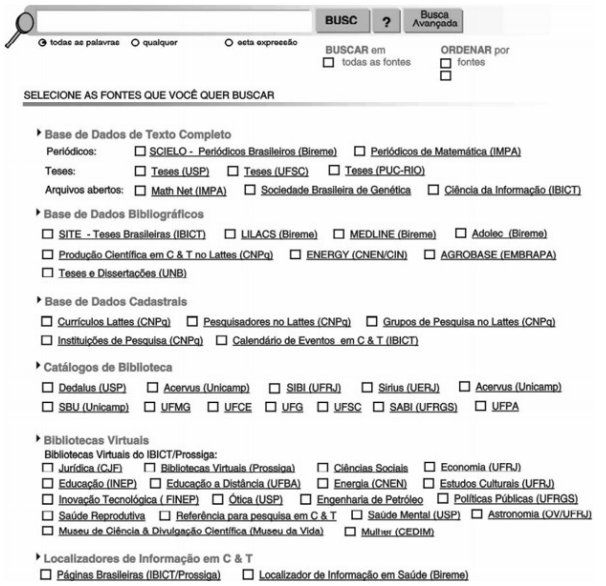
No intuito de disseminar mais facilmente as tecnologias de publicação na Internet entre as diversas comunidades de conhecimento, sempre que possível serão

14 Disponível em: http://www.arc.cs.odu.edu. Acesso em: 1 dez. 2001.

adotados *software* de domínio público, preservada a qualidade, documentação e mantenabilidade dos mesmos. Isto se deve à constatação de que há uma diversidade surpreendente de *software* livres, confiáveis e de qualidade que estão sendo adotados por instituições importantes na área de informação.

O modelo de interoperabilidade proposto para a BDB aproxima-se bastante dos modelos do portal da NDLTD (SULEMAN, 2001) e do Arc – *Cross Archive Searching Service15*. Ambos os sistemas fazem *harvesting* de metadados de provedores de dados, alimentando uma base de dados central de metadados. O portal da BDB na Internet será a materialização da Biblioteca Digital Brasileira em C&T. Trata-se de um site que, através de diferentes mecanismos de interoperabilidade, possibilitará ao pesquisador acesso unificado e integrado a diferentes recursos informacionais brasileiros de interesse para C&T, heterogêneos e distribuídos, sem a necessidade de navegar e consultar cada recurso individualmente. A figura 1 dá uma ideia da proposta do portal da BDB.

Figura 1- Proposta de interface de busca heterogênea para a BDB.



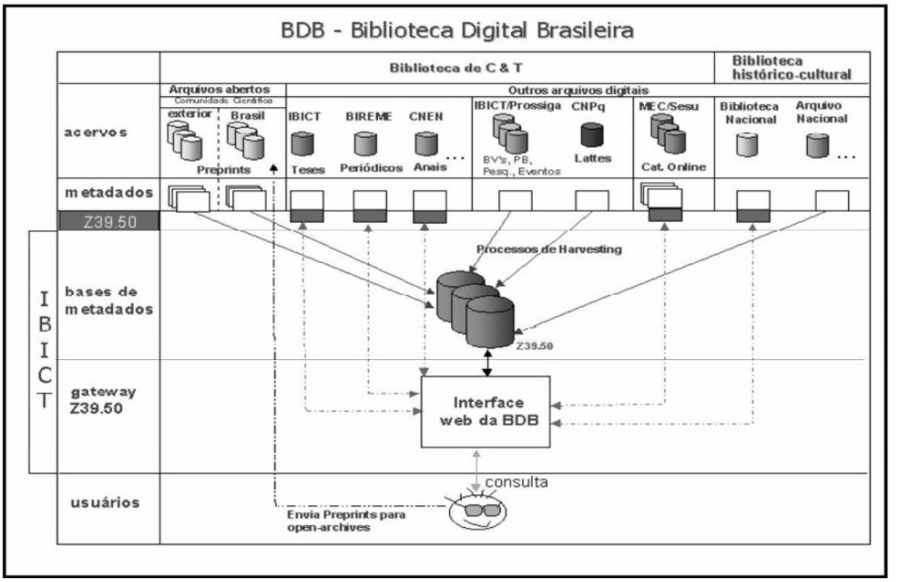
Fonte :os autores.

Entre estes recursos informacionais se incluirão, já na versão inicial do portal, periódicos eletrônicos brasileiros que fazem parte do portal Scielo, mantido pela

15 Disponível em: http://www.arc.cs.odu.edu. Acesso em: 1 dez. 2001.

Bireme, anais eletrônicos de eventos brasileiros em C&T a serem disponibilizados pela CNEN/CIN, bancos de teses eletrônicas hoje já existentes na USP, Unicamp, UFSC, PUC-Rio, ENSP/Fiocruz, repositórios de *e-prints* brasileiros que começam a ser disseminados na Internet, como o do Impa. Na figura 2, é mostrado o modelo geral de interoperabilidade da BDB.

Figura 2 - Proposta de interface de busca heterogênea para a BDB.



Fonte: os autores

Em um ambiente de publicação e acesso a documentos publicados na rede como o da BDB, a *Open Archives Initiative* reconhece dois atores institucionais fundamentais: provedores de dados e provedores de serviço. Estas definições serão utilizadas aqui para explicitar a solução de interoperabilidade adotada, com base no que é estabelecido na *Open Archives Initiative*:

Provedores de dados: de uma forma ampla, seriam todas as instituições brasileiras que possuem site Internet que disponibiliza documentos eletrônicos em C&T. Eventualmente, este site abriga também um ambiente de submissão/publicação de documentos em texto completo; um autor registra seu documento no site através de um conjunto de metadados e, opcionalmente, armazena aí seu documento em formato eletrônico. O site do provedor de dados provê facilidades de busca para acesso aos documentos nele armazenados e, caso seja aderente ao padrão *Open Archives Harvest Protocol*, permitirá também que os metadados dos documentos

do seu acervo sejam visíveis a um programa de harvest. Exemplos típicos seriam o SCIELO, os diversos arquivos eletrônicos existentes no mundo, como o *CogPrints*[16], ou o arquivo aberto do IMPA[17].

Provedor de serviços: instituições que provêem serviços de valor agregado sobre documentos eletrônicos disponibilizados por um ou mais provedor de dados. Exemplos destes serviços seriam a montagem de bases de dados qualificadas, o acesso unificado a documentos armazenados em diferentes provedores de dados, revisão e avaliação de documentos publicados em um ou mais provedores de dados, linkagem de recursos informacionais. Exemplos típicos seriam o serviço Arc[18] que provê acesso unificado a diferentes arquivos abertos e o objeto deste projeto, a própria BDB.

Dado o papel integrador da proposta da BDB, seu modelo de interoperabilidade se baseia em dois elementos: mecanismos de submissão de consultas, a partir da interface única do portal, aos diferentes recursos informacionais que comporão a BDB e conjunto de metadados que descreverão e fornecerão uma visão unificada dos diferentes conjuntos de documentos.

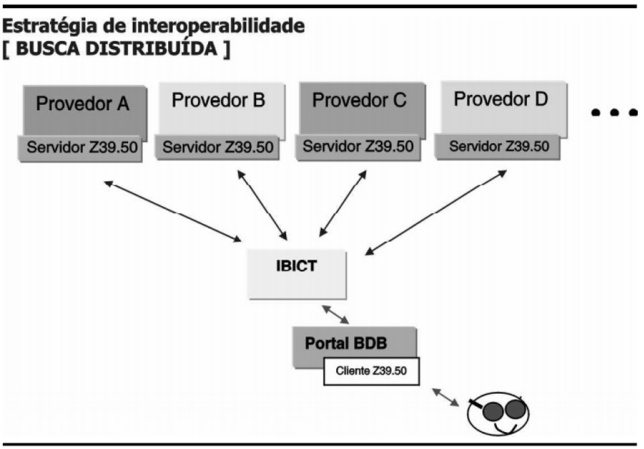
Com relação aos mecanismos de submissão de consultas, a BDB deverá incorporar as principais alternativas tecnológicas discutidas anteriormente, buscas distribuídas e busca em uma base centralizada de metadados obtida mediante coleta automática (*harvesting*). Para isso, o portal da BDB disporá de um programa cliente Z39.50, que lhe permitirá acesso integrado por meio da distribuição de consultas aos catálogos na Internet das principais bibliotecas universitárias do país e estrangeiras servidas pelo protocolo Z39.50. Qualquer outro recurso informacional na Internet que seja servido por este protocolo também poderá ser acessado do portal da BDB, como, por exemplo, um arquivo de *e-prints* ou o Scielo, o qual planeja implementar um servidor Z39.50 para sua base. Esta opção é ilustrada na figura 3.

16 Disponível em: http://cogprints.org/. Acesso em: 23 ago. 2021.

17 Disponível em: https://impa.br/biblioteca/servidores-de-pre-publicacoes/. Acesso em: 23 ago. 2021.

18 Disponível em: http://www.arc.cs.odu.edu. Acesso em: 1 dez. 2001.

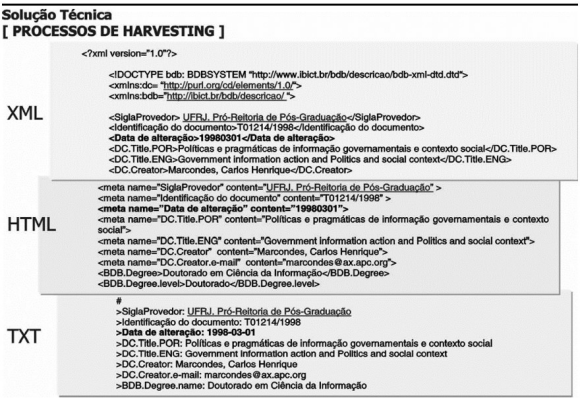
Figura 3 – Estratégia de ineroperabilidade [busca distribuída].



Fonte: os autores

Além de integrar, via consultas distribuídas, os recursos informacionais servidos pelo protocolo Z39.50, a BDB manterá em seu site uma base comum de metadados, obtida pelo processo de *harvesting* dos metadados dos recursos/serviços de informação que não tiverem servidor Z39.50. Estes recursos/serviços serão objeto de coleta automática periódica, usando o *OAI Harvesting protocol* dos provedores de dados compatíveis com este protocolo. Outras soluções foram pensadas de modo a não onerar tecnicamente provedores de dados não compatíveis com este protocolo, como coleta de metadados via FTP em arquivos HTML ou arquivos texto. Os formatos de arquivos passíveis de coleta automática são ilustrados na figura 4.

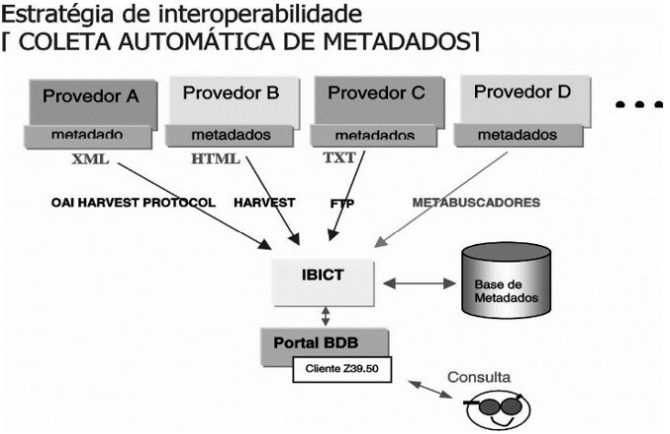
Figura 4 – Solução técnica [processos de *harvesting*].



Fonte: os autores

Estes metadados serão processados e armazenados na base comum de metadados, mantida no site da BDB; esta base, por sua vez, contará também com um servidor protocolo Z39.50, que a tornará acessível a partir do portal da BDB como qualquer outro recurso servido por este protocolo. Arquivos acadêmicos de *preprints* eletrônicos que poderão ser implantados em departamentos de instituições de ensino superior, institutos de pesquisa, sociedades científicas, periódicos eletrônicos, como o próprio Ciência da Informação do IBICT, ou projetos específicos interinstitucionais, como Projeto Genoma, poderão se integrar à BDB segundo esta opção. Ela é ilustrada na figura 5.

Figura 5 – Estratégia de interoperabilidade [coleta automática de metadados]



Fonte: os autores

O segundo elemento do esquema de interoperabilidade da BDB é o conjunto de metadados. Ele deverá contemplar e integrar, em uma descrição unificada, as diferentes tipologias de documentos originários dos diferentes recursos informacionais que comporão a BDB. A referência emergente nesta área, avalizada pela comunidade de informação, é *Dublin Core Element Set*. É resultado de intenso trabalho de discussão e padronização em nível internacional, mantida por um ativo grupo e fórum internacionais, a *Dublin Core Metadata Initiative*, que já realizou diversos encontros; é usado em diferentes sistemas, inclusive *na Open Archives Initiative* e se encontra em pleno desenvolvimento.

O conjunto de metadados Dublin Core é composto de 13 elementos descritivos (DUBLIN CORE, 1999), suporta qualificadores para especificar o significado de formatos como HTML e XML (COX; MILLER; POWELL, 2000; BECKETT; MILLER; BRICKLEY, 2000, 2001). Como é explicitado na própria proposta Dublin Core, o

conjunto de metadados deve ser tão simples e intuitivo a ponto de permitir que o próprio autor descreva seu trabalho. É exatamente assim que funcionam os ambientes de submissão de trabalhos: ao submeter seu trabalho, o autor preenche um formulário com os metadados pertinentes.

Dada a sua característica integradora de diferentes recursos informacionais heterogêneos, que armazenam diferentes tipos de documentos, a BDB usará o conjunto Dublin Core, expandido com qualificadores para suportar características especiais de alguns tipos de documentos. Detalhes da estrutura de metadados da BDB serão objeto de um trabalho posterior.

4 Conclusões

As experiências internacionais em torno da questão das publicações na rede e da interoperabilidade entre de bibliotecas digitais são bastante recentes, contemporâneas mesmo; as soluções para estes problemas são hoje o foco das maiores atenções por parte dos pesquisadores de Ciência da Informação e dos diversos sistemas de informação em C&T. Apesar de toda esta ebulição, está patente que já existe um conjunto de padrões e tecnologias maduros e consolidados que são bases para vários sistemas de informação importantes já em operação pelo mundo. As questões de interoperabilidade endereçadas pelo projeto da BDB são ainda bastante restritas, limitadas a um enfoque tecnológico. A interoperabilidade entre recursos informacionais heterogêneos na Internet tem várias outras dimensões – semântica, política/humana, entre comunidades, internacional, interlinguística (POWELL; FOX, 1998) –, como alerta Miller (2001).

O sucesso do projeto em uma perspectiva a médio e longo prazo vai depender de o IBICT adotar uma nova postura institucional com relação ao acompanhamento das tendências e padrões tecnológicos envolvidos, à pesquisa, ao desenvolvimento e à adaptação de tecnologias. Este quadro é tão amplo que seria impossível ao IBICT acompanhá-lo sozinho. Para fazer frente a este desafio, o IBICT deve assumir o papel de articulador entre diferentes parceiros nacionais e a comunidade acadêmica para um trabalho conjunto em torno de uma agenda de pesquisas que inclua questões de interesse do projeto da BDB. Para isso, o projeto sugere:

> Acompanhamento e articulação com os fóruns internacionais que discutem questões como metadados, interoperabilidade, publicações na rede, comunicação científica via rede, preservação de documentos eletrônicos, direitos autorais em documentos eletrônicos, linkagem de recursos informacionais etc (IBICT, 2001, p.7)

Como *World Wide Web Consortiun*, *Dublin Core Metadata Initiative*, *Open Archives Initiative*, *Digtal Library Federation*. Com o objetivo de acompanhar o desenvolvimento das tecnologias associadas a bibliotecas digitais e o desenrolar das controvérsias sobre estes assuntos e seus desdobramentos, é necessária uma aproximação com as organizações e/ou redes internacionais que operam e/ou desenvolvem experiências com bibliotecas digitais heterogêneas distribuídas. É também de grande importância a participação do IBICT nos principais fóruns internacionais que discutem as questões relacionadas ao tema. Sugere-se (IBICT, 2001, p.16):

- propor uma agenda incluindo os temas de pesquisa citados anteriormente à comunidade de pesquisas brasileira e uma linha de fomento correspondente, envolvendo universidades, institutos de pesquisa e instituições parceiras;
- propor um fórum nacional sobre o tema amplo de bibliotecas digitais que sirva para discussão e troca de experiências;
- propor um programa de formação de quadros.

O projeto da BDB, embora tenha um compromisso pragmático com a prestação de serviços à comunidade acadêmica brasileira, coloca também questões relativas ao planejamento de ICT no país. Desde fins da década de 80, com a Ação Programa de Informação em Ciência e Tecnologia (BRASIL, 1984) e com os PADCTs (BRASIL, 1985), não surgiram documentos abrangentes de planejamento de ICT no Brasil. Os atuais Livros Verdes, o da Sociedade da Informação (2000) e o de Ciência, Tecnologia e Inovação (2001), não contemplam questões relativas à ICT.

Hoje, no entanto, a comunicação científica é cada vez mais fortemente dependente das tecnologias de informação. O projeto da BDB menciona itens que constituiriam a infraestrutura necessária para viabilizar um ambiente de informação integrado. Planejar e implantar esta infraestrutura seria claramente papel do IBICT. Um ambiente como este garantiria aos usuários "a informação na ponta dos dedos" e simultaneamente ampla visibilidade à produção acadêmica brasileira. Componentes desta infraestrutura seriam: ambientes para submissão e armazenamento de documentos eletrônicos, mecanismos de linkagem de documentos eletrônicos entre si, de modo que um usuário pudesse ter acesso imediatamente a referências e fontes citadas em um documento eletrônico, endereços eletrônicos persistentes, sem os problemas de links inválidos, base de autoridades, linguagens de descrição, esquemas de classificação temática e sistemas de metadados para interoperabilidade entre sistemas e descoberta de informações, armazenamento e preservação de documentos eletrônicos por longo tempo.

Um empreendimento amplo e com um caráter integrador como a proposta da BDB deve ser gerido por um comitê dirigente que inclua as parcerias estratégicas do IBICT neste projeto e representantes da comunidade de C&T. Este Comitê Dirigente deve se reunir periodicamente para analisar e aprovar o Relatório de Atividades da BDB e seu Plano de Trabalho para o próximo período. Entre as reuniões do comitê dirigente, a BDB será gerida por um comitê executivo.

Embora os objetivos da BDB sejam ambiciosos, mostram- se também plenamente viáveis em termos tecnológicos; a implantação da BDB pode se iniciar com um investimento baixo, beneficiando-se da experiência e das metodologias já desenvolvidas nos principais centros internacionais. Estes objetivos também são necessários para que a comunidade acadêmica brasileira possa dispor de uma infraestrutura informacional compatível com os padrões internacionais. Isto permitirá ao país se inserir plenamente nos fóruns científicos internacionais dentro do paradigma atual da comunicação científica. E permitirá principalmente que os sistemas de informação brasileiros se integrem ao fluxo mundial de informações, dando maior visibilidade à produção brasileira em C&T. O interesse do pesquisador é conseguir a maior visibilidade possível para sua produção acadêmica. Os serviços de informação em C&T, entre eles as bibliotecas digitais como a BDB, devem ajudar os autores dos documentos neles armazenados a obter a máxima visibilidade da sua produção, adotando mecanismos que maximizem a integração e interoperabilidade amplas entre serviços de informação.

Referências

BECKETT, Dave; MILLER, Eric; BRICKLEY, Dan. **An XML encoded of simple Dublin core metadata**. [S. l.]: Dublin Core Metadata Initiative, 2000. Disponível em: https://www.dublincore.org/specifications/dublin-core/dcmes-xml/2000-12-01/. Acesso em: 6 ago. 2021.

BECKETT, Dave; MILLER, Eric; BRICKLEY, Dan. **Using Dublin core in XML**. [S. l.]: Dublin Core Metadata Initiative, 2000. Disponível em: http://xml.coverpages.org/dcmes-xml-20000714a.html. Acesso em: 6 ago. 2021.

BRASIL. Ministério da Ciência e Tecnologia. **PADCT**: documento base [e] documentos sínteses dos subprogramas. Brasília: CNPq, 1985. 97 p.

BRASIL. Presidência da República. Secretaria de Planejamento. **III PBDCT**: III Plano Básico de Desenvolvimento Científico e Tecnológico: informação em ciência e tecnologia. Brasília: CNPq, 1984, 69 p.

CIÊNCIA, tecnologia e inovação. **Desafio para a sociedade brasileira**: livro verde. Brasília: Ministério da Ciência e Tecnologia, Academia Brasileira de Ciências, 2001.

CLEVER PROJECT. Hypersearching the web. **Scientific American**, New York, n. 6, 1999. Disponível em: https://www.cs.cornell.edu/home/kleinber/sciam99.html. Acesso em: 6 ago. 2021.

COX, Simon; MILLER, Eric; POWELL, Andy**. Recording qualified Dublin core metadata in HTML meta elements**. [S. l.] : Dublin Core Initiative, 2000. Disponível em: https://www.dublincore.org/specifications/dublin-core/dcq-html/2000-08-15/. Acesso em: 6 ago. 2021.

DAVIS, James R. Creating a networked computer science technical report library. **D-Lib Magazine**, Sept. 1995. Disponível em: http://www.dlib.org/dlib/september95/09davis.html. Acesso em: 6 ago. 2021.

DUBLIN core metadata elements set, version 1.1: reference description. [S. l.]: **Dublin Core Initiative**, 1999. Disponível em: https://www.dublincore.org/specifications/dublin-core/dces/. Acesso em: 6 ago. 2021.

EXTENSIBLE markup language (XML). [S. l.]: World Wide Web Consortium, 2000. Disponível em: http://www.w3.org/XML/. Acesso em: 6 ago. 2021.

GINSPARG, P. Winners and losers in the global research village. In: CONFERENCE ON ELECTRONIC PUBLISHING IN SCIENCE, 1996, Paris. **Proceedings**... Disponível em: https://www.tandfonline.com/doi/pdf/10.1300/J123v30n03_13. Acesso em: 6 ago. 2021.

IMESH. The IMesh toolkit: an architecture and toolkit for distributed subject gateways**. InternatiOnal Digital Libraries**, n. 19, jan. 1999.

IBICT. **Projeto técnico da biblioteca digital brasileira em C&T**. Brasília, 2001. 40 p.

LAWRENCE, Steve. **Free online availability substantially increases a paper's impact**. Disponível em: http://www.nature.com/nature/debates/e-access/Articles/lawrence.html. Acesso em: 6 ago. 2021.

LEINER, Barry M . "The NCSTRL approach to open architecture for the confederated digital libraries." **D-Lib Magazine**, 1998. Disponível em:<http://dlib.org/dlib/december98/leiner/12leiner.html> Acesso em 24 ago. 2021.

LIU, Xiaoming *et al*. Arc: an OAI service provider for digital library federation**. D-Lib Magazine**, v. 7, n. 4, Apr. 2001. Disponível em: http://www.dlib.org/dlib/april01/liu/04liu.html. Acesso em: 6 ago. 2021.

MILLER, Paul. **UK interoperability focus**: Bath, UK, UKOLN. Disponível em: http://www.ukoln.ac.uk/interop-focus/about/. Acesso em: 6 ago. 2021.

PAEPCKE, Andreas *et al*. Search middleware and the simple digital library interoperability protocol . **D-Lib Magazine**, v. 6, n. 3, Mar. 2000. Disponível em: http://www.dlib.org/dlib/march00/paepcke/03paepcke.html. Acesso em: 6 ago. 2021.

PINHEIRO, Lena Vânia Ribeiro; LOUREIRO, José Mauro Matheus. Traçados e limites da ciência da informação. **Ciência da Informação**, Brasília, v. 24, n. 1, p. 42-53, jan./abr. 1995. Disponível em: http://revista.ibict.br/ciinf/article/view/609/611. Acesso em: 6 ago. 2021.

POWELL, James; FOX, Edward A. Multilingual federated searching across heterogeneous collections. **D-Lib Magazine**, Sept. 1998. Disponível em: http://dlib.org/dlib/september98/powell/09powell.html. Acesso em: 6 ago. 2021.

RNP. **Boletim Edupage**. Achando uma agulha (ou 7.079 páginas em uma agulha) na web. 1998.

SHNEIDERMAN, Ben; BYRD, Don; CROFT, W. Bruce. Clarifying search: a user-interface framework for text searches. **D-Lib Magazine**, Jan. 1997. Disponível em: https://www.dlib.org/dlib/january97/retrieval/01shneiderman.html. Acesso em: 6 ago. 2021.

SOCIEDADE da informação no Brasil. **Livro verde**. Brasília: Ministério da Ciência e Tecnologia, 2000.

SULEMAN, Hussein *et al*. Networked digital library of theses and dissertations: bringing the gap for global access: part 1: mission and progress. **D-Lib Magazine**, v. 7, n. 9, Sept. 2001. Disponível em: http://www.dlib.org/dlib/september01/suleman/09suleman-pt1.html. Acesso em: 6 ago. 2021.

TROLL, Denise; MOEN, Bill. **Report to the DLF on the Z39.50 Implementers' Group**. DLF: 2001. Disponível em: https://old.diglib.org/architectures/zig0012.htm. Acesso em: 6 ago. 2021.

ZIMAN, John. **Conhecimento público**. Belo Horizonte: Itatiaia/São Paulo: USP, 1979.

Afinal, o que é biblioteca digital?

Luís Fernando Sayão[1]

1 Introdução

Começando na década passada, a ideia de biblioteca digital, tal qual genericamente imaginamos, tem se transformado no paradigma reiterado dos sistemas e serviços bibliotecários. Desde então, um número impressionante de iniciativas, que incluem produtos e serviços de informação digital, infraestrutura técnica, normativa e comercial, consórcios em escala global, tem sido continuamente desenvolvido em torno dessa ideia.

As bibliotecas digitais surgem num contexto que sobrepõe, por um lado, a integração e uso das tecnologias de informação e de comunicação, das redes de computadores, das tecnologias de apresentação e o barateamento dos meios de armazenamento em massa; e, por outro, a disponibilidade crescente de conteúdos digitais em escala planetária, a possibilidade de digitalização a um custo economicamente viável de conteúdos em mídias convencionais e, ainda, o fenômeno conhecido como coerência das mídias digitais, que abre a possibilidade singular para a concepção de novos serviços de informação a partir da integração de objetos digitais heterogêneos.

Esse contexto de rápidas transformações oferece as condições primordiais para o estabelecimento de uma infraestrutura técnica que viabiliza o surgimento de diversas atividades centradas no conhecimento e na informação globalmente distribuídos. Além do mais, a ambientação tecnologicamente favorável exerce uma forte influência sobre a reconfiguração dos mercados de conteúdos e no delineamento de uma nova dinâmica para a economia da informação, que rapidamente vai incorporando novos patamares de consumo de conteúdos digitais. Assim sendo, surgem diversos produtos e serviços de informação – resultados de concepções inéditas ou de inovações sobre serviços já consagrados. Dentre eles, impulsionada por um conjunto heterogêneo de forças, está a ideia de biblioteca digital.

Verdadeiramente, nós estamos ainda nos primeiros estágios do entendimento de todas as implicações dessa tecnologia e das suas potencialidades. Termos como

1 Centro de Informações Nucleares. Comissão Nacional de Energia Nuclear. luis.sayao@cnen.gov.br.

"redes de conhecimento", "colaboratórios" e bibliotecas digitais têm significados que se misturam. Há uma forte sobreposição de significados entre a pesquisa e o desenvolvimento em arquitetura para federação de coleções digitais autônomas e seus serviços associados e a arquitetura necessária ao comércio eletrônico em larga escala, como apontava Dan Atkins (1998). Dez anos não foram suficientes para resolver essas ambiguidades. A observação de Atkins continua viva nas agendas das comunidades envolvidas com a questão.

Não obstante a intensa atividade de pesquisa, de utilização e de exploração comercial das bibliotecas digitais, não se tem um consenso estável do que seja exatamente uma biblioteca digital e das suas vinculações com a biblioteca tradicional e com a biblioteconomia. O termo "biblioteca digital" vem sendo aplicado a uma variedade extraordinária de coisas – do catálogo online de comércio eletrônico à coleção de programas de computadores –, grande parte delas desvinculada do conceito que temos de biblioteca.

A busca por uma definição mais precisa e consensual para biblioteca digital esbarra também na existência de três termos – biblioteca digital, biblioteca eletrônica e biblioteca virtual – que possuem diferentes significados, mas que são usados frequentemente para designar a mesma coisa.

Ainda há mitos que envolvem as bibliotecas digitais, o que aumenta o grau de mal-entendidos sobre o problema. Kuny e Cleveland (1998) analisaram a irrealidade em torno das bibliotecas digitais num artigo cujo objetivo explícito era ser uma provocação, uma refutação ao tecnologismo exacerbado e aos excessos retóricos que caracterizavam as expectativas em torno do tema "biblioteca digital". Segundo os autores, essa mistificação foi impulsionada pelas companhias tecnológicas, políticos e revistas vanguardistas – as mesmas forças que nos deram mitos como o *paperless office* (escritório sem papel) e vaticinaram o fim dos livros.

A diversidade de atores envolvidos na curta história das bibliotecas digitais bem como o complexo de tecnologias necessárias ao seu pleno funcionamento parecem ser o motivo mais óbvio para ideia pouco precisa do que seja uma biblioteca digital. Entretanto, outros fatores intervêm nas sobreposições conceituais, um deles é o extraordinário potencial de crescimento em diversos domínios e as expectativas geradas nos mais diversos segmentos da sociedade (sayão, 2008). Além do mais, a realização desses potenciais, numa perspectiva social e humanística, constitui um desafio que requer uma interação sofisticada entre várias disciplinas tecnológicas e sociais.

Por outro lado, a linha que separa a concepção – quase ingênua – de biblioteca digital como um mero sistema computacional para armazenamento e acesso a informações eletrônicas tem sido rapidamente pulverizada pela ideia

avassaladora de um ambiente voltado para a criação e para o compartilhamento de informações digitais. Esse ambiente é formado por um complexo de serviços e de coleções de conteúdos distribuídos, gerenciados de forma autônoma, contudo interoperáveis. Nesse patamar, viabiliza-se também o surgimento de uma nova economia da informação e de modelos de negócios que vão moldando as novas possibilidades de distribuição de conteúdos de toda natureza via rede de computadores.

É sobre o que é biblioteca digital, a ótica como ela é vista por diversos grupos sociais e a natureza da sua vinculação e apropriação pela biblioteca tradicional, que vamos discutir rapidamente neste texto.

2 Visões sobre as bibliotecas digitais

Criou-se, historicamente, uma enorme expectativa em torno das potencialidades das bibliotecas digitais, não somente em termos de um novo paradigma de sistema de informações, de busca e recuperação, mas também como um recurso estratégico dentro de contextos altamente institucionalizados, como governo, educação, cidadania, negócios e pesquisa científica.

O conceito de uma biblioteca digital meramente equivalente a uma coleção de objetos digitalizados, assistida por uma ferramenta de gestão de informação, torna- se tosco e já não cabe nas utopias desses inúmeros setores. A ideia de biblioteca digital como um "ambiente distribuído que integra coleções, serviços e pessoas na sustentação do ciclo de vida completo de criação, disseminação, uso e preservação de dados, informação e conhecimento" (DUGUID, 1997) – conforme preconizado pelo relatório final do *Santa Fé Planning Workshop on Distributed Knowledge Work Environments* –, talvez esteja mais próxima do que se almeja para bibliotecas digitais agora e num futuro possível.

A complexidade das bibliotecas digitais em termos tecnológicos e organizacionais, somada ao seu universo vasto e variado de usuários e à multiplicidade de visões – reais e imaginárias – sobre as suas possibilidades e a sua extensão, impacta significativamente a construção de uma definição comum. "Apesar das intensas atividades de pesquisa e de desenvolvimento em torno das várias vertentes do problema, não se tem absolutamente claro o significado do termo biblioteca digital" (HARTER, 1997).

Passada mais de uma década, a afirmação de Harter continua sendo irritantemente verdadeira: biblioteca digital é uma ideia em movimento, ainda se desenvolvendo e tomando forma. "Nós estamos agora na adolescência das bibliotecas digitais", confirmam Lagoze e seus colaboradores (2005, p. 1) pensando nos motivos de preocupação e otimismo que essa fase turbulenta representa.

A impossibilidade de uma definição de consenso acontece por vários motivos, porém, o mais importante deles é que o termo "biblioteca digital" é usado para denotar um número extraordinário de coisas – de coleções pessoais até a internet inteira. Na maioria das vezes essas coisas só têm em comum uma remota manipulação de recursos informacionais digitalizados (HARTER, 1997). Somam-se ainda o grande número de atores que contribuíram para o desenvolvimento e a implementação de bibliotecas digitais e aqueles que estão envolvidos profissionalmente no seu uso, além, é claro, do dinamismo próprio da ambientação tecnológica que sustenta essas bibliotecas. Biblioteca digital representa um espaço sinérgico de um grande número de áreas da tecnologia da informação e várias outras disciplinas e campos de pesquisa, como biblioteconomia, ciência da informação, museologia, arquivologia e gestão do conhecimento, para citar algumas das mais importantes (CANDELA *et al*., 2007).

Dessa forma, a maioria das definições é fortemente influenciada pela percepção e pontos de vista particulares de pessoas e de organizações de diversas áreas que estiveram envolvidas em empreendimentos voltados para a construção e o uso de bibliotecas digitais. A diversidade de contribuições que tanto serviu para o enriquecimento da área criou, ao mesmo tempo, uma zona obscura de indefinições. Para ilustrar essa pluralidade de visões e possibilidades de uso, um resumo da ótica dos cientistas da informação e bibliotecários, cientistas da computação, arquivistas, políticos e governantes, editores, educadores e professores, comunidades da área cultural e do comércio eletrônico é apresentado a seguir tendo como base o artigo de Urs (2007).

A comunidade de biblioteconomia e ciência da informação visualiza a biblioteca digital menos como um sistema de computação – uma máquina – e mais como uma instituição, como uma extensão lógica do que as bibliotecas vêm fazendo desde os tempos imemoriais, ou seja, adquirindo, organizando e disseminando conhecimento usando as tecnologias correntes. O que o bibliotecário deseja é a ampliação dos recursos e dos serviços disponíveis e também a audiência das bibliotecas. Na sua perspectiva prática, o acesso simultâneo a um mesmo documento digital por um número indefinido de usuários significa o fim da lista de empréstimo. Para ele a biblioteca digital é um estágio a mais no desenvolvimento contínuo de novos meios de publicação – em que a biblioteca soma a responsabilidade de também ser uma publicadora web –, bem como uma nova infraestrutura tecnológica e organizacional voltada para potencializar a sua missão de disseminar informação e conhecimento. Porém, enquanto os profissionais de informação têm uma perspectiva de continuidade evolutiva em relação às bibliotecas digitais, outras visões importantes se sobrepõem.

Os profissionais da área de ciência da computação enxergam as bibliotecas digitais como uma extensão dos sistemas de computadores em rede – um sistema que oferece facilidades informacionais. Essas visões se fragmentam à medida que se analisa com um grau a mais de detalhes as diferentes áreas que compõem o domínio da ciência da computação. Por exemplo, enquanto os pesquisadores da área de Recuperação da Informação (RI) veem as bibliotecas digitais como uma ampliação dos sistemas de recuperação de informação em que os documentos e sua representação (ou descrição) são diferentes da RI tradicional, quem trabalha com sistemas multimídia considera as bibliotecas digitais uma aplicação dessa tecnologia; para pesquisadores da área de base de dados, a biblioteca digital é tão somente uma ampla base de dados.

Apesar das controvérsias apaixonadas, a maioria dos políticos e governantes percebe a biblioteca digital como parte da infraestrutura tecnológica necessária para a superação da desigualdade informacional e de acesso, e como mais um recurso para apoio dos programas de inclusão digital. Consideram, com maior ênfase, a biblioteca digital como um insumo básico para a pesquisa, o ensino superior e a pós-graduação e como um instrumento para a maior visibilidade de bens e instituições culturais. Os governantes, com intensidade variável, têm investido em infraestrutura computacional e de redes que beneficiam diretamente as iniciativas na área de bibliotecas digitais. Grande parte dos projetos mais relevantes são iniciativas do poder público, financiados por suas agências e, não raro, apoiados por segmentos da iniciativa privada interessada em expandir suas áreas de atuação.

Os editores, desde a revolução de Gutemberg, têm continuamente desempenhado um papel fundamental na facilitação da produção e distribuição de informação. A percepção da indústria editorial em relação à nova mídia representada pelas bibliotecas digitais é ambivalente: em contrapartida às novas oportunidades mercadológicas existem as ameaças representadas pelas novas formas de autopublicação e o movimento crescente em torno do acesso livre, o que exige uma adaptação permanente a um meio que se renova constantemente. Numa visão otimista, para o mundo editorial, a biblioteca digital constitui um novo modo de distribuição de conteúdos e um novo mercado – bastante competitivo – a ser conquistado, num contexto de mudança da economia da informação. Para isso os editores estão se adaptando ao paradigma da publicação eletrônica, integrando mídias, criando novos modelos de negócio, como os portais agregadores, e estabelecendo parcerias com organizações mais próximas ao mundo da internet.

Para os educadores e os professores que sempre tiveram uma relação de colaboração quase que simbiótica com as bibliotecas tradicionais, as bibliotecas digitais podem ser um meio de ampliar essa relação clássica. Para eles, as bibliotecas digi-

tais constituem um novo recurso de aprendizado, apoiados por conteúdos multimídia, interatividade e integração de informações heterogêneas de que o ensino e, particularmente, o ensino a distância não podem prescindir. As bibliotecas digitais abrem possibilidades extraordinárias para a educação e o ensino, mudando paradigmas e estabelecendo novas metodologias pedagógicas. São as áreas que mais podem se beneficiar dessa nova tecnologia.

Para os arquivistas, as bibliotecas digitais rompem com a relação quase antagônica entre a preservação e o acesso existente no mundo do papel e dos demais materiais analógicos (sayão, 2005). Isso acontece na medida em que a digitalização se torna um meio de preservar os conteúdos raros, únicos ou frágeis, ao mesmo tempo em que proporciona acesso universal a representações digitais desses conteúdos através das bibliotecas e arquivos digitais. A digitalização é vista pelos arquivistas como uma alternativa à microfilmagem tradicional com a ressalva dos problemas de integridade e confiabilidade dos conteúdos digitais, ou seja, do seu valor de prova e de sua preservação de longo prazo, que é uma preocupação constante de toda a comunidade arquivística.

Para os pesquisadores, a colaboração é a chave para a pesquisa e o desenvolvimento. Nesse sentido, eles percebem a biblioteca digital como um espaço dinâmico voltado para a geração, o compartilhamento e a disseminação de conhecimento. Através das bibliotecas digitais, os dados de pesquisa agora podem ser acessados em escala planetária pelos pesquisadores interessados. Essa característica é de grande importância para o surgimento do conceito de "colaboratórios" – resultado da contração das palavras "colaboração" e "laboratório", significando um centro de pesquisa sem paredes onde os pesquisadores interagem eletronicamente no desenvolvimento de projetos inovadores. Projetos como Genoma Humano, baseados em compartilhamento internacional de dados de pesquisa e análises, são exemplos significantes da ideia de um colaboratório.

Ainda há a perspectiva da biblioteca digital como forma de apropriação do mundo da informação pelo comércio eletrônico. Para as organizações comerciais, as bibliotecas digitais estabelecem um novo mercado global, constituindo, para alguns autores, um caso específico de economia da informação (schäuble; smeaton, 1998). Um dado importante é que os desenvolvedores de bibliotecas digitais têm deliberadamente incorporado modelos econômicos e de preços nas arquiteturas de bibliotecas digitais.

No campo cultural, o que se observa é que a biblioteca digital é um meio privilegiado de dar visibilidade global a manifestações culturais antes circunscritas às suas comunidades e sem canais de comunicação para fora delas. O desenvolvimento de metodologias e técnicas para recuperação multilíngue de informação,

somado ao desenvolvimento de recursos linguísticos para serem acoplados às bibliotecas digitais, vai ajudar as comunidades que se expressam em outros idiomas que não o inglês a superarem as barreiras linguísticas no acesso e na disseminação de informações.

3 A biblioteca versus a "googlização"

Tentando interpretar essa pluralidade de entendimentos e expectativas sobre o que é biblioteca digital, Harter (1997) contrapõe as duas visões extremas sobre a natureza das bibliotecas digitais: uma visão abrangente que toma a biblioteca digital tal como a web é hoje – anárquica e individualista; e uma visão que toma a biblioteca digital como uma metáfora, ou mesmo uma extensão, da biblioteca tradicional. No espaço entre esses limites são discutidas as diferenças essenciais: propriedades de localização física, de conteúdo, de critérios de seleção, de organização, controle de autoridades, de autoria, de acesso, de grupos de usuários-alvo, de serviços, de taxação e de fixidade – conceito que está relacionado com a integridade e a segurança dos conteúdos e suas propriedades de permanência.

Num extremo, está a "googlização" das bibliotecas digitais, referindo-se à incômoda e errônea concepção de que o Google[2] representa a apoteose da informação digital e que os problemas existentes nesse domínio já foram resolvidos ou serão resolvidos por esse serviço ou por outra ferramenta semelhante. Esse estreitamento das discussões conduz à visão míope de que a biblioteca digital está limitada à busca e ao acesso – funções essenciais (e ainda desafiadoras), mas que são somente parte do ambiente informacional circunscrito pela ideia plena de biblioteca, seja ela imaginária ou real (LAGOZE *et al.*, 2005). Essa visão está turvada pelo fato de mais e mais pessoas estarem usando a internet como a principal fonte de informação. De fato, a internet tem sido referida por muitos como "uma vasta biblioteca, contendo todo tipo de informação conhecida pelos seres humanos" (WALLACE, 1999). Entretanto, essa constatação não pode ser ignorada como elemento de compreensão do seu contrário, pois, diferentemente das bibliotecas tradicionais, onde as fontes de informação adicionadas às coleções são cuidadosamente selecionadas, organizadas e descritas – classificadas, catalogadas, indexadas, resumidas –, isso não acontece com frequência nas coleções encontradas na internet. Porém, a infraestrutura oferecida pela internet é um veículo de dramática importância para a distribuição de informação de qualidade para os usuários, e é parte essencial da infraestrutura tecnológica de que as bibliotecas digitais não podem prescindir.

No outro extremo, observa-se uma tendência convergente na direção do enqua-

2 Disponível em: http://www.google.com.br. Acesso em: 1 fev. 2009.

dramento das bibliotecas digitais aos cânones biblioteconômicos, principalmente no que concerne à organização e à representação dos recursos informacionais e também às relações orgânicas com suas comunidades-alvo. Isso parece indicar que as bibliotecas digitais devem se equiparar às bibliotecas tradicionais, ao mesmo tempo em que criam condições técnicas para expandir os limites, as formulações e o alcance espacial e temporal do que sempre conhecemos como biblioteca. Entretanto, é importante assinalar que vai ficando cada vez mais nítido que essa visão expandida de biblioteca exige novas reflexões sobre os modelos de informação e de serviços sobre os quais elas estarão baseadas.

Essa convergência para a biblioteconomia pode ser justificada de várias maneiras, porém a mais convincente delas é também a mais óbvia: biblioteca digital continua sendo biblioteca.

> O progresso tecnológico mudou a maneira como as bibliotecas fazem o seu trabalho, mas não a razão do seu trabalho. Ainda que desenvolvimentos tecnológicos mais contundentes – como a conexão de um computador a outro numa cadeia contínua pelo mundo afora – possam alterar o conceito fundamental de biblioteca no século XXI, podemos supor que a tecnologia não vai mudar substancialmente o negócio das bibliotecas, que é conectar pessoas com informações (KUNY; CLEVELAND, 1998, p. 1, tradução nossa).

É imprescindível compreender que a tecnologia atual está focada na conversão de papel para formatos digitais e não na conversão da biblioteca *in toto* para formatos digitais (BROWN, 2005). Assim como uma biblioteca de audiovisual ou de microfilmes continua sendo uma biblioteca, o conceito atual de biblioteca digital constitui um subconjunto de um conceito mais extenso de biblioteca, e não um substituto para ele. Todos os valores e funções da biblioteca continuam válidos, o que muda são os objetos físicos que formam a biblioteca e, naturalmente, o instrumental tecnológico para manipulá-los. As mídias digitais devem ser vistas como um novo suporte na longa lista de materiais que a civilização tem, ao longo da história, utilizado para registrar e transmitir o conhecimento para gerações futuras. Como os outros materiais, nós podemos esperar que eles sejam utilizados na proporção em que a sua disponibilidade local, as tecnologias de apoio, seu custo e a sua confiabilidade sejam adequados e suficientes para armazenar e disseminar informação e conhecimento de acordo com as exigências do seu tempo. As novas gerações de bibliotecas digitais não devem ser consideradas como meros repositórios de informações estáticas. Antes disso, elas devem ser reconhecidas como

núcleo inicial do que, num estágio futuro, constituirá uma parte substancial do conhecimento humano (THANOS, 2004).

4 Biblioteca digital: invenção ou reinvenção?

O conceito de biblioteca digital não é algo que desponta desvinculado da ideia ancestral que temos de biblioteca. Ao contrário, ele se desenvolve tendo como fundamento uma analogia direta com a biblioteca tradicional e com a sua missão de organizar coleções impressas e outros artefatos, de operar serviços e sistemas que facilitem o acesso físico e intelectual – e também o acesso de longo prazo – aos seus estoques informacionais.

Assim como no surgimento de outras concepções da era digital, que são recriações de ideias já estabelecidas, como é, por exemplo, o correio eletrônico, a biblioteca digital, num primeiro momento, espelha-se na biblioteca tradicional, para em seguida expandir esse conceito já consagrado através da apropriação e uso das tecnologias disponíveis.

"Adicionando o adjetivo 'digital' ao nome 'biblioteca', o futuro parece estar reconciliado com o passado" (LYMAN, 1996, p.1). Alegorias futurísticas como bibliotecas digitais e publicações eletrônicas são tranquilizadoras porque elas sugerem uma continuidade institucional entre o passado e o futuro. Pois, se é verdade que a inovação tecnológica geralmente começa imitando o passado, não são as novas ferramentas que constituem inovação, mas sim as novas instituições. "Elas acalmam e ocultam a tensão latente que existe entre tecnologia digital e as instituições de uma sociedade industrial, tensões que levam a questões importantes sobre a natureza das bibliotecas digitais" (LYMAN, 1996, p. 1). Em outras palavras, bibliotecas digitais parecem oferecer-nos toda a conveniência, a eficiência, a sofisticação da tecnologia digital dentro da ideia familiar e confortável de uma biblioteca (MCPHERSON, 1997). Nessa direção, a biblioteca digital parece antes querer reforçar os fundamentos da biblioteca e da biblioteconomia do que aniquilá-los, como temem alguns.

O produto que gerenciamos nas bibliotecas tradicionais é informação, e o seu invólucro que nos é mais familiar, o padrão códice, tem influência decisiva sobre a arquitetura da biblioteca e sobre o seu funcionamento, mas ele não define por si só o que é uma biblioteca. "Nós não estamos preocupados em qualificar nossas bibliotecas chamando-as de 'bibliotecas de tabletes' ou 'bibliotecas de rolos de papiros', por que então temos que qualificar as bibliotecas digitais?", interroga-se Braude (1999, p. 86, tradução nossa) num artigo com um título interessante: "Virtual ou Real: o Termo Biblioteca É o Bastante".

Mas, apesar de a biblioteca digital ser, na maioria das vezes, um serviço vinculado à biblioteca tradicional, fica claro que existe uma distinção que deve ser feita

entre elas. Os invólucros físicos e monolíticos em que a informação está fixada – por exemplo, um livro – são adequados ao acesso direto pelos nossos sentidos e podem ser manuseados fisicamente; por outro lado, dados digitais são constituídos de sinais eletrônicos que independem de mídias, mas que dependem de máquinas e programas de computadores que os interpretem antes de qualquer interação humana com eles. A transição do impresso para o digital implica também a criação de camadas de funcionalidades, de modos diferenciados de disseminação e entrega da informação e na forma como nos relacionamos com ela. A informação digital não é antagônica à informação impressa, porém, no patamar atual, também não é a sua mímica. O seu surgimento muda muita coisa.

A passagem inicial do impresso para o digital teve como ênfase a conversão retrospectiva direta de conteúdos impressos para formatos digitais, por exemplo, a conversão de documentos raros, frágeis ou muito consultados. A versão digitalizada dos estoques informacionais da biblioteca tradicional proporcionou a possibilidade inédita do acesso independente de distância e de tempo, o compartilhamento por mais de um usuário de uma mesma obra a um custo muito baixo e, é claro, o acesso instantâneo e fácil a uma versão digital do texto completo.

Muito além da mera conversão retrospectiva, a emergência da web acelerou o surgimento de novos gêneros de tipos de documentos que não tinham equivalência no domínio da informação impressa e existiam somente no domínio da computação e da comunicação em rede. No contexto ciberespaço, a informação digital pode ser transportada na velocidade da luz, armazenada em densidade atômica, e convergir em novos tipos de documentos que combinam texto, imagem, gráficos, vídeo, áudio, *hiperlinks*, *applets* e tudo mais que a inovação tecnológica e força do mercado possam proporcionar.

As bibliotecas digitais incluem as funcionalidades das bibliotecas tradicionais, mas potencialmente vão além em escopo e significado. O ambiente da biblioteca digital é um espaço dinâmico, constituído de informações eletrônicas, com níveis diferenciados de granularidade, e serviços que possibilitam inúmeras configurações nas suas formas de disseminação e uma gama extraordinária de usos e reusos para os seus estoques informacionais e para as representações correspondentes.

A substituição de papel pelo documento eletrônico está assentada em algumas importantes diferenças: no armazenamento distribuído em formas digitais, na comunicação direta, online, na obtenção do material via redes de computadores, na multiplicidade de cópias a partir de uma versão original, no nível de granularidade que é possível tratar as informações digitais e nas suas possibilidades de reuso. Essas diferenças se desdobram em transformações tão profundas que eventualmente

deixam a biblioteca digital distante de uma mera expressão da biblioteca tradicional (mcpherson, 1997). Essas diferenças e transformações – que ainda estão em curso – as tentativas de definição tentam traduzir.

5 Uma definição possível

Os sonhos e as utopias, juntamente com as realidades, expressam-se de diversas formas quando adicionamos o termo "digital" à ideia precisa que a maioria de nós tem sobre o que é biblioteca. No momento em que se analisa a multiplicidade de ideias sobre o que é biblioteca digital, o único consenso possível de se distinguir é que o conceito de biblioteca digital não é equivalente a uma mera coleção digitalizada apoiada por uma ferramenta de gestão de informação. Esse primeiro patamar na evolução das bibliotecas digitais foi substituído por um conceito mais sofisticado que envolve um ambiente onde estão reunidas coleções, serviços e pessoas com a missão de dar apoio ao ciclo completo de criação, disseminação, uso e preservação de dados, informação e conhecimento, como propõe Paul Duguid (1997).

É necessário trazer, portanto, o debate para dentro dos limites da realidade onde atuam os bibliotecários e demais profissionais do conhecimento e onde se desenrolam pesquisas e as práticas mais importantes para a área. Localizar os possíveis atributos e propriedades das bibliotecas digitais nesse espaço de pesquisas e práticas pode ajudar a definir as vinculações das bibliotecas digitais ao universo das bibliotecas. Os autores Savanur e Nagaraj (2004) e Urs (2007) trilharam esse caminho e definiram conjuntos de características que contornam o desafio de uma definição mais formal. Antes deles, a *Association of Research Libraries* (arl) lançou mão da mesma estratégia, registrada no documento *Definition and Purposes of a Digital Library* (arl, 1995). Essas características foram reunidas a seguir:

- as bibliotecas digitais são a contraparte digital das bibliotecas tradicionais e incluem materiais eletrônicos (digitais) bem como materiais impressos e ainda outros materiais – por exemplo, áudio, vídeo e objetos que não se enquadram na mídia impressa e nem podem ser disseminados em formato digital ainda;
- uma biblioteca digital possui e controla a informação. Ela oferece acesso à informação, e não apenas aponta para ela;
- uma biblioteca tem uma estrutura organizacional unificada com pontos consistentes para acesso aos dados;
- uma biblioteca digital não é uma entidade única, ela pode também oferecer acesso a materiais digitais e recursos de outras bibliotecas digitais;

- bibliotecas digitais apoiam o acesso rápido e eficiente a uma grande quantidade de fontes de informação distribuídas, porém vinculadas por links e que são plenamente integradas;
- bibliotecas digitais têm coleções que: a) são volumosas e persistentes ao longo do tempo; b) são bem organizadas e bem gerenciadas; c) contêm formatos variados; d) contêm objetos e não somente a sua representação; e) contêm objetos que não podem ser obtidos de outra forma;
- bibliotecas digitais incluem todos os processos e serviços oferecidos pelas bibliotecas tradicionais, embora esses processos tenham que ser revisados para acomodar diferenças entre mídias digitais e impressas;
- as bibliotecas digitais cumprem o paradigma do acesso onipresente, a qualquer hora e em qualquer lugar. Existe uma biblioteca onde houver um computador pessoal conectado a uma rede. As bibliotecas digitais estão sempre disponíveis;
- as bibliotecas digitais intensificam o conceito de compartilhamento de recursos provenientes das bibliotecas tradicionais;
- as bibliotecas digitais se dirigem a uma ou a um conjunto de comunidades de usuários.

A *Digital Library Federation* (DLF) foi mais adiante. Ela registra na sua página web uma definição abrangente que institucionaliza a visão biblioteconômica das bibliotecas digitais:

> Bibliotecas digitais são organizações que disponibilizam os recursos, incluindo pessoal especializado, para selecionar, estruturar, oferecer acesso intelectual, interpretar, distribuir, preservar a integridade e assegurar a persistência ao longo do tempo de coleções de trabalhos digitais, de forma que eles estejam pronta e economicamente disponíveis para uso de uma comunidade definida ou um conjunto de comunidades (DLF, 2009, tradução nossa).

A própria DLF (2009) oferece uma interpretação para a definição que ela estabelece: "Pode- se, naturalmente, revisar, refinar e de outra forma melhorar essa definição abrangente. Entretanto, o que se propõe aqui é principalmente sugerir que existe um conjunto de atributos que confere coerência ao conceito de biblioteca digital". Esses atributos incluem funções de coleções, organização, preservação, acesso e economia. O que significa dizer que os projetos que envolvam bibliotecas digitais precisam ser definidos e mensurados segundo o desenvolvimento desses

atributos. Porém, é importante deixar claro que a definição proposta também enfatiza que as bibliotecas digitais devem ser definidas e mensuradas em relação às comunidades a que elas servem.

Esses atributos, de certa forma, dão densidade ao conceito proposto pela DLF, ao mesmo tempo em que jogam luz sobre outras funções biblioteconômicas importantes para o desenvolvimento de bibliotecas digitais, que não somente a busca e o acesso a coleções digitais. As propostas da ARL e da DLF revelam que a biblioteca digital não está isenta de administrar os serviços de uma biblioteca no seu elenco de funções, e que a sua vertente digital deve conviver com as outras modalidades de informação disponíveis, estabelecendo uma possível convergência entre os reinos digital e o impresso.

A definição da DLF tem sido adotada amplamente por grande parte das comunidades vinculadas às áreas de biblioteconomia e de ciência da informação. Tem sido adotada também como marco primordial dos principais projetos de pesquisa das muitas áreas que permeiam os estudos em biblioteca digital. Entretanto, como vimos anteriormente, ela revela apenas uma das muitas faces do que é universalmente discutido e entendido como biblioteca digital.

6 À guisa de conclusão

A ideia de biblioteca digital tem muitas faces, mas nenhuma delas a define completamente e esgota todos seus significados. As definições de biblioteca digital se reconfiguram de acordo com os seus inúmeros protagonistas que se espalham por muitas áreas.

Mesmo no contexto mais restrito da biblioteconomia e da ciência da informação, há uma multiplicidade de visões sobre a natureza das bibliotecas digitais que se sobrepõem, e de práticas que se concretizam em harmonia com essas visões. O que se pode concluir com algum risco é que o conceito de biblioteca digital é algo ainda no estágio transiente de evolução e que provavelmente guardará significados distintos ou receberá denominações distintas à medida que as atuais sobreposições conceituais se resolvam. O breve passado das bibliotecas digitais não foi capaz de resolver essas ambiguidades, quem sabe o futuro seja rápido em harmonizá-las.

No entanto, o que se observa é que a divergência se instala menos em relação à natureza dos serviços, produtos e interações que uma biblioteca digital pode oferecer – isso parece cada vez mais claro e consensual –, e mais em relação à natureza da sua vinculação com a biblioteca e seus fundamentos.

Essa questão deve ser ressaltada, pois está reiteradamente explícito nas definições correntes que biblioteca digital e biblioteca tradicional são coisas separadas

e distintas. Elas não incluem a perspectiva simples de que a biblioteca pode ser as duas coisas: impressa e digital.

O digital não é o antagônico do impresso, como o rolo de papiro não é o antagônico do livro. Para cumprir o seu papel ancestral a biblioteca sempre se apropriou das mais avançadas tecnologias disponíveis e vem continuamente evoluindo no ritmo dessas tecnologias. Assim foi com a tecnologia de microfilme, com a computação e agora com a web. Desde o surgimento dessas tecnologias, percebeu- se que elas trariam um ganho extraordinário de produtividade e de amplitude nas funções administrativas, técnicas e de intercâmbio de informação e conhecimento no mundo das bibliotecas.

Portanto, a biblioteca digital é mais um marco – que não traz aniquilamentos e nem pontos de singularidade – na continuidade evolutiva das bibliotecas, que caminham rapidamente para se tornarem palácios híbridos de acesso à informação e ao conhecimento distribuído, para onde convergem e se integram todos os tipos de mídias.

Referências

ATKINS, Dan. Vision for Digital Libraries. *In*: SCHÄUBLE, Peter; SMEATON, Alan. **An International Research Agenda for Digital Libraries:** Summary Report of the Series of Joint NSF-EU Working Groups on Future Directions for Digital Libraries Research. 1998. Disponível em: https://www.ercim.eu/publication/ws-proceedings/DELOS-B/dl_sum_report.pdf. Acesso em 28 jul. 2021.

ARL. **Definition and purposes of a digital libraries**. 1995. Disponível em: http://yunus.hacettepe.edu.tr/~tonta/courses/fall99/kut655/DL-definition.htm. Acesso em 28 jul. 2021.

BRAUDE, Robert M. Virtual or actual: the term library is enough. **Bulletin of the Medical Librarians Association**, v. 87, n. 1, 1999, pp. 85-7.

BROWN, Mary E. **History and Definition of Digital Libraries**. New Haven, C. T., Southern Connecticut State University, 2005. Disponível em: www.southernct.edu/~brownm/dl_history.html. Acesso em: 30 set. 2008.

CANDELA, Leonardo *et al*. Setting the foundations of digital libraries. **D-Lib Magazine**, v. 13, n. 3/4, Mar.-Apr./2007. Disponível em: https://dlib.org/dlib/marcho7/castelli/03castelli.html. Acesso em 28 jul. 2021.

DFL. **About**. Disponível em: http://www.diglib.org/about/dldefinition.html. Acesso em: Acesso em: 1 fev. 2009.

DUGUID, Paul. **Report of the Santa Fe Planning Workshop on Distributed Knowledge Work Environments**: Digital Libraries. University of Michigan School of Information, Sept./1997. Disponível em: http://www.si.umich.edu/

SantaFe/. Acesso em: 25 maio 2008.
HARTER, Stephen. Scholarly communication and the digital library: problem and issues. **Journal of Digital Information**, v. 1, n. 1, 1997. Disponível em: http://journals.ecs.soton.ac.uk/jodi/Articles/vo1/io1/Harter/. Acesso em 28 jul. 2021.
KUNY, Terry; CLEVELAND, Gary. The digital library: myths and challenges. **IFLA Journal**, v. 24, n. 2, 1998. Disponível em: https://journals.sagepub.com/doi/abs/10.1177/034003529802400205?journalCode=iflb. Acesso em 28 jul. 2021.
LAGOZE, Carl *et al*. What is a digital library anymore, anyway? **D-Lib Magazine**, v. 11, n. 11, Nov./2005. Disponível em: https://www.immagic.com/eLibrary/GENERAL/DIG_LIB/D051100L.pdf. Acesso em 28 jul. 2021.
LYMAN, Peter. What is a digital library? Technology, intellectual property, and the public interest. **Daedalus**, v. 125, n. 4, 1996.
MCPHERSON, Madelaine. Managing Digital Libraries. **CSIRO Information, Management & Technology Conference**, 1997. Disponível em: http://www.usq.edu.au/users/mcpherso/csiro.htm. Acesso em: 31 mar. 2008.
SAVANUR, Kiran P.; NAGARAJ, M. N. Design and implement of digital library: an overview. *In*: TALAWAR, V. G.; BIRADAR, B. S. (ed.). **ASSIST National Seminar**. Kuvempu University, Shimoga, Karnataka, Índia. 2004, pp. 47-53. Disponível em: http://eprints.rclis.org/archive/00007842/01/ASSIST.pdf. Acesso em: 15 nov. 2008.
SAYÃO, Luis Fernando. Preservação Digital no Contexto das Bibliotecas Digitais. *In*: MARCONDES, Carlos Henrique; KURAMOTO, Hélio; TOUTAIN, Lidia Brandão; SAYÃO, Luis Fernando (org.). **Bibliotecas digitais**: saberes e práticas. Salvador/Brasília, UFBA/IBICT, 2005, p. 115-49.
SAYÃO, Luis Fernando. Bibliotecas Digitais e Suas Utopias. **Ponto de Acesso**, v. 2, n. 2, 2008. Disponível em http://www.portalseer.ufba.br/index.php/revistaici/article/view/2661. Acesso em 28 jul. 2021.
SCHÄUBLE, Peter; SMEATON, Alan. **An International Research Agenda for Digital Libraries**: Summary Report of the Series of Joint NSF-EU Working Groups on Future Directions for Digital Libraries Research. 1998. Disponível em: http://www.ercim.org/publication/ws-proceedings/DELOS-B/dl_sum_report.pdf. Acesso em 28 jul. 2021.
THANOS, Costantino. DELOS: a network of excellence on digital libraries**.** **DELOS Newletter**, n. 1, 2004. Disponível em: https://www.ercim.eu/publication/Ercim_News/enw57/thanos.html. Acesso em 28 jul. 2021.
URS, Shalini. **Digital libraries**: the road ahead. International Caliber. Panjab University, Chandigharh, 2007. Disponível em: https://ir.inflibnet.ac.in:8443/ir/bitstream/1944/509/1/1-11.pdf. Acesso em 28 jul. 2021.

WALLACE, Koehler. Digital libraries and world wide web sites and page persistence. **Information Research,** v. 4, n. 4, July/1999. Disponível em: http://informationr.net/ir/4-4/paper60.html. Acesso em 28 jul. 2021.

Bibliotecas digitais e suas utopias

Luís Fernando Sayão[1]

1 Introdução

Sempre houve o sonho de uma biblioteca total que reunisse toda a sabedoria, toda experiência e toda a literatura humana. Quando o escritor e bibliotecário Jorge Luis Borges (1941) escreveu o conto *A Biblioteca de Babel*, ele quis ir ainda mais adiante nesse sonho. A biblioteca infinita de Borges se confundia com o próprio universo e guardava em espaços hexagonais intermináveis todos os livros possíveis – os escritos e os por serem escritos –, em todos os idiomas e dialetos – os decifráveis e os indecifráveis –, fruto das combinações de vinte e poucos símbolos. Não obstante, a biblioteca de Borges, ou antes, o Universo, pode ser reduzida a um único livro "que contém o código que resume todo o resto e se assemelha a um deus" (p. 51, tradução nossa). Ele é como um Aleph que reúne em um único ponto todas as experiências de todas as vidas (borges, 1949). O livro total de Borges só encontra paralelo no sonho do poeta Stéphane Mallarmé, que perseguia a ideia de dar forma a um livro integral, múltiplo, que teria uma arquitetura inovadora, sem começo, meio e fim, e que contivesse todos os livros possíveis. Um gerador de textos, de poesias.

A ficção literária está povoada de bibliotecas vastas e imaginárias. Palácios de saberes que ambicionam reunir todos os livros escritos em todos os tempos ou que dominam totalmente o universo dos personagens /que estão à sua volta. Além da biblioteca de Borges, há também a biblioteca criada por Cervantes para Alonso Quijano – Dom Quixote – que era o alimento da sua loucura; a biblioteca repleta de passagens secretas e espelhos, que a tornava virtualmente infinita, imaginada por Umberto Eco em *O Nome da Rosa*; a biblioteca fantástica do Capitão Nemo e muitas e muitas outras (saldanha, 2001). Essas bibliotecas têm em comum, além da vastidão indeterminada de seus acervos, o fato de serem manifestações de um desejo atávico da humanidade de ter ao alcance dos sentidos todo o conhecimento possível.

1 Centro de Informações Nucleares. Comissão Nacional de Energia Nuclear. luis.sayao@cnen.gov.br.

No mesmo plano, exceto pelo fato de realmente ter existido, está a imagem mítica da Biblioteca de Alexandria, fundada provavelmente no século III a.C., que talvez seja a mais antiga referência na concretização da busca secular pela totalização do conhecimento, mas paradoxalmente tornou-se antes o símbolo da impermanência e da fragilidade dos tesouros que acumulava. Os seus milhares de rolos de papiro, pergaminhos, gravuras e livros que registravam a cultura e a ciência da antiguidade desapareceram em sucessivos incêndios que pareciam dramaticamente apontar para a impossibilidade da sua ambição de ser a guardiã de todos saberes.

A ideia de um repositório que se desdobre ao infinito registrando e organizando todo o conhecimento humano parece ser um sonho obsessivamente renovado ao longo do tempo. As mentes mais criativas e ousadas, como a de H.G. Wells e a de Paul Otlet, e as mais avançadas tecnologias de todas as épocas sempre estiveram a serviço da sua concretização. Assim é que nesse contínuo, hoje a Internet e a Web oferecem uma infraestrutura tecnológica que tornam possível mais uma etapa, talvez a mais importante, da longa história do desejo humano de registrar a totalidade das informações e o conhecimento que ela gerou.

Mas é possível também que a Web – com a sua memória de tudo, que é ao mesmo tempo o seu poderio e a sua fragilidade –, possa ser, no seu estágio atual, tão somente um registro absoluto e total, sem inteligência, como a memória de Irineu Funes, o memorioso, outro personagem da ficção fantástica de Borges, que registra tudo, mas é incapaz de pensar e dar significado às suas memórias. Funes é incapaz de generalizações, de abstração e até do esquecimento. "Suspeito entretanto, que não era muito capaz de pensar. Pensar é esquecer as diferenças, é generalizar, abstrair" (BORGES, 1944, p. 72, tradução nossa). Insistindo nessa comparação arriscada, o que diz Borges sobre seu personagem parece sugerir que precisamos ainda de conferir alguma inteligência e uma força ordenadora e integradora que se sobreponha à memória caótica e fragmentada da Web, para que ela finalmente cumpra as suas utopias. A conexão entre memória virtual e inteligência, que parece ser um desafio gigantesco para as áreas de estudo de tecnologias de informação, especialmente as tecnologias semânticas, e para diversas outras áreas, guarda um papel determinante para os futuros serviços de informação, incluindo as bibliotecas.

Os conceitos subjacentes à ideia de biblioteca digital – tecnologias abertas, interoperabilidade e recursos distribuídos – sugerem que a biblioteca universal é algo possível, sem que para isso seja necessário que todas as informações estejam reunidas em um único lugar. Mas para tal, é necessário se dispor de um conjunto de técnicas e de metodologias que viabilizem a invenção de uma metáfora de uma biblioteca em que livros, imagens, músicas, filmes e outros recursos inéditos de in-

formação, distribuídos por todo o mundo, pareçam estar perfeitamente integrados e organizados em estantes feitas de *bits*.

A biblioteca digital pode ser somente um outro passo na busca contínua do santo graal da biblioteca universal. Os novos conceitos envolvidos no amplo domínio das suas áreas de pesquisa, como interoperabilidade, recuperação de informação, preservação digital e tecnologias semânticas são expressões renovadas desse desejo ancestral. Nesse conceito ainda em formação, cuja história ainda está se desenrolando, cabem muitas expectativas.

Nessa direção, o objetivo deste texto é analisar rapidamente os fatos que antecederam o surgimento das bibliotecas digitais e o que os vários segmentos da sociedade esperam delas num futuro próximo.

2 A pré-história das bibliotecas digitais

As perspectivas abertas pelo desenvolvimento da microfotografia formaram as bases tecnológicas que inspiraram as primeiras ideias de se construir repositórios universais de conhecimentos. A possibilidade técnica de se armazenar informações em uma mídia de alta densidade – 300 caracteres por polegada – instigaram algumas mentes visionárias nessa direção. Entretanto, essas ideias permaneceram somente no campo da abstração e só puderam se materializar (ou se virtualizar?) nos dias de hoje, apoiadas pelo complexo de tecnologias da informação que recria, sobre os alicerces dessas tecnologias, a metáfora de uma memória total.

É impossível, e também injusto, falar das ideias que antecederam as bibliotecas digitais sem nos surpreender com as reflexões perturbadoramente atuais de H.G. Wells (1866- 1946), autor de clássicos da ficção científica – ou romances científicos, como ele próprio chamava – como *A Guerras dos Mundos*, publicado em 1898 e adaptado para o cinema três vezes, sendo a última muito recentemente em 2005. Wells, ainda em 1937, delineou o que seria uma *Permanent World Encyclopaedia*, "[...] um repositório onde conhecimento e ideias são recebidas, ordenadas, sumarizadas, classificadas, analisadas e comparadas" (WELLS, 1938, p. 49). Estes repositórios, que integrariam toda a inteligência do mundo, teriam sua base de conhecimento apoiada na tecnologia de microfilmes, na época, ainda em sua infância.

> Os especialistas americanos em microfilmes estão produzindo fac-símiles de livros raros, de manuscritos, de imagens e de amostras, que podem ser facilmente acessíveis na tela de projeção da biblioteca. Por meio do microfilme, os documentos e artigos mais raros e mais complexos podem ser agora estudados diretamente da fonte original, simultaneamente, em salas de projeções (WELLS, 1937, tradução nossa).

Os microfilmes poderiam ser duplicados e enviados para qualquer lugar, onde seriam então ampliados possibilitando que estudantes e pesquisadores pudessem estudar os registros em todos os seus detalhes. "Toda a memória humana pode ser, e provavelmente o será a curto prazo, acessível para cada indivíduo" (WELLS, 1937, tradução nossa). "Qualquer estudante em qualquer parte do mundo, sentado em seu estúdio com o seu projetor, no momento mais conveniente poderá examinar uma réplica exata de qualquer livro ou qualquer documento" (WELLS, 1938, p. 49, tradução nossa).

Wells também imaginava essa memória universal duplicada e distribuída como uma forma de proteção à fragilidade dos registros humanos expostos à violência e à destruição provocadas pela frequência cada vez maior das guerras.

> Ela não necessita estar concentrada num único lugar. Ela não necessita ser venerável como é o cérebro humano ou como é coração humano. Ela pode ser reproduzida exata e completamente no Peru, na China, na Islândia, na África Central ou em qualquer lugar que ofereça segurança contra o perigo e a interrupção (WELLS, 1937, tradução nossa).

Wells com a sua mente visionária tocava em alguns dos principais desafios que estão sendo equacionados hoje pela área de bibliotecas digitais: integração das informações, universalidade e democratização do acesso, fontes de informação distribuída, informação persistente e ainda a preservação, além de aplicações importantes para a pesquisa e o ensino. Substituindo "microfilme" por "arquivos digitais" verificamos o quanto eram exatas as suas utopias.

Outro protagonista da saga do acesso universal ao conhecimento é Paul Otlet (1868- 1944) – o homem que queria classificar o mundo –, figura central no desenvolvimento da Documentação e autor do livro monumental *Traité de Documentacion* (1934). Enquanto para Borges a biblioteca universal era uma abstração da ficção literária, para Otlet era algo possível de se tornar real (WRIGHT, 2007). Toda a sua trajetória foi direcionada para a realização do sonho de reunir a totalidade do conhecimento mundial e classificá-lo de acordo com o sistema desenvolvido por ele e seu amigo Henri La Fontaine – a Classificação Decimal Universal (CDU). Outlet lutou incansavelmente, por décadas, para encontrar uma solução para os problemas técnicos, teóricos e organizacionais que tornassem o conhecimento registrado disponível para aqueles que necessitam dele, para ele um problema crucial para a sociedade. Na construção de suas utopias antecipou alguns dos problemas importantes para os sistemas de informação de hoje, como o estabelecimento de relações entre documentos que dava margem a formação

de um "livro universal", *réseau* (teia) de conhecimento humano, acesso remoto a bases de dados via "telescópio elétrico", dispositivo com conexão por linha telefônica.

Porém, enquanto alguns pensadores perseguiam a utopia da totalização universal do conhecimento segundo a perspectiva do "acesso", existiam outros que sonhavam com sistemas capazes de intensificar a memória humana através de armazenamento personalizado e entrelaçamento de informações (URS, 2007). O amplificador de memória, a máquina utópica concebida por Vanevar Bush em 1945, denominada por ele, ao acaso, de memex – por querer indicar, talvez, *memory extender* – é uma referência obrigatória para todos os que se debruçam sobre os antecedentes das bibliotecas digitais. O memex se contrapõe à ideia das invenções humanas voltadas somente para a amplificação do poderio físico das pessoas, como, por exemplo, um microscópio ampliando o olhar; se contrapõe também à rigidez dos sistemas de informação organizados linearmente de forma hierárquica por catálogos que devem ser percorridos por ordem alfabética, numérica ou por classes ou subclasses, de forma não natural ao cérebro humano. O engenho abre possibilidade da ampliação do poder mental, da capacidade da memória e do seu potencial de associação para um indivíduo, posto que memex é, na sua essência, uma máquina pessoal. Nas próprias palavras de Bush, o

> memex é um dispositivo através do qual um indivíduo armazena todos os seus livros, seus registros e suas comunicações, ele é mecanizado de forma que pode ser consultado com extraordinária velocidade e flexibilidade. O memex é um suplemento pessoal ampliador da memória desse indivíduo (BUSH, 1945, p. 4, tradução nossa).

É quase impossível não pensar num computador pessoal. A tecnologia subjacente à máquina conceitual de Bush, que envolvia uma combinação de controles eletromecânicos, câmeras e leitores de microfilmes integradas em uma mesa de trabalho, permite a exibição de livros, imagens, jornais armazenados em rolos de microfilmes e a ligação com uma biblioteca. Contemplava ainda vínculos, chamados "trilhas", entre as informações, possibilitando uma leitura não linear, por associação, que sugeria algo como *links*, ou mais precisamente, referências cruzadas entre quadros de microfilmes. O memex foi uma primeira inspiração para o hipertexto, porém, ao contrário do que muitos afirmam, a máquina de Bush não estabelecia a ideia de hipertexto da maneira e no nível de granularidade que hoje conhecemos. Isto foi estabelecido nos anos 1960 por outro visionário: Theodor Holm Nelson, mais conhecido como Ted Nelson.

Ted Nelson é um gênio inconformado com os rumos da Web, cujos fundamentos imprescindíveis foram inventados por ele, o hipertexto e a hipermídia, incluindo os próprios termos. Esses conceitos foram desenvolvidos no contexto do seu projeto chamado Xanadu, um paradigma amplo voltado para a implementação de um sistema de hipermídia distribuído, que foi iniciado a partir de 1960 e está até hoje inacabado. A irritação de Nelson é principalmente com o padrão de navegabilidade da Web que ainda imita o papel e não aproveitou a riqueza tridimensional dos *links* concebidos no seu projeto.

O projeto Xanadu – nome dado em homenagem à cidade mítica onde ficava o palácio do imperador mongol Kubla Khan – partiu da premissa de que

> nós precisamos de uma forma das pessoas armazenarem informação não somente por meio de arquivos individuais, mas em rede, permitindo criar, acessar e manipular essa extensa e rica base de dados de forma a poder interligar a informação aí contida de uma maneira segura e eficaz. Os documentos devem permanecer sempre acessíveis, salvos de qualquer tipo de perda, dano, modificação ou censura, preservando dessa forma os direitos do seu autor (FELDMAN, 1990, tradução nossa).

O objetivo do Xanadu era estabelecer o conceito fundacional de "Docuverse", um sistema de bases de dados onde os escritores podiam publicar diretamente os seus textos vinculandos-os a outros documentos, constituindo, dessa forma, uma biblioteca eletrônica universal online de documentos interconectados; um lugar mágico para a memória literária da humanidade em que todas as obras se interligariam. O Docuverse permitia a criação de cópias virtuais de qualquer informação existente, sem problemas de direitos autorais. Isso porque o autor da informação consultada recebia uma determinada quantia automaticamente sempre que alguém acessasse a sua obra. Essa forma de edição – que nos remete imediatamente às formas alternativas de publicação eletrônica, foco de intensas controvérsias nos dias de hoje – já apontava para a publicação diretamente no sistema hipertextual, sem a intermediação das editoras convencionais (ARAÚJO, [200-?]).

Ainda que Xanadu nunca tenha ultrapassado o patamar de um protótipo e nunca tenha sido comercializado – fato que marcou, algumas vezes cruelmente, a trajetória de Nelson e o coloca como um *outsider* –, o sistema foi submetido a constantes desenvolvimentos, que, porém, nunca foram postos em prática na sua plenitude. Entretanto, Nelson modelou muitos dos conceitos fundamentais aos sistemas de hipermídia, incluindo a própria *World Wide Web*, apesar de todas as homenagens terem ficado com Tim Berners-Lee, apontado de forma absoluta como o seu criador.

Desde o início da computação ficou claro que a automação – ou mecanização, como se chamava na época – das bibliotecas traria um extraordinário ganho de produtividade aos processos biblioteconômicos por conta da natureza e do volume de dados tratados pelas bibliotecas. As primeiras aplicações concretas de computadores no apoio a funções de bibliotecas aconteceram no início da década de 1950, por iniciativa da corporação americana IBM. Esse primeiro esforço estava voltado para a utilização de cartões perfurados para dar suporte às operações de processos técnicos da biblioteca (HARTER, 1997); e, no início dos anos 1960, para o desenvolvimento do MARC, sigla para *Machine Readable Cataloguing*, formato legível por computador para representação e intercâmbio de dados bibliográficos. Apesar dos anos, o MARC e suas inúmeras variantes lograram acompanhar todas as mudanças e têm forte presença mundial até hoje.

Ainda nos anos 60, no contexto de um trabalho pouco conhecido no mundo da informação, J.C.R. Licklider (1915-1990) cunhou a expressão "biblioteca do futuro" referindo-se à sua visão de uma biblioteca completamente baseada em computador. Licklider, considerado um dos mais influentes pesquisadores na história da ciência da computação, principalmente por sua atuação na criação e no desenvolvimento da Internet, registrou essas ideias no seu livro *Libraries of the Future* (LICKLIDER, 1965), onde estavam delineadas as características dessas bibliotecas do futuro, que era, em pouquíssimas palavras, uma continuação do exercício de imaginar aplicações para o computador. Nessa direção, Licklider discute no livro como a informação podia ser armazenada e recuperada eletronicamente.

"Lick", como era conhecido por seus colegas, vinha do mundo da ciência da computação, mas tinha também uma forte formação em psicologia, esse fato lhe conferia uma visão única, uma perspectiva inigualável dos problemas que ele estava envolvido. A sua obra mais importante tem o título de *Man Computer Symbiosis* (LICKLIDER, 1960), que não deixa dúvida sobre onde repousava o seu sistema de referências. Nesse livro ele proclamava que computadores tinham que ser desenvolvidos com o objetivo de tornar possível que homens e computadores cooperem na tomada de decisões e no controle de situações complexas, ou seja, que os computadores expandam o intelecto humano. As preocupações de Licklider ultrapassavam os limites do registro, do processamento e da recuperação de dados, ele já pensava em termos de conhecimento e seus fluxos, refletindo que isso poderia se tornar o patamar para uma nova concepção de sistema de biblioteca. Ao sistema teórico de informação originado desses princípios ele deu o nome de "sistema procognitivo" (*progonitive system*), que soa muito familiar a Web de Tim Berners-Lee (HAUBEN, 2007).

Uma década depois, Lancaster (1978) publicou um livro cujo título não deixa dúvidas sobre o teor do seu conteúdo: *Toward paperless information system*, onde

proclamava que, no contexto de uma sociedade sem papel (*paperless society*), em breve as bibliotecas tradicionais teriam seus acervos substituídos totalmente por formas eletrônicas. Lancaster – por estar, talvez, mais próximo do rumo que se desenhava para as tecnologias que potencialmente impactariam os serviços e sistemas de informação – antecipa com precisão muitas das facilidades que as redes de computadores, as publicações eletrônicas e as bibliotecas digitais viabilizam hoje, como a submissão online, a revisão e comentários via rede, a substituição da economia de assinaturas pela leitura por demanda, a interoperabilidade, identificadores persistentes, etc. Porém, a coleção da biblioteca aprisionada nos seus limites físicos e a legitimidade das coleções digitais são questões centrais para ele, que coloca a desmaterialização da coleção como uma questão filosófica para as novas bibliotecas. "O que é a coleção da biblioteca?" (LANCASTER, 1982, p. 8). Ele mesmo conclui imediatamente que a coleção inclui tudo o que a biblioteca pode tornar acessível quando necessário para o seu usuário. Parece que é desta forma que hoje se equaciona essa questão no domínio das bibliotecas digitais.

Lancaster conclui que "por volta do ano 2000, parece inteiramente razoável esperar que as bibliotecas, como as conhecemos hoje, desapareçam. Tudo o que restará serão umas poucas instituições que preservarão os registros impressos do passado" (LANCASTER, 1982, p. 10). Isso não se cumpriu e tudo indica que não será assim num futuro que se pode equacionar. A realidade é que hoje um dos grandes desafios da área de bibliotecas digitais é precisamente integrar a diversidade crescente de objetos digitais e as fontes impressas, fornecendo ao usuário uma visão unificada dos estoques de informação. São as bibliotecas híbridas, que gerenciam coleções digitais e convencionais que despontam como as vitoriosas, ou antes, a Biblioteconomia, que logrou o reconhecimento do seu poder de ordenação a uma Web que parecia caminhar rumo ao caos.

3 História recente

Não está muito claro quando surgiu a primeira biblioteca digital, mas o conceito não apareceu antes do início da década de 1980 (LI, [200-?]) e a área de estudo de bibliotecas digitais só se configurou como um campo explícito de pesquisa a partir de 1990 (DELOS, 2003, p. 1). Biblioteca digital, no sentido tal qual ela é percebida hoje, é visto frequentemente como um fenômeno decorrente do surgimento da Web (URS, 2007), posto que a rede, no seu sentido mais amplo, é que define as condições tecnológicas e ambientais para a sua concretização enquanto um constructo tecnológico e também social. Em 1989 o projeto da *World Wide Web* estava preliminarmente proposto, e desde meados de 1993 começou a crescer em taxas exponenciais. No entanto, os alicerces do desenvolvimento das bibliotecas digitais são mais profundos e antecedem a Web e a própria Internet.

Muitos dos primeiros sistemas de informação chamados de "bibliotecas digitais" eram apenas tipos de coleções digitais e de serviços de informação desenvolvidos de forma isolada: recursos de informação pessoais, coleções de informações organizacionais e de grupos de trabalho e ambientes colaborativos. Porém, não obstante a infraestrutura imprescindível subjacente às bibliotecas digitais atuais providas pela Web, algumas experiências importantes se desenrolaram no período anterior ao surgimento da Internet, entre elas estão o *Project Mercury* da *Carnegie Mellon University* (1989-1992). O projeto usava uma configuração moderna de computadores distribuídos para oferecer acesso a uma grande variedade de bases de dados textuais, incluindo texto completo (ARMS *et al.*, 1992).

Quando se percorre toda a linha temporal da evolução técnica das bibliotecas digitais, se torna claro que as suas bases teóricas e práticas estão fortemente vinculadas às pesquisas desenvolvidas pela área da computação denominada de recuperação da informação. As bibliotecas digitais evoluíram baseadas nas técnicas e princípios desenvolvidos por pesquisadores desse domínio ainda no princípio da década de 1950. Entre eles estão Calvin Mooers – que em 1951 inventou o termo "recuperação da informação" (*information retrieval*) –, James Perry, Allen Kent, Mortiner Taube, Hans Peter Luhn (SARACEVIC, 1999, p.1057); porém, de extraordinária importância está Gerald Salton (1927-1995), que foi antes de tudo um pesquisador, um cientista da computação e pai da moderna recuperação da informação. Salton foi um pioneiro no desenvolvimento das técnicas de indexação automática e sistemas de busca cujos conhecimentos as atuais bibliotecas não podem prescindir. O cerne intelectual das bibliotecas digitais foi "construído sobre sólidos alicerces consolidados em mais de três décadas de pesquisa em recuperação da informação" (URS, 2001, p. 3, tradução nossa).

A emergência e o desenvolvimento das bibliotecas digitais nos primeiros estágios foram impulsionados por duas forças principais: em primeiro lugar, o rápido desenvolvimento das tecnologias de informação, especialmente a multimídia e as redes de computadores, que ofereciam formas mais eficientes e, às vezes, inovadoras de processar, gerenciar e apresentar a informação; em segundo, as pessoas, principalmente, os acadêmicos, que desejavam compartilhar com maior eficiência informações importantes, tais como material bibliográfico, base de dados científicos e resultados de pesquisa. Dessa forma, impulsionados por um contexto tecnológico favorável, os pesquisadores de diversas áreas vislumbravam aplicar ou criar tecnologias que potencializassem o uso e o compartilhamento de informações em formatos digitais num ambiente de rede (LI, [200-?]).

Na primeira metade dos anos 1990, a área de bibliotecas digitais – de um objeto de preocupação quase obscuro limitado a umas poucas pessoas da área de Ciência

da Computação e de profissionais de Biblioteconomia – tornou-se rapidamente um pólo de intensa atração de interesses e de financiamentos, transformando-se numa área altamente institucionalizada. Esse fato teve como desdobramento o surgimento de um grande número de projetos importantes cuja característica mais destacada eram as visões diversificadas que apresentavam. O campo de estudos e práticas de bibliotecas digitais atraiu a atenção de grupos de pesquisa de um amplo espectro de disciplinas e profissões. Esse fato marcou definitivamente a área: vários domínios da academia, da indústria, das empresas, do governo e outros se tornaram parceiros ativos no desenvolvimento e na consolidação do que hoje chamamos bibliotecas digitais. O ritmo intenso de crescimento da área de bibliotecas digitais e o reconhecimento da sua relevância comercial, estratégica e acadêmica refletiram-se no número de edições especiais dos mais importantes periódicos em ciência da informação e em ciência da computação, e também no número crescente de *workshops* e conferências acontecidos na década passada, estendendo-se aos dias atuais (URS, 2001).

Grande parte desse interesse foi alimentado pelo governo americano que em 1994, impelido pela repentina explosão de crescimento da Web e pelo desenvolvimento de navegadores gráficos (*Web browsers*), vislumbrou a oportunidade de estender os recursos e os serviços de bibliotecas além de seus limites físicos e além das suas comunidades, facilitando o compartilhamento de recursos informacionais escassos e alcançando públicos mal servidos por estes recursos (BROWN, 2005). Nessa direção, as agências americanas *National Aeronautics and Space Administration* (NASA), *Defense Advanced Research Projects Agency* (DARPA) e *National Science Foundation* (NSF) passaram a considerar as bibliotecas digitais como um dos focos principais do esforço de pesquisa em prol da Infra- estrutura Nacional de Informação (NII - *National Information Infrastructure*) – um plano amplo para interconectar indústria, governo, pesquisa, educação e cada lar através de redes de telecomunicações avançadas e de recursos e tecnologias de informação (McLOUGHLIN, 2000).

Traduzindo este interesse estratégico em apoio financeiro à pesquisa nas áreas circunscritas pelas bibliotecas digitais, essas instituições poderosas investiram o montante de 24,4 milhões de dólares na constituição de um programa multiagência, denominado *Digital Library Initiative* (DLI). O programa foi planejado para quatro anos (1994-1998), mas, devido principalmente aos bons resultados alcançados, ele se estendeu em uma segunda fase que se desenrolou no período de 1999 a 2004. Na primeira fase, conhecida pela sigla DLI-1, o programa colocou em foco a perspectiva de compartilhamento de informações. A ideia era "avançar dramaticamente nos meios de coletar, armazenar e organizar informação em forma digital, e

torná-la disponível para busca, recuperação e processamento via redes de comunicação – tudo isso de forma amigável para o usuário", conforme explicitado na página Web do programa[2]. O DLI-1, considerado o maior e o mais importante programa de pesquisa em bibliotecas digitais até hoje estruturado, contemplou seis grandes projetos de pesquisa sediados em diferentes universidades americanas, cada qual com características distintas em termos de conteúdos e tecnologias. Seus resultados formam o corpo de conhecimento que apoiou a consolidação e a operacionalização das principais iniciativas em escala mundial, incluindo as experiências no Brasil (MARCONDES; SAYÃO, 2001). Foram os seguintes os projetos iniciados com apoio do programa DLI-1:

- *Carnegie Mellon University – Informedia Digital Vídeo Library*[3];
- *University of Illinois at Urbana-Champaign – Federation Repositories of Scientific* Literature[4];
- *University of California at Berkeley Electronic – Environmental Planning and Geographical Information Systems*[5];
- *University of California at Santa Barbatra – Alexandria Digital Library Project: Spatially-referenced Map Information*[6];
- *University of Michigan Digital Library Project (*UMDL*) - Intelligent Agents for Information Location*[7];
- *University of Satanford Digital Library Project – Interoperation Mechanisms Among Heterogeneus Services*[8].

O sucesso do DLI-1 pode ser mensurado em termos dos avanços que proporcionou nas pesquisas e nas práticas de biblioteconomia digital e, não menos importante, no interesse gerado entre as comunidades acadêmicas, os formuladores da política de Ciência e Tecnologia e do público usuário em geral. O êxito do programa assegurou apoio contínuo e necessário para área, que foi traduzido, especialmente, pela instalação do DLI-2, um empreendimento de alcance ainda maior, envolvendo outros importantes patrocinadores como a *Library of Congress*, a *National Library*

2 Disponível em: <http://www.dli2.nsf.gov/dlione/. Acesso em: 1 set. 2008.
3 Disponível em: https://www.ri.cmu.edu/project/informedia-digital-video-library/. Acesso em: 31 jul. 2021.
4 Disponível em: http://dli.grainger.uiuc.edu/. Acesso em: 1 set. 2008.
5 Disponível em: http://elib.cs.berkeley.edu/. Acesso em: 1 set. 2008.
6 Disponível em: http://alexandria.sdc.ucsb.edu/. Acesso em: 1 set. 2008.
7 Disponível em: http://www.si.umich.edu/UMDL/. Acesso em: 1 set. 2008.
8 Disponível em: http://dbpubs.stanford.edu:8091/diglib/. Acesso em: 1 set. 2008.

of Medicine e o *National Endowment for the Humanities*. Tendo como diretriz a biblioteca digital como um sistema centrado no ser humano, o DLI-2 significou uma expansão em relação à primeira fase da iniciativa em todas as dimensões do seu escopo, refletindo o crescimento do número e da diversidade de agências e de interesses envolvidos. As intenções do DLI-2 ultrapassavam as fronteiras das especificidades das comunidades de computação e de comunicação e propunham incluir acadêmicos, médicos e estudantes, não somente de ciências e engenharia, mas também de artes e humanidades. Esse fato era fruto do reconhecimento de que avanços significantes nas áreas de tecnologia eram resultados de perspectivas, métodos e práticas de domínios não científicos (GRIFFIN, 1998).

Como parte do *Human Centered Systems* (HuCS) – programa que tinha como objetivo tornar os sistemas de computadores e redes de comunicação mais acessíveis e usáveis para todas as comunidades de usuários –, as expectativas em torno dos projetos da DLI-2 eram que envolvessem conteúdos em áreas temáticas que cobrissem todo o universo de interesse humano. Com essa perspectiva, o DLI-2 estabeleceu como ênfase a interoperabilidade e as tecnologias de integração, a gestão e o desenvolvimento de conteúdos e de coleções digitais, a infraestrutura operacional e de aplicações e a compreensão das bibliotecas digitais em domínios específicos e sua contextualização social, econômica e internacional. As pesquisas e práticas incluíam pontos como: a) tipos de mídias incluídas – som, música, dados econômicos, *software*, imagens, vídeos e material textual; b) diversidade de conteúdo, incluindo imagens e modelos antropológicos, manuscritos literários, prontuários médicos entre outros; c) exploração de novos recursos tecnológicos como aqueles voltados para a interoperabilidade, segurança, classificação automática etc (GRIFFIN, 1998; FOX, 1999).

Outra iniciativa importante – tanto do ponto vista tecnológico quanto histórico – é o projeto *Networked Computer Science Technical Report Library*, mais conhecido pela sigla NCSTRL (pronunciada como a palavra inglesa *ancestral*). O NCSTRL constitui uma rede de bibliotecas digitais distribuídas que provê acesso a documentos da área de ciência da computação. A importância dessa rede é conferida pela sua contribuição significativa para o desenvolvimento de tecnologias e ferramentas voltadas para bibliotecas digitais. A rede NCSTRL começou a operar no final de 1995, fruto da fusão de dois outros projetos: *Wide Area Technical Report Service* (Waters) e o *Dienst*, em cujo âmbito foram especificados dois importantes elementos para interoperabilidade de repositórios digitais: uma arquitetura conceitual aberta para bibliotecas digitais federadas e um protocolo para comunicação no domínio dessa arquitetura (DAVIS; LAGOZE, 2000).

O interesse pela área de bibliotecas digitais alimentado pelas iniciativas do governo americano determinou não somente uma evolução contínua da área, mas

despertou também a atenção de outros países. Os projetos americanos começaram a se expandir internacionalmente quando, em 1999, a *National Science Foundation* (NSC) fez uma aproximação do seu programa de pesquisa em bibliotecas digitais com as atividades similares na Inglaterra, capitaneadas pelo *U.K. Joint Information Systems Committee* (JISC). O resultado dessa colaboração foi JISC-NSF *International Digital Library Initiative*, um programa de três anos que tinha como objetivos imediatos: a) integrar coleções inacessíveis por barreiras técnicas, fragmentação, distância etc.; b) criar novas tecnologias orientadas para usuários distribuídos e; c) avaliar o impacto dessas novas tecnologias e seus benefícios em escala internacional (SUN MICROSYSTEM, 2002; WISEMAN; RUSBRIDGE; GRIFFIN, 1999).

O desdobramento imediato da propagação do interesse por outros países pela área de bibliotecas digitais foi a constituição de novos contextos, enfoques, práticas e visões que universalizaram e enriqueceram a área. Enquanto nos Estados Unidos as pesquisas estavam voltadas majoritariamente para a construção de bibliotecas digitais – como consequência, talvez, do grande envolvimento da comunidade de Ciência da Computação –, no Reino Unido, um outro cenário de crescimento e evolução se apresentava caracterizado por um comprometimento intenso da comunidade de Biblioteconomia e Ciência da Informação. Esse fato determinou uma ênfase na extensão dos serviços das bibliotecas tradicionais para as bibliotecas digitais. A Europa, como um todo, distinguia-se por um modelo diferente, focado no esforço de digitalização, desenvolvimento de coleções, preservação de materiais legados e questões relacionadas à linguagem (URS, 2001).

Dessa forma, parcialmente estimuladas pelas atividades americanas, as bibliotecas digitais na Europa começaram a se distinguir como um campo de pesquisa nos meados da década de 1990. Observa-se o surgimento de diversas iniciativas importantes em âmbito nacional, como por exemplo, o *eLib Programme*, no Reino Unido e o *Medoc Project*, na Alemanha. Essas iniciativas foram desdobramentos da terceira e quarta edições do *Framework Programme for Research and Technological Development*, da Comissão Européia, que estimulava a criação de projetos de bibliotecas digitais européias, especialmente no contexto do Programa *Telematic for Libraries* (1990-1998) (BORBINHA, 2007; DELOS, 2003). O *Telematic for Libraries* tinha o objetivo ambicioso de unificar o acesso a informações dos países europeus. Para tal, apoiou diversos projetos que convergiam para esse objetivo. O programa estava estruturado em torno de quatro linhas de ação complementares: bibliografias computadorizadas, redes de bibliotecas e interconexão de sistemas, serviços inovadores de bibliotecas e produtos e serviços de bibliotecas baseados em tecnologia. São alguns exemplos dos tipos de projetos desenvolvidos no âmbito do programa: *Controlled Access to Digital Libraries in Europe* (CANDLE), *Digitised European*

Periodicals (DIEPER), *Networked European Deposit Library* (NEDLIB) (TELEMATIC FOR LIBRARIES, [199-?]; LIU, 2005)

Em continuidade, a Comissão Europeia reconhecia a necessidade de apoiar a criação de uma comunidade europeia de pesquisa em biblioteca digital de caráter integrado. Por essa razão, a partir de 1997, já dentro do escopo do *Fifth Framework Program*, apoiou a fundação do DELOS: a *Network of Excellence on Digital Library* – que iniciou como um grupo de trabalho. O DELOS tem sido considerado um sucesso em estimular as atividades coordenadas de pesquisa na Europa e em promover a construção de expertise em bibliotecas digitais e em áreas correlatas, mantendo a pesquisa e o desenvolvimento em bibliotecas digitais na Europa em níveis globalmente competitivos (CANDELA *et al.*, 2007; THANOS, 2004).

Desde então, muitas outras organizações importantes se envolveram na expansão das tecnologias e práticas de bibliotecas digitais, incluindo a *European Union*, *Association for Computing Machinery* (ACM), o *Institute of Electrical and Electronics Engineers* (IEEE), a *International Federation of Library Association* (IFLA), a *American Library Association* (ALA), a *Coalition for Networked Information* (CNI) e a *Digital Library Federation* (DLF).

4 Afinal, o que é uma biblioteca digital?

Pelo que vimos até aqui, criou-se historicamente uma enorme expectativa em torno das potencialidades das bibliotecas digitais, não somente em termos de um novo paradigma de sistema de informações, de busca e recuperação, mas também como um recurso estratégico dentro de contextos altamente institucionalizados, como governo, educação, cidadania, negócios e pesquisa científica. O conceito de uma biblioteca digital meramente equivalente a uma coleção de objetos digitalizados assistida por uma ferramenta de gestão de informação torna-se tosco e já não cabe nas utopias desses inúmeros setores. A ideia de biblioteca digital como um "ambiente distribuído que integra coleções, serviços e pessoas na sustentação do ciclo de vida completo de criação, disseminação, uso e preservação de dados, informação e conhecimento" (DUGUID, 1997) – conforme preconizado pelo relatório final do *Santa Fé Planning Workshop on Distributed Knowledge Work Environments* –, esteja, talvez, mais próxima do que se almeja para bibliotecas digitais agora e num futuro possível.

Coerente com essa visão, a *Digital Library Federation* (DLF) estabelece na sua página Web, pensando menos numa formalização e mais numa definição operacional, que:

> Bibliotecas digitais são organizações que disponibilizam os recursos, incluindo pessoal especializado, para selecionar, estruturar, oferecer

> acesso intelectual, interpretar, distribuir, preservar a integridade e assegurar a persistência ao longo do tempo de coleções de trabalhos digitais, de forma que eles estejam prontamente e economicamente disponíveis para uso de uma comunidade definida ou um conjunto de comunidades (DLF, 2008, tradução nossa).

Essa definição tem sido adotada amplamente por grande parte das comunidades vinculadas às áreas de Biblioteconomia e de Ciência da Informação. Entretanto, ela revela apenas uma das muitas faces do que é universalmente discutido e entendido como biblioteca digital.

A complexidade das bibliotecas digitais em termos tecnológicos e organizacionais, somado ao seu universo vasto e variado de usuários e à multiplicidade de visões – reais e imaginárias – sobre as suas possibilidades e a sua extensão impactam significativamente a construção de uma definição comum. "Apesar das intensas atividades de pesquisa e de desenvolvimento em torno das várias vertentes do problema, não se tem absolutamente claro o significado do termo biblioteca digital" (HARTER, 1997, tradução nossa).

Passada mais de uma década, a afirmação de Harter continua sendo irritantemente verdadeira: biblioteca digital é uma ideia em movimento, ainda se desenvolvendo e tomando forma. "Nos estamos agora na adolescência das bibliotecas digitais", confirma Lagoze e seus colaboradores (2005, p. 1, tradução nossa) pensando nos motivos de preocupação e otimismo que essa fase turbulenta representa.

A impossibilidade de uma definição de consenso acontece por vários motivos, porém o mais importante deles é que o termo "biblioteca digital" é usado para denotar uns números extraordinários de coisas – de coleções pessoais até a Internet inteira. Na maioria das vezes essas coisas só têm em comum uma remota manipulação de recursos informacionais digitalizados (HARTER, 1997). Somam-se ainda o grande número de atores que contribuíram para o desenvolvimento e a implementação de bibliotecas digitais e aqueles que estão envolvidos profissionalmente no seu uso, além, é claro, do dinamismo próprio da ambientação tecnológica que sustenta essas bibliotecas. Biblioteca digital representa um espaço sinérgico de um grande número de áreas da Tecnologia da Informação e várias outras disciplinas e campos de pesquisa como Biblioteconomia, Ciência da Informação, Museologia, Arquivologia e Gestão do Conhecimento, para citar algumas das mais importantes (CANDELA *et al.*, 2007).

Além do mais, a busca por uma definição mais precisa e consensual para biblioteca digital esbarra também na existência de três termos – biblioteca digital, biblioteca eletrônica e biblioteca virtual – que possuem diferentes significados, mas que são usados frequentemente para designar a mesma coisa (SAUNDERS, 1996).

Dessa forma, a maioria das definições é fortemente influenciada pela percepção e pontos de vista particulares de pessoas e de organizações de diversas áreas que estiveram envolvidas em empreendimentos voltados para a construção e uso de bibliotecas digitais. A diversidade de contribuições que tanto serviu para o enriquecimento da área criou, ao mesmo tempo, uma zona obscura de indefinições.

Não obstante as indefinições sobre o termo, o conceito de biblioteca digital não é algo absolutamente novo. De fato, a ideia central que ele encerra precede o desenvolvimento do primeiro computador (BROWN, 2005). Segundo Harter (1997), o uso do termo "biblioteca digital" – que é o mais recente para denotar uma ideia quase ancestral – surge no decorrer do estabelecimento da primeira fase da *Digital Library Initiative* (DLI-1), em 1994.

> O termo foi rapidamente adotado pelos cientistas da computação, bibliotecários e outros. Assim, enquanto o termo "biblioteca digital" é relativamente novo, o trabalho de trazer recursos digitais de informação para as bibliotecas (ou pensar em recurso de informação digital como biblioteca) tem uma história que se estende por várias décadas (HARTER, 1997, p. 2, tradução nossa).

Tentando interpretar a diversidade de entendimento, Harter (1997) contrapõe as duas visões extremas sobre a natureza das bibliotecas digitais: uma visão abrangente que toma a biblioteca digital tal como a Web é hoje – anárquica e individualista; e uma visão que toma a biblioteca digital como uma metáfora, ou mesmo uma extensão, da biblioteca tradicional. No espaço entre esses limites são discutidas as diferenças essenciais: propriedades de localização física, de conteúdo, de critérios de seleção, de organização, controle de autoridades, de autoria, de acesso, de grupos de usuários alvo, de serviços, de taxação e de fixidade – conceito que está relacionado com a integridade e a segurança dos conteúdos e suas propriedades de permanência.

Num extremo está a "googlização" das bibliotecas digitais, referindo-se à incômoda e errônea concepção de que o Google representa a apoteose da informação digital e que os problemas existentes nesse domínio já foram resolvidos ou serão resolvidos por esse serviço ou por outra ferramenta semelhante. Esse estreitamento das discussões conduz à visão míope de que a biblioteca digital está limitada à busca e ao acesso – funções essenciais (e ainda desafiadoras), mas que são somente parte do ambiente informacional circunscrito pela ideia plena de biblioteca, seja ela imaginária ou real (LAGOZE *et al.*, 2005). Essa visão está turvada pelo fato de mais e mais pessoas estarem usando a Internet como a prin-

cipal fonte de informação. De fato a Internet tem sido referida por muitos como "uma vasta biblioteca, contendo todo o tipo de informação conhecida pelos seres humanos" (WALLACE, 1999). Entretanto, essa constatação não pode ser ignorada como elemento de compreensão do seu contrário, pois diferentemente das bibliotecas tradicionais onde as fontes de informação adicionadas às coleções são cuidadosamente selecionadas, organizadas e descritas – classificadas, catalogadas, indexadas, resumidas – isso não acontece com frequência nas coleções encontradas na Internet. Porém, a infraestrutura oferecida pela Internet é um veículo de dramática importância para a distribuição de informação de qualidade para os usuários, e é parte essencial da infraestrutura tecnológica que as bibliotecas digitais não podem prescindir. No outro extremo, observa-se uma tendência convergente na direção do enquadramento das bibliotecas digitais aos cânones biblioteconômicos, principalmente no que concerne à organização e à representação dos recursos informacionais e também às relações orgânicas com suas comunidades-alvo. Isso parece indicar que as bibliotecas digitais devem se equiparar às bibliotecas tradicionais, ao mesmo tempo em que criam condições técnicas para expandir os limites, as formulações e o alcance espacial e temporal do que sempre conhecemos como biblioteca. Entretanto, é importante assinalar que vai ficando cada vez mais nítido que essa visão expandida de biblioteca exige novas reflexões sobre os modelos de informação e de serviços sobre os quais elas estarão baseadas.

Essa convergência para a Biblioteconomia pode ser justificada de várias maneiras, porém a mais convincente delas é também a mais óbvia: biblioteca digital continua sendo biblioteca.

> O progresso tecnológico mudou a maneira *como* as bibliotecas fazem o seu trabalho, mas não a *razão* do seu trabalho. Ainda que os desenvolvimentos tecnológicos mais contundentes – como a conexão de um computador a outro numa cadeia continua pelo mundo afora – possam alterar o conceito fundamental de biblioteca no século 21, podemos supor que a tecnologia não vai mudar substancialmente o negócio das bibliotecas que é conectar pessoas com informações (KUNY; CLEVELAND, 1998, p. 1, tradução nossa).

É imprescindível compreender que a tecnologia atual está focada na conversão de papel para formatos digitais e não na conversão da biblioteca *in toto* para formatos digitais (BROWN, 2005). Assim como uma biblioteca de áudio-visual ou de microfilmes continua sendo uma biblioteca, o conceito atual de biblioteca

digital constitui um subconjunto de um conceito mais extenso de biblioteca, e não um substituto para ele. Todos os valores e funções da biblioteca continuam válidos, o que muda são os objetos físicos que formam a biblioteca, e, naturalmente, o instrumental tecnológico para manipulá-los. As mídias digitais devem ser vistas como um novo suporte na longa lista de materiais sobre os quais a civilização tem continuamente utilizado para registrar e transmitir o conhecimento para gerações futuras. Como os outros materiais, nós podemos esperar que eles sejam utilizados na proporção em que a sua disponibilidade local, as tecnologias de apoio, seu custo e a sua confiabilidade sejam adequados e suficientes para armazenar e disseminar informação e conhecimento de acordo com as exigências do seu tempo.

"Adicionando o adjetivo "digital" ao nome "biblioteca" o futuro parece estar reconciliado com o passado" (LYMAN, 1996, p.1). Alegorias futurísticas como bibliotecas digitais e publicações eletrônicas são tranquilizadoras porque elas sugerem uma continuidade institucional entre o passado e o futuro. Pois, se é verdade que a inovação tecnológica geralmente começa imitando o passado, não são novas ferramentas que constituem inovação, mas sim novas instituições. "Elas acalmam e ocultam a tensão latente que existe entre tecnologia digital e as instituições de uma sociedade industrial, tensões que levam a questões importantes sobre a natureza das bibliotecas digitais" (LYMAN, 1996, p. 1, tradução nossa). Em outras palavras, bibliotecas digitais parecem oferecer-nos toda a conveniência, eficiência, a sofisticação da tecnologia digital dentro da ideia familiar e confortável de uma biblioteca (MCPHERSON, 1997).

5 Por que bibliotecas digitais?

Logo no título de um de seus artigos, Michael Lesk faz a pergunta primordial: "Por que bibliotecas digitais?" (*Why digital libraries*?). A resposta, simples e direta, vem em seguida.

Existem muitas razões para que as bibliotecas digitais sejam algo desejável. Elas podem tornar as pesquisas mais fáceis para os acadêmicos. Elas podem aliviar a pressão orçamentária sobre as bibliotecas. Elas podem resolver nosso problema urgente e crescente de preservação, ou elas podem ajudar as bibliotecas a estender as coleções para novas mídias. Mas, talvez, a maior vantagem das bibliotecas digitais seja a capacidade de ajudar a sociedade a tornar a informação mais disponível, melhorando a sua qualidade e aumentando a sua diversidade. As bibliotecas digitais podem desempenhar este papel? Isso vai depender de como nós financiamos, regulamos e gerenciamos as bibliotecas digitais e a nova infraestrutura de comunicação e as novas tecnologias que as impulsionam (LESK, 1995, p. 1, tradução nossa).

Pensando nas justificativas para as bibliotecas digitais além da agregação de valores significante e sem paralelo aos serviços de biblioteca, verificamos que elas caminham rapidamente para se tornar um ponto concentrador de tecnologias e metodologias voltadas para o apoio à pesquisa e à comunicação científica, às diversas modalidades de ensino e à disseminação de informações, de toda a natureza, para o cidadão comum.

> As bibliotecas digitais representam uma nova infraestrutura e ambientação que tem sido concretizada por vários fatores, principalmente a integração e uso de um conjunto de tecnologias de informação e de comunicação, disponibilidade de conteúdos digitais em escala global e uma forte demanda por parte de usuários online. As bibliotecas digitais estão destinadas a se tornarem uma parte essencial da infraestrutura de informação do século 21 (THANOS, 2004, p. 1, tradução nossa).

Essa visão abrangente está expressa em vários documentos importantes. Alguns deles delineiam visões estratégicas advindas de setores governamentais no exercício de prospectar tecnologias-chave e transformações que nortearão este século que ainda se inicia.

Esse parece ser o caso do relatório publicado em 2001 pelo PITAC – sigla para US *President´s Information Technology Advisory Committee* – sobre bibliotecas digitais. O relatório recebeu um título que não deixa dúvida sobre o seu conteúdo e a ênfase que os conselheiros do, até hoje, Presidente George W. Bush queriam lhe transmitir: "*Digital Libraries: Universal Access to Human Knowledge*". Os conselheiros identificam um conjunto de "Transformações Nacionais Desafiadoras", itens que seriam pré-requisitos essenciais para capacitar todos os cidadãos no contexto da sua sociedade a participar e usufruir dos benefícios da Era da Informação. O PITAC reconhece que as transformações apontadas são desafios cruciais que não podem prescindir dos avanços das tecnologias de bibliotecas digitais (PITAC, 2001). "Nós estamos especialmente satisfeitos em remeter este relatório [...] pela profunda relevância dessa tecnologia para o avanço da qualidade da educação em cada escola, em cada centro de aprendizagem e em cada lar no país" (PITAC, 2001, p. 1, tradução nossa), dizia a carta de encaminhamento do relatório ao Presidente.

Para os europeus, que têm como riqueza a diversidade cultural e linguística, a ideia de integração e acesso notadamente prevalece. Nessa direção, a Comunidade Européia reconhece a necessidade de estimular a criação de uma biblioteca digital

europeia integrada voltada para a comunidade de pesquisa. Essa é a razão para a criação do DELOS[9], cuja visão de longo prazo é que as

> bibliotecas digitais devem capacitar qualquer cidadão acessar todo o conhecimento humano a qualquer momento e em qualquer lugar, de uma forma amigável, de várias maneiras, de forma efetiva e eficiente, rompendo as barreiras da distância, da linguagem e culturais. Utilizando para tal múltiplos dispositivos conectados via Internet (DELOS, 2003, tradução nossa).

Para a DELOS (2003) as novas gerações de bibliotecas digitais não devem ser consideradas como meros repositórios de informações estáticas. Antes disso, elas devem ser reconhecidas como núcleo inicial do que, num estágio futuro, constituirá uma parte substancial do conhecimento humano (THANOS, 2004).

As razões para se criar um forte ordenamento e a institucionalização das pesquisas em bibliotecas digitais são da mesma natureza que o imaginário utópico que estimula a reinvenção das bibliotecas totais sob óticas distintas através da história e da ficção literária. Muito se espera dessa nova formulação de biblioteca que chamamos hoje de biblioteca digital. Entretanto, surge aqui o mesmo problema identificado na definição de bibliotecas digitais discutidos anteriormente: cada uma das comunidades envolvidas no desenvolvimento e/ou no uso das bibliotecas digitais têm pontos de vista e expectativas diferentes em relação a elas. O ambiente de serviços de biblioteca digital é um espaço de informações eletrônicas que suporta visões altamente diferenciadas e uma gama extraordinária de usos para o seu universo de informações em rede (GREENSTEIN, 2002). Como ilustração dessa pluralidade de visões e possibilidades de uso, segue uma breve análise, baseada em artigo de Urs (2001), sobre a ótica dos cientistas da informação e bibliotecários, cientistas da computação, arquivistas, políticos e governantes, editores, educadores e professores, comunidades da área cultural e do comércio eletrônico.

A comunidade de Biblioteconomia e Ciência da Informação visualiza a biblioteca digital menos como um sistema de computação – uma máquina – e mais como uma instituição; como uma extensão lógica do que as bibliotecas vêm fazendo desde os tempos imemoriais, ou seja, adquirindo, organizando e disseminando conhecimento usando as tecnologias correntes. O que o bibliotecário deseja é a ampliação dos recursos e dos serviços disponíveis e também a audiência das bibliotecas. Na sua perspectiva prática, o acesso simultâneo a um mesmo documento di-

9 Disponível em: www.delosinfo.com.br. Acesso em: 1 set. 2008.

gital por um número indefinido de usuários significa o fim da lista de empréstimo. Para ele a biblioteca digital é um estágio a mais no desenvolvimento contínuo de novos meios de publicação – em que a biblioteca soma a responsabilidade de também ser uma publicadora Web –, bem como uma nova infraestrutura tecnológica e organizacional voltada para potencializar a sua missão de disseminar informação e conhecimento. Porém, enquanto os profissionais de informação têm uma perspectiva de continuidade evolutiva em relação às bibliotecas digitais, outras visões importantes se sobrepõem.

Os profissionais da área de Ciência da Computação enxergam as bibliotecas digitais como uma extensão dos sistemas de computadores em rede – um sistema que oferece facilidades informacionais. Essas visões se fragmentam à medida que se analisa com um grau a mais de detalhes as diferentes áreas que compõem o domínio da Ciência da Computação. Por exemplo, enquanto os pesquisadores da área de Recuperação da Informação (RI) veem as bibliotecas digitais como uma ampliação dos sistemas de recuperação de informação em que os documentos e sua representação (ou descrição) são diferentes da RI tradicional, quem trabalha com sistemas multimídia considera as bibliotecas digitais uma aplicação dessa tecnologias; para pesquisadores da área de base de dados, a biblioteca digital é tão somente uma ampla base de dados.

Apesar das controvérsias apaixonadas, a maioria dos políticos e governantes percebe a biblioteca digital como parte da infraestrutura tecnológica necessária para a superação da desigualdade informacional e de acesso, e como mais um recurso para apoio dos programas de inclusão digital. Consideram, com maior ênfase, a biblioteca digital como um insumo básico para a pesquisa, o ensino superior e a pós-graduação e como um instrumento para a maior visibilidade de bens e instituições culturais. Os governantes, com intensidade variável, têm investido em infraestrutura computacional e de redes que beneficiam diretamente as iniciativas na área de bibliotecas digitais. Como já vimos, grande parte dos projetos mais relevantes são iniciativas do poder público, financiados por suas agências e, não raro, apoiado por segmentos da iniciativa privada interessada em expandir suas áreas de atuação.

Mesmo considerando as mudanças atuais nos papéis de autor, editor e outros atores e nos limites entre eles, proporcionados principalmente pelos avanços da Internet, os editores, desde a revolução de Gutenberg, têm continuamente desempenhado um papel fundamental na facilitação da produção e distribuição de informação. A percepção da indústria editorial em relação à nova mídia representada pelas bibliotecas digitais é ambivalente: em contrapartida às novas oportunidades mercadológicas existem as ameaças representadas pelas novas formas de autopu-

blicação e o movimento crescente em torno do acesso livre, o que exige uma adaptação permanente à um meio que se renova constantemente. Numa visão otimista, para o mundo editorial, a biblioteca digital constitui um novo modo de distribuição de conteúdos e um novo mercado – bastante competitivo – a ser conquistado, num contexto de mudança da economia da informação. Para isso os editores estão se adaptando ao paradigma da publicação eletrônica, integrando mídias, criando novos modelos de negócio, como os portais agregadores, e estabelecendo parcerias com organizações mais próximas ao mundo Internet.

Para os educadores e os professores que sempre tiveram uma relação de colaboração quase que simbiótica com as bibliotecas tradicionais, as bibliotecas digitais podem ser um meio de ampliar essa relação clássica. Para eles as bibliotecas digitais constituem um novo recurso de aprendizado, apoiados por conteúdos multimídia, interatividade e integração de informações heterogêneas que o ensino e, particularmente, o ensino à distância não pode prescindir. As bibliotecas digitais abrem possibilidades extraordinárias para a educação e o ensino, mudando paradigmas e estabelecendo novas metodologias pedagógicas. São as áreas que mais podem se beneficiar dessa nova tecnologia.

Para os arquivistas, as bibliotecas digitais rompem com a relação quase antagônica entre a preservação e o acesso existente no mundo do papel e dos demais materiais analógicos (sayão, 2005). Isso acontece na medida em que a digitalização se torna um meio de preservar os conteúdos raros, únicos ou frágeis, ao mesmo tempo em que proporcionam acesso universal a suas representações digitais através das bibliotecas e arquivos digitais. A digitalização é vista pelos arquivistas como uma alternativa à microfilmagem tradicional com a ressalva dos problemas de integridade e confiabilidade dos conteúdos digitais, ou seja, do seu valor de prova e de sua preservação de longo prazo que é uma preocupação constante de toda a comunidade arquivística.

Para os pesquisadores, a colaboração é a chave para a pesquisa e o desenvolvimento, nesse sentido eles percebem a biblioteca digital como um espaço dinâmico voltado para a geração, o compartilhamento e a disseminação de conhecimento. Através das bibliotecas digitais, os dados de pesquisa agora podem ser acessados em escala planetária pelos pesquisadores interessados. Essa característica é de grande importância para o surgimento do conceito de "colaboratórios" – resultado da contração das palavras "colaboração" e "laboratório", significando um centro de pesquisa sem paredes onde os pesquisadores interagem entre si eletronicamente no desenvolvimento de projetos inovadores. Projetos como Genoma Humano, baseados em compartilhamento internacional de dados de pesquisa e análises, são exemplos significantes da ideia de um colaboratório .

Ainda há a perspectiva da biblioteca digital enquanto forma de apropriação do mundo da informação pelo comércio eletrônico. Para as organizações comerciais, as bibliotecas digitais estabelecem um novo mercado global, constituindo, para alguns autores, um caso específico de economia da informação (SCHÄUBLE; SMEATON, 1998). Um dado importante é que os desenvolvedores de bibliotecas digitais têm deliberadamente incorporado modelos econômicos e de preços nas arquiteturas de bibliotecas digitais.

No campo cultural, o que se observa é que a biblioteca digital é um meio privilegiado de dar visibilidade global a manifestações culturais antes circunscritas às suas comunidades e sem canais de comunicação para fora delas. O desenvolvimento de metodologias e técnicas para recuperação multilingue de informação somado ao desenvolvimento de recursos linguísticos para serem acoplados às bibliotecas digitais vai ajudar as comunidades que se expressam em outros idiomas que não o inglês a superarem as barreiras linguísticas no acesso e na disseminação de informações.

6 Problemas e desafios

Como parte de uma matriz complexa de serviços de informação baseada em rede de computadores, espera-se que as bibliotecas digitais estabeleçam uma ampla estrutura de intermediação entre recursos informacionais heterogêneos e distribuídos e as comunidades de usuários, um universo tão amplo, diversificado e mutante quanto são os interesses humanos. Para cumprir as expectativas e o que se planeja para a futura geração de bibliotecas digitais, um conjunto de desafios deve ser superado pela pesquisa e inovação que se estendem por várias áreas de conhecimento.

6.1 Arquitetura para bibliotecas digitais

Uma exigência imprescindível para as novas bibliotecas digitais é o desenvolvimento de uma arquitetura, que se constitua numa infraestrutura comum, que possa ser customizada segundo as necessidades de diferentes setores e aplicações. Essa infraestrutura tem que apoiar o estado da arte e também os modelos e técnicas inovadores que irão surgir; tem que ser altamente customizável, configurável e adaptativa, refletindo a diversidade de aplicações que se espera para as bibliotecas digitais.

6.2 Desenvolvimento de coleção digital

As bibliotecas, ao longo do tempo, têm coletado informações publicadas em vários formatos – livros, periódicos, CD-ROM, fitas de áudio e de vídeos e discos.

Nos últimos anos, a esse conjunto crescente de mídias as bibliotecas estão crescentemente incorporando repositórios de informações digitais. Vias de regra, as bibliotecas não estão substituindo mídias analógicas por mídias digitais, mas estão coletando-as também em complementação as mídias já estabelecidas (KUNY; CLEVELAND, 1998).

Processos tradicionais desempenhados pelas bibliotecas, tais como desenvolvimento de coleções e referência, embora formem uma base potencial para o funcionamento da biblioteca digital, devem ser revisados para acomodar as diferenças determinadas pela natureza digital dos recursos informacionais. O desenvolvimento de coleções digitais compreende todos os problemas da formação e gestão de coleções convencionais, como políticas e estratégias de seleção e aquisição. Porém, compreende também os problemas decorrentes da condição digital da informação, como a conversão de material impresso para digital, a geração de material unicamente digital, as barreiras tecnológicas que impedem o acesso e a usabilidade dos objetos, a sustentabilidade das coleções digitais, a gestão de direito, a criação e novos gêneros de objetos digitais e, naturalmente, a preservação digital.

Entretanto, o maior desafio que se impõe à formação das coleções digitais é a integração dos diversos tipos e formatos de objetos digitais que temos atualmente – e dos novos objetos que cotidianamente vão aparecendo – com os materiais tradicionais, oferecendo uma visão coerente de todo o acervo. No contexto desse problema surge a ideia de "coerência digital", significando que todos os objetos numa biblioteca digital, sejam eles registros sonoros, imagens, texto, vídeo ou qualquer outro, podem ser tratados essencialmente da mesma forma. Essa forma de tratamento é diferente do que normalmente é praticado pelas bibliotecas, em que cada mídia recebe um tratamento diferente, por exemplo, biblioteca de fitas de vídeo. Em outras palavras: coerência digital é o mecanismo que permite uma forma de equalização entre vários recursos informacionais no ambiente de uma biblioteca digital. Essa equalização no tratamento é um desafio a ser superado de grande importância na distribuição e integração das informações (BROWN, 2005).

6.3 Metadados

Apesar de ser um conceito familiar para as bibliotecas – posto que uma das suas atividades básicas é a criação de catálogos descrevendo documentos –, metadado é uma questão crucial no desenvolvimento de bibliotecas digitais. No ambiente de uma biblioteca digital, os objetos digitais são descritos, estruturados, resumidos, identificados, gerenciados, preservados e suas representações manipuladas por meio de uso de metadados; os metadados também são imprescindíveis na descoberta de recursos e na utilização dos documentos digitais. Portanto, as coleções

digitais exigem esquemas de metadados bem estruturados que sejam capazes de descrever os objetos digitais e seus conteúdos em diversos níveis de grunularidade – de uma coleção como um todo até uma ilustração em um livro. Um dos maiores desafios com relação aos metadados é a diversidade de formatos de informação digital e a maneira como eles devem ser descritos no contexto de diferentes coleções dirigidas a diferentes públicos-alvo. Isso leva à questão de mapeamento entre diferentes esquemas de metadados constituir um dos problemas mais interessantes da área, especialmente, no que concerne à interoperabilidade entre bibliotecas digitais (SHIRI, 2003).

6.4 Interoperabilidade

As várias bibliotecas digitais são desenvolvidas segundo diferentes arquiteturas e tecnologias, são gerenciadas por organizações distintas, submetidas a diferentes padrões de qualidade. Esse ambiente distribuído e heterogêneo introduz um alto grau de complexidade na conquista de uma visão integrada das coleções digitais. A complexidade aumenta ainda mais quando consideramos que cada coleção é caracterizada pela diversidade de conteúdos informacionais, representados por vocabulários específicos em termos de metadados e por formatos de apresentação próprios.

A exigência imprescindível por algum grau de interoperabilidade entre as bibliotecas digitais decorre do fato de que grande parte das aplicações mais sofisticadas que toda a sociedade espera dessas bibliotecas, especialmente as áreas de ensino, de pesquisa e cultural, depende da interação efetiva entre as diversas bibliotecas e o fornecimento de uma visão unificada das informações ao usuário como resultado de uma operação de busca. O desafio da interoperabilidade é caracterizado pela multiplicidade de facetas que ela possui: interoperabilidade técnica, interoperabilidade semântica, interoperabilidade política e humana e muitas outras. As soluções em pauta passam quase sempre pela aplicação de padrões e protocolos comuns e pelos arranjos sociais e organizacionais que só podem ser estabelecidos pela cooperação e pelo consenso (SAYÃO; MARCONDES, 2008).

6.5 Interfaces e usabilidade

O desenvolvimento de interfaces inovadoras para bibliotecas digitais constitui uma linha de pesquisa bastante explorada. As bibliotecas digitais se dirigem para diferentes contextos e audiências – ambientes acadêmicos, escolas, governo, negócios –, portanto é necessário que elas se reconfigurem de acordo com a familiaridade, habilidades, faixa etária e percepções de cada um dos segmentos de usuários. Essa área pode incluir ainda as questões de usabilidade e questões comportamen-

tais, compreendendo interação com as bibliotecas digitais, acessibilidade, aceitação por parte do usuário, interação homem-computador, entre outras.

6.6 Descoberta de recursos

A informação digital publicada na Internet é caracterizada pelo fato de que os documentos digitais podem existir em várias formas, possivelmente em várias versões e instâncias, identificados por esquemas frágeis e localizados em endereços pouco fixados. Isso torna o recurso volúvel e transitório, criando sérios obstáculos aos processos de descoberta de recursos digitais. Os serviços de indexação e busca genéricos tais como Google, Yahoo e outros oferecem ferramentas básicas que ajudam o usuário a achar a informação que procura. Entretanto, esses serviços não têm o nível de especificidade, de desempenho e, sobretudo, de tratamento biblioteconômico exigido para a maioria dos empreendimentos. Além do mais, a qualidade das informações recuperadas pode se diluir no mar de resultados irrelevantes e de indesejáveis duplicações (KUNY; CLEVELAND, 1998).

Nessa nova etapa estabelecida pelas bibliotecas digitais, os processos de descoberta de recursos não podem prescindir das metodologias de organização de conhecimento – num sentido mais geral, conjunto de ferramentas usadas para ordenamento, classificação e recuperação de conhecimento –, e das tecnologias semânticas. Na pesquisa por metodologias para a realização de busca integrada entre bibliotecas digitais heterogêneas, um dos desafios importantes é o mapeamento e a interoperabilidade entre vários sistemas de organização de conhecimento (SAYÃO; MARCONDES, 2008).

6.7 Preservação

As coleções impressas podem sobreviver inercialmente por anos, armazenadas com pouco ou nenhum controle. Esses recursos permanecem viáveis, ou seja, legíveis e interpretáveis inercialmente, décadas depois. Mas não é esse o caso com os equivalentes digitais. Manter os conteúdos digitais viáveis para uso de futuras gerações requer um esforço intencional e um monitoramento e investimentos contínuos. Isso acontece porque a informação digital depende, na sua mais pura essência, de um aparato tecnológico para ser acessada e, sobretudo, corretamente interpretada. Mas esse aparato tecnológico de intermediação – formado por *hardware*, *software*, mídias e formatos – está em constante mutação, em ciclos de obsolescência cada vez mais rápidos, determinados principalmente pelo dueto inovação e competição. Contribui ainda grandemente para esse problema o fato dos meios de armazenamento serem muito frágeis e extremamente suscetíveis à degradação física.

A preservação digital não é uma ação fixada no tempo, é um processo que se desenrola indefinidamente. Além dos desafios técnicos representados pelas estratégias, procedimentos e padrões voltados para a preservação, é necessário pensá-la também como um desafio gerencial e organizacional (SAYÃO, 2005).

6.8 Gestão de direitos autorais

As leis de direitos autorais (*copyright*) constituem um instrumento de equilíbrio entre os interesses do criador e as obrigações da sociedade de facilitar o livre fluxo de informação, salvaguardando o interesse privado e o interesse público. Entretanto, no ambiente digital, as regras atuais de *copyright* falham porque o controle de cópias foi perdido, os objetos digitais são menos fixados, são facilmente copiados e acessíveis remotamente e simultaneamente por muitos usuários em escala mundial.

Há um consenso absoluto por parte de toda a comunidade envolvida de que a gestão de direitos é um dos mais complexos e desafiadores problemas que a área de bibliotecas digitais tem que enfrentar. Discutir direitos conduz forçosamente para os territórios legal e de negócios, os quais as bibliotecas, cuidadosamente, procuraram evitar no passado (COYLE, 2004a; COYLE, 2004b).

O direito autoral é considerado uma das barreiras mais relevantes no desenvolvimento das bibliotecas digitais. Isso porque as bibliotecas são, na maioria dos casos, simplesmente custodiantes da informação e não detêm os direitos autorais sobre o material que está sob o seu controle. É improvável, portanto, que bibliotecas possam livremente digitalizar e prover acesso a materiais detentores de *copyright* da sua coleção. Ao invés disso terão que desenvolver mecanismos para gerenciar esses direitos, procedimentos que permitam que elas disponibilizem informação sem violar as regras do direito autoral e da propriedade intelectual – estes procedimentos são chamados coletivamente de gestão de direitos autorais.

É necessário caminhar na direção da ampliação dos modelos atuais de gestão de direitos quando for possível e desenvolver novos modelos que preservem os conceitos de *fair use* e da *first sale doctrine*, essenciais para o desenvolvimento no contexto acadêmico; desenvolver modelos automatizados de *Digital Right Management* (DRM) que considerem além dos direitos dos proprietários de materiais protegidos por *copyright*, os direitos de acesso individuais e institucionais dos usuários, preservando privacidade dos mesmos (SAYÃO; MARCONDES, 2008).

6.9 Personalização

Na medida em que as bibliotecas digitais se tornam mais universais e seus serviços e conteúdos mais diversificados, os seus usuários – cada vez mais experien-

tes – esperam serviços mais sofisticados e mais talhados às suas necessidades e às suas habilidades, e que considerem também os seus direitos de acesso, tanto individualmente, como na qualidade de membros de uma ou mais comunidades e/ou organizações.

As buscas tradicionais, que são módulos comuns para todas as bibliotecas, não fazem frente à complexidade crescente das necessidades dos usuários e ao volume exponencial de informações que devem ser gerenciadas. As bibliotecas digitais precisam se deslocar, num futuro próximo, do atual estado passivo, onde oferece um grau mínimo de adaptação aos seus usuários, para um estágio mais proativo e dinâmico nos processos de entrega e de conformação da informação para usuários individuais e grupos de usuário e no apoio ao esforço de comunidades em capturar, estruturar e compartilhar conhecimento (CALLAN; SMEATON, 2003).

Muitas outras questões importantes para o pleno desenvolvimento das bibliotecas digitais estão assinaladas nas agendas de pesquisa da área. Uma parte considerável desses itens não são questões de fundo tecnológico, como modelos econômicos e sustentabilidade e administração e gestão de bibliotecas digitais.

7 À guisa de conclusão

O desejo antiquíssimo da humanidade de construir uma memória total que está expresso reiteradamente na literatura, às vezes na forma de um personagem coadjuvante, mas muitas vezes como um protagonista, ou nos projetos de inúmeras mentes visionárias – parece finalmente tomar o caminho da concretização. São muitos os fatores que contribuem para que esse fato ficcional e histórico se torne um fenômeno do nosso tempo. Mas a conjunção das tecnologias de informação e comunicação e a diminuição drástica dos custos de criação, armazenamento online, manipulação e transmissão de conteúdos digitais, combinados com o fenômeno da convergência de todos os tipos de mídias digitais, foram e continuam sendo fatores determinantes para o estabelecimento de uma infraestrutura tecnológica propícia ao surgimento de um conceito, ao mesmo tempo inovador e tradicional, de biblioteca.

Pela primeira vez é possível construir serviços em larga escala onde coleções de informações são armazenadas em formatos digitais, distribuídas em escala mundial e recuperadas através de redes de computadores por usuários através de computadores pessoais, nas suas casas ou escritórios, ou onde houver uma rede disponível, através de equipamentos móveis: *notebooks*, *palm-tops*, telefones celulares e tudo mais que o futuro permitir. As bibliotecas digitais cumprem a utopia ancestral das bibliotecas totais integrando globalmente repositórios multilingues e multiculturais de dados, informações e conhecimento de toda natureza, dirigido

a um universo de usuários igualmente diversificado, sem que para isso os seus recursos informacionais estejam guardados em um único lugar e sem os limites do tempo e do espaço.

São muitas as expectativas em torno das bibliotecas digitais, um exemplo recorrente é o acesso universal aos objetos únicos, raros, frágeis e remotos como os Manuscritos do Mar Morto, ou a um exemplar da Bíblia de Gutenberg, ou o retrato de Mona Lisa, através de representações digitais perfeitas. O fato de milhares de usuários poderem, ao mesmo tempo, acessar o mesmo recurso é, por si só, uma revolução sem precedentes. Porém, as potencialidades das bibliotecas digitais não estão restritas somente à busca e ao acesso à informação. Elas oferecem também um ambiente completo para administração, curadoria, comercialização, preservação, geração de aplicações que promovem e asseguram o uso adequado de suas coleções. Por exemplo, a reordenação e o reuso de conteúdos digitais oferecem oportunidades extraordinárias para a criação de serviços inovadores na área da educação, da arte, da cultura e dos negócios e, sobretudo, da pesquisa científica.

Refletimos apenas sobre a face mais visível das atuais aplicações de bibliotecas digitais. Porém, um largo espectro de aplicações potencialmente possíveis num futuro próximo – que serão viabilizadas por um grau crescente de integração de novas tecnologias, inovação, padronização e de disponibilização exponencial de conteúdos de qualidade e o equacionamento de problemas críticos como *copyright* e preservação digital – cria um quadro de otimismo justificado para que as bibliotecas digitais se tornem uma parte essencial da infraestrutura mundial de informação.

Referências

ARAUJO, Rute. **Estudos sobre multimídia**. [200-?] Disponível em: http://www.citi.pt/estudos_multi/. Acesso em: 30 maio 2008.

ARMS, William Y. *et al*. The design of the Mercury Electronic Library. **EDUCOM Review**, v. 27, n. 6, p. 38-41, nov./dec. 1992.

BORBINHA, José. Bibliotecas, arquivos e outras coisas digitais. *In*: CONGRESSO NACIONAL DE BIBLIOTECÁRIOS, ARQUIVISTAS E DOCUMENTALISTAS, 9., 2007,

Ponta Delgada. **Anais...** Disponível em: https://www.bad.pt/publicacoes/index.php/congressosbad/article/view/564. Acesso em: 28 jul. 2021.

BORGES, Jorge Luis. La biblioteca de babel. *In*: BORGES, Jorge Luis. **El jardín de senderos que se bifurcan**. Buenos Aires: Editorial Sur, 1941.

BORGES, Jorge Luis. Funes el memorioso. *In*: BORGES, Jorge Luis. **Ficciones**. Buenos Aires : Editorial Sur, 1944.

BORGES, Jorge Luis. **El Aleph**. Buenos Aires: Emecé Editores, 1949.
BROWN, Mary E. **History and definition of digital libraries**. New Haven, C.T.: Southern Connecticut State University, 2005. Disponível em: www.southernct.edu/~brownm/dl_history.html. Acesso em: 30 mar. 2008.
BUSH, Vanevar. As we may think. **Atlantic Montly**. 1945. Disponível em: http://www.theatlantic.com/doc/194507/bush. Acesso em: 28 jul. 2021.
CALLAN, Jamie; SMEATON, Alan. **Personalisation and Recommender Systems in Digital Libraries**: Joint NSF-EU DELOS Working Group Report. 2003. Disponível em: http://www.ercim.org/publication/ws-proceedings/Delos-NSF/Personalisation.pdf. Acesso em: 28 jul. 2021.
CANDELA, Leonardo *et al*. Setting the foundation of digital libraries. **D-Lib Magazine**, v. 13, n. 3/4, Mar./Apr. 2007. Diponível em: http://www.dlib.org/dlib/march07/castelli/03castelli.html. Acesso em: 28 jul. 2021.
COYLE, Karen. The "Rights" in the Digital Rights Management. **D-Lib Magazine**, v. 10, n. 9, September 2004a. Disponível em: http://www.dlib.org/dlib/september04/coyle/09coyle.html. Acesso em: 28 jul. 2021.
COYLE, Karen. Rights Management and Digital Library Requirements. **Ariadne**, n. 40, July 2004b. Disponível em: http://www.ariadne.ac.uk/issue40/coyle. Acesso em: 28 jul. 2021.
DAVIS, James R.; LAGOZE, Carl. NCSTRL: Design and deployment of a globally distributed digital library. **Journal of American Society for Information Science**, v. 51, n. 3, p. 273-280. 2000. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/(SICI)1097-4571(2000)51:3%3C273::AID-ASI6%3E3.0.CO;2-6. Acesso em: 28 jul. 2021.
DELOS. **DELOS Summary and objectives**. 2003. Disponível em: http://delos-old.isti.cnr.it/workdocs/ta_4_5_ist_relevance.pdf. Acesso em: 20 maio 2008.
DFL. **About**. Disponível em: http://www.diglib.org/about/dldefinition.html. Acesso em: Acesso em: 1 set. 2008.
DUGUID, Paul. **Report of the Santa Fe Planning Workshop on Distributed Knowledge Work Environments**: digital libraries. University of Michigan School of Information, Sept. 1997. Disponível em: http://www.si.umich.edu/SantaFe/. Acesso em: 25 maio 2008.
FOX, Edward A. The Digital Library Initiative: update and discussion. **Bulletin of the American Society for Information Science**, v. 26, n. 1, Oct./Nov. 1999. Disponível em: http://www.asis.org/Bulletin/Oct-99/fox.html. Acesso em: 15 maio 2008.

FELDMAN, Ian. Ted Nelson 1990 world tour. **TidBITS**, n. 30, 15 nov. 1990. Disponível em: http://www.xanadu.com.au/media/nelson90.html. Acesso em: 28 jul. 2021.
GREENSTEIN, Daniel. **Next-generation digital libraries**. *In*: JOINT CONFERENCE ON DIGITAL LIBRARIES, 2., 2002, Portland. Disponível em: http://www.vala.org.au/vala2002/2002pdf/01Grnstn.pdf. Acesso em: 28 jul. 2021.
GRIFFIN, Stephen M. NSF/DARPA/NASA Digital Libraries Initiative: a program manager's perspective. **D-Lib Magazine**, Jul./Aug. 1998. Disponível em: http://www.dlib.org/dlib/july98/07griffin.html. Acesso em: 28 jul. 2021.
HARTER, Stephen. Scholarly communication and the digital library: problem and issues. **Journal of Digital Information**, v.1, n.1, 1997. Disponível em: https://journals.tdl.org/jodi/index.php/jodi/article/view/jodi-3. Acesso em: 28 jul. 2021.
HAUBEN, Jay. Vanevar Bush and JCR Licklider: libraries of the future 1945-1965. **The Amateur Computerist**, v. 15, n. 2, Spring 2007. Disponível em: http://www.ais.org/~jrh/acn/ACn15-2.pdf. Acesso em: 28 jul. 2021.
KUNY, Terry; CLEVELAND, Gary. The digital library: myths and challenges. **IFLA Journal**, v. 24, n. 2, 1998. Disponível em: https://journals.sagepub.com/doi/abs/10.1177/034003529802400205. Acesso em: 28 jul. 2021.
LANCASTER, F.W. **Toward paperless information systems**. New York: Academic Press, 1978.
LANCASTER, F.W. The evolving paperless society and its implication for libraries.
International Forum on Information & Documentation , v. 7, n. 4, p. 3-10, 1982.
LAGOZE, Carl *et al.* What is a digital library anymore, anyway? **D-Lib Magazine**, v. 11, n. 11, Nov. 2005. Disponível em: http://www.dlib.org/dlib/november05/lagoze/11lagoze.html. Acesso em: 28 jul. 2021.
LESK, Michael. **Why digital libraries?** UKOLN, 1995. Disponível em: http://www.lesk.com/mlesk/follett/follett.html. Acesso em: 28 jul. 2021.
LI, Bin. **The history of digital library**. [200-?]. Disponível em: www.ils.unc.edu/~lib/digital-library.html. Acesso em: 30 mar. 2008.
LICKLIDER, J.C.R. **Libraries of the future**. Cambridge, Mass.: MIT Press, 1965.
LICKLIDER, J.C.R. Man Computer Symbiosis. **IRE Transactions on Human Factors in Electronics**, v. 1, n.1, p. 4-11, 1960. Disponível em: http://memex.org/licklider.pdf. Acesso em: 28 jul. 2021.
LIU, Jia. Digital library activities in Europe: a brief overview. **Journal of Educational Media & Library Science**, v. 4, n. 42, Jun. 2005. Disponível em: http://joemls.dils.tku.edu.tw/fulltext/42/42-4/455-469.pdf. Acesso em: 28 jul. 2021

LYMAN, Peter. What is a digital library? Technology, intellectual property, and the public interest. **Daedalus**, v. 125, n. 4, Fall 1996. Disponível em: https://sci-hub.se/10.2307/20027384. Acesso em: 31 jul. 2021.

MARCONDES, Carlos Henrique; SAYÃO, Luis Fernando. Integração e Interoperabilidade no Acesso a Recursos Informacionais Eletrônicos em C&T: a proposta da Biblioteca Digital Brasileira. **Ciência da Informação**, v. 30, n. 3, p. 24-33, set./dez. 2001. Disponível em http://www.scielo.br/pdf/ci/v30n3/7283.pdf. Acesso em: 28 jul. 2021

MCLOUGHLIN, Glenn. **The National Information Infrastructure**: the federal role. Washington, D.C. : Congressional Research Service Report, 2000. Disponível em: http://www.ncseonline.org/NLE/CRSreports/Science/. Acesso em: 30 maio 2008.

MCPHERSON, Madelaine. **Managing digital libraries**. *In*: CSIRO INFORMATION, MANAGEMENT & TECHNOLOGY CONFERENCE, 1997, Gold Coast. Disponível em: http://www.usq.edu.au/users/mcpherso/csiro.htm. Acesso em: 31 mar. 2008.

OTLET, Paul. **Traité de documentation**: le livre sur le livre, theorie et pratique. Bruxelles: Editiones Mundaneum, 1934. Disponível em: https://3lib.net/book/3108359/3c0cb9. Acesso em: 31 jul. 2021.

PITAC - PRESIDENT INFORMATION ADVISORY COMMITTEE. **Digital libraries**: universal access to human knowledge. February, 2001. Disponível em: http://www.nitrd.gov/pubs/pitac/pitac-dl-9feb01.pdf. Acesso em: 31 jul. 2021.

SALDANHA, Luis Cláudio Dallier. Bibliotecas imaginárias e o livro eletrônico: possibilidades do texto no ciberespaço. **Revista Philologus**, v. 7, n. 21, 2001. Disponível em: http://www.filologia.org.br/rph/ANO07/21/003.pdf. Acesso em: 31 jul. 2021.

SAUNDERS, Laverna M. **The evolving virtual library:** visions and case studies. Medford, NJ.: Information Today, 1996.

SARACEVIC, Tefko. Information Science. **Journal of the American Society for Information Science**, v. 50, n. 12, 1999. Disponível em: https://tefkos.comminfo.rutgers.edu/SaracevicInformationScienceELIS2009.pdf. Acesso em: 31 jul. 2021.

SAYÃO, Luis Fernando. Preservação digital no contexto das bibliotecas digitais. *In*: MARCONDES , Carlos Henrique; KURAMOTO, Hélio; TOUTAIN, Lidia Brandão; SAYÃO, Luis Fernando (org.). **Bibliotecas digitais**: saberes e práticas. Salvador/Brasília: UFBA/IBICT, 2005, p. 115-149.

SAYÃO, Luis Fernando; MARCONDES, Carlos Henrique. O desafio da interoperabilidade e as novas perspectivas para as bibliotecas digitais. **Transinformação**, v. 20, n. 2, 2008. Disponível em: https://www.scielo.br/j/tinf/LSxTfhK6NfX54t4ypBK87kM/?format=pdf&lang=pt. Acesso em: 31 jul. 2021.

SCHÄUBLE, Peter; SMEATON, Alan. **An international research agenda for digital libraries**: Summary Report of the Series of Joint NSF-EU Working Groups on Future Directions for Digital Libraries Research. 1998. Disponível em: http://www.ercim.org/publication/ws-proceedings/DELOS-B/dl_sum_report.pdf. Acesso em: 31 jul. 2021.
SHIRI, Ali. Digital library research: current developments and trends. **Library Review**, v. 52, n. 5, p. 198-2002, 2003. Disponível em: http://eprints.rclis.org/4905/1/ASLRcolumn.pdf. Acesso em: 31 jul. 2021.
SUN MICROSYSTEMS. **Digital library technology trends**. Sun Microsustems, 2002. Disponível em: https://adlsn.org/greenstone/collect/toolbox/index/assoc/HASHb1cc.dir/Digital%20Library%20Technology%20Trends.pdf. Acesso em: 31 jul. 2021.
TELEMATICS FOR LIBRARIES. **Creating a European Library Space Telematics for Libraries Programmes 1990-1998**. [199-?]. Disponível em: http://cordis.europa.eu/libraries/en/intro.html. Acesso em: 30 maio 2008.
THANOS, Costantino. DELOS: a network of excellence on digital libraries. **DELOS Newletter**, n. 1, 2004. Disponível em: https://www.ercim.eu/publication/Ercim_News/enw57/thanos.html. Acesso em: 31 jul. 2021.
URS, Shalini. **Digital libraries**: an overview. *In*: JOINT WORKSHOP ON DIGITAL LIBRARIES, 2001. Mysore: United States Educational Foundation in India, DRTC/Indan Statistical Institute, 2001.
URS, Shalini. **Digital Libraries**: the road ahead. *In*: INTERNATIONAL CALIBER, 2007, Panjab University, Chandigharh. Disponível em: http://210.212.200.226/shaliniurs_files/caliber.pdf. Acesso em: 30 mar. 2008.
WALLACE, Koehler. Digital libraries and World Wide Web sites and page persistence. **Information Research**, v. 4, n. 4, July 1999. Disponível em: http://informationr.net/ir/4- 4/paper60.html. Acesso em: 31 jul. 2021.
WELLS, H.G. **World brain**: the idea of a permanent world encyclopaedia. *In*: ENCYCLOPÉDIE FRANCAISE. August 1937. Disponível em: https://sherlock.ischool.berkeley.edu/wells/world_brain.html. Acesso em: 31 jul. 2021.
WELLS, H.G. World brain. [S.l.]: Meuthuen & Co. Limited, 1938.
WISEMAN, Norman; RUSBRIDGE, Chris; GRIFFIN, Stephen M. The Joint NSF/JISC International Digital Library Initiative. **D-Lib Magazine**, v. 5, n. 6, June 1999. Disponível em: http://www.dlib.org/dlib/june99/06wiseman.html. Acesso em: 31 jul. 2021.
WRIGHT, Alex. O antepassado esquecido: Paul Otlet. **Revista ExtraLibris**. Tradução de Moreno Barros. 2007. Disponível em:https://mo-re-no.medium.com/o-antepassado-esquecido-paul-otlet-3495bde0b602. Acesso em: 31 jul. 2021.

O desafio da interoperabilidade e as novas perspectivas para as bibliotecas digitais

Luis Fernando Sayão e Carlos Henrique Marcondes

1 Introdução

A informação é produzida de forma dispersa no espaço e no tempo, e as bibliotecas sempre foram, ao longo da História, lugares ou instituições que se opunham a essa dispersão, concentrando a informação num lugar físico para servir a uma determinada comunidade de usuários. Esse é o valor que agregam. Como, porém, as bibliotecas tinham limites físicos, paredes, coleções em estantes, etc., o alcance desses serviços ficava restrito aos membros da comunidade que conseguia ter acesso presencial a ela.

Com o surgimento da Internet, essa situação evolui de forma avassaladora. Nos nossos dias, mais e-mails registros da cultura humana são produzidos já diretamente em formato digital como música, imagens, vídeos, material textual – inclusive aquele a ser impresso em papel –, e ainda as novas formulações e concepções de registros proporcionados pela tecnologia. A Internet é o retrato vivo da "explosão informacional", do "caos informacional", mas o acúmulo imenso e desordenado de informação é, ao mesmo tempo, a sua fortaleza e a sua fragilidade, sendo necessário, portanto, algum grau de ordenamento e intermediação.

O ser humano individualmente não tem capacidade cognitiva em absorver esse volume crescente de informação colocado à sua disposição. Portanto são necessários sistemas de intermediação que se interponham entre o usuário e as fontes de informação que agreguem valor ao avaliar/selecionar/filtrar a informação. Com a realização do conceito de biblioteca digital, além da capacidade de coletar e concentrar informações dispersas ter aumentado enormemente, acrescentou-se ainda – apoiado fortemente em tecnologia da informação – algo de crucial importância: o potencial de atender a uma comunidade, que não se restringe mais a quem tem acesso presencial à biblioteca. Além disso, abriu-se a possibilidade inédita de um número ilimitado de usuários poder acessar simultaneamente a mesma cópia de um documento digital.

Essas características das bibliotecas digitais ampliam de forma extraordinária as suas possibilidades, tornando-as um instrumento de grande efetividade para distribuição, cooperação e acesso ao conhecimento, atendendo e podendo servir de pólo agregador para comunidades segmentadas e distribuídas geograficamente. Isso, se devidamente explorado, pode representar um impacto importante para um país como o Brasil, que se caracteriza pela diversidade em todas as suas dimensões. Pode-se resumir que a principal vantagem das bibliotecas digitais sobre as bibliotecas físicas é a capacidade que elas têm de multiplicar o alcance – geográfico e temporal - em termos das comunidades que elas são capazes de atingir e servir.

Como ilustração, imagine o seguinte caso de uso: como integrar numa biblioteca digital conteúdos em cultura brasileira e língua portuguesa sobre "o Brasil colônia e a influência da cultura negra", em formato digital disponibilizados na Internet, para fins de apoiar atividades educacionais de professores brasileiros? Conteúdos de interesse para essa proposta vão desde coleções custodiadas por museus, arquivos e bibliotecas, em diferentes graus de tratamento técnico informatizado, usando nenhuma ou diferentes tecnologias, padrões, protocolos e estágios de interoperabilidade, até páginas Web simples de grupos Afro, ONGs, grupos de teatro e programas governamentais. Esses diferentes conteúdos só poderão ser integrados e reusados, no sentido de terem aproveitadas as sinergias uns dos outros, se estiverem ancorados por sistemas que permitam um alto grau de interoperabilidade.

Nos últimos anos, a interoperabilidade tem sido um dos itens mais críticos para quem pensa no desenvolvimento e operação de sistemas de repositórios e de bibliotecas digitais distribuídos funcionando em rede. O conceito de interoperabilidade, entretanto, está longe de ser uma novidade no domínio das bibliotecas: há muito tempo, desde meados do século xx, para fazer frente ao fenômeno social da "explosão informacional", as bibliotecas sempre estabeleceram serviços cooperativos, trocaram informações, criaram um ordenamento universal dessas informações. Toda uma estrutura global foi montada em torno da ideia do compartilhamento e da cooperação entre bibliotecas.

Com a concretização e a consolidação do conceito de bibliotecas digitais – que se localiza na interseção entre biblioteconomia, ciência da computação e tecnologias de rede – a interoperabilidade torna-se um foco de grande interesse para muitos atores. Uma razão para que ela tenha recebido tanta atenção é que o problema permeia quase todos os aspectos circunscritos pela ideia de biblioteca digital, enquanto sistema implementado segundo uma arquitetura distribuída. Outro dado relevante é o crescente interesse da indústria de conteúdos, principalmente por parte dos setores ligados à editoração científica, de incorporar nos seus modelos de negócio as novas formas de disseminação dos repositórios digitais como meio de

distribuição de seus produtos no ambiente de rede. Isso se desenrola no contexto de uma economia da informação em transição, que está saindo de um modelo baseado em assinaturas para modelos baseados no acesso e uso. O acesso via *pay-per-view* a um artigo científico serve como exemplo de uma modalidade que se está popularizando. Nesse escopo está incluído ainda o interesse pelas tecnologias automatizadas de controle de direitos autorais como parte integrante das estratégias de interoperabilidade.

Há ainda o interesse de instâncias governamentais nas potencialidades dos novos serviços que se tornam possíveis com a integração de acervos digitais heterogêneos e distribuídos. Uma das principais áreas onde a integração e o compartilhamento são altamente demandados é a da educação, particularmente no ensino a distância. A adoção generalizada das tecnologias Internet como canal para a educação e treinamento tem resultado numa abundância de recursos educacionais e de treinamento em formato digital pronto para utilização via Web. Não obstante a sua aparente onipresença, a localização e reuso desses objetos educacionais são prejudicados pela falta de esforços coordenados e políticas voltadas para o armazenamento, o tratamento técnico e a gestão de direitos visando a uma integração enriquecedora desses conteúdos (HATALA *et al.*, 2004).

Em paralelo às bibliotecas, os arquivos e ainda os museus têm também vislumbrado a Internet como um mecanismo por meio do qual é possível ampliar de uma forma nunca antes possível as fronteiras físicas do mundo convencional, criando novas possibilidades de acesso e disseminação de suas coleções. Em relação a esse tema, pode ser interessante consultar Warren, Thurlow, Alsmeyer (2006), Projeto DigiCult[1] e página "*Archives & Museums Informatics*"[2].

Nesse contexto, tendo a percepção mais imediata das vantagens dos repositórios digitais, dispondo de alguns recursos financeiros, humanos e metodológicos e tendo ferramentas de *software*, via de regra, livremente disponíveis, as organizações da área de conhecimento – principalmente as bibliotecas, arquivos e museus – têm crescentemente migrado seus estoques de informação para repositórios digitais. Caracteristicamente esses repositórios são estanques e fragmentados e as buscas a eles devem ser realizadas individualmente, por meio de interfaces e formulações distintas. Quem desejar integrar dados de diferentes locais terá que fazê-lo manualmente. Esse enfoque tem produzido muitos sistemas de repositórios digitais heterogêneos, mas torna a interoperabilidade, o reuso, o intercâmbio e o desenvolvimento cooperativo extremamente difíceis de se alcançar. Isso não é exa-

1 Disponível em: http://www.digicult.info/pages/info.php. Acesso em: 23 ago. 2021.

2 Disponível em: http://www.archimuse.com. Acesso em: 23 ago. 2021.

tamente o que se espera como resultado dessa transição para um novo conceito de integração de bibliotecas.

O acompanhamento do estado-da-arte na área de bibliotecas digitais têm mostrado como evoluiu o enfoque da integração de dados nesse domínio: no primeiro momento caminhou-se no sentido de se criar bibliotecas digitais isoladas. Nesse patamar, as bibliotecas são mantidas por uma única instituição e as coleções de dados são autocontidas, enquanto os conteúdos são formalmente localizados e gerenciados de forma centralizada. O passo seguinte está relacionado à federação em rede de várias bibliotecas independentes, possivelmente organizadas em torno de um tema ou área comum, formando uma rede de bibliotecas acessível por meio de uma única interface (PIRRI; PETTENATI; GIULI, 2002).

Existem diferentes modelos, tecnologias e metodologias para se alcançar o nível necessário de interoperabilidade entre bibliotecas digitais e seus componentes de serviço, tendo como perspectiva as razões e os fundamentos do sistema de informação como um todo. É esse tema, precisamente, que vai ser discutido aqui.

2 Antecedentes – o problema

Há um consenso absoluto de que a interoperabilidade é um problema de vasto domínio. Contudo, ela tem sido investigada tipicamente dentro de escopos específicos, tais como os circunscritos por uma comunidade particular, por exemplo: bibliotecas, empresas, comunidades científicas; dentro de uma classificação particular de informação, por exemplo, registros eletrônicos, relatórios técnicos, *software*; ou ainda dentro de área particular de tecnologia da informação (TI), como bases de dados relacionais, imagens digitais, visualização de dados. Escapando desse fracionamento, as pesquisas mais recentes sobre interoperabilidade no âmbito da arquitetura de bibliotecas digitais estão concentradas no desafio de criar uma infraestrutura para acesso e integração de informação transversalmente a esses domínios específicos. Um objetivo comum desses esforços é permitir que diferentes comunidades, com diferentes tipos de informação e usando diferentes tecnologias, consigam um nível geral de compartilhamento de informação e, por meio de processos de agregação apoiados por tecnologia da informação, criem novos e mais poderosos tipos de serviços de informação (PAYETTE *et al.*, 1999).

Idealmente, uma biblioteca digital deve ser capaz de armazenar uma variedade de tipos tradicionais de conteúdo – livros, periódicos, relatórios técnicos, *softwares* –, bem como entidades multimídia complexas que misturam texto, imagens, vídeo e dados. Para que o acesso a essas informações seja efetivamente viável, o sistema no qual elas estão armazenadas deve ser capaz de gerar processos que sejam interoperáveis com os sistemas que estão à sua volta. Uma organização verdadeiramente

interoperável é capaz de maximizar o valor e o potencial de reuso da informação que está sob o seu controle. É também capaz de intercambiar efetivamente estas informações com outras organizações igualmente interoperáveis, permitindo que novos conhecimentos possam ser gerados a partir da identificação de relacionamentos entre conjuntos de dados previamente não relacionados. Na perspectiva do usuário, as interfaces devem apresentar para o usuário uma visão unificada em termos semânticos de diferentes recursos informacionais heterogêneos, ou seja: como nomeá-los, como referenciá-los, como utilizá-los em buscas, como acessá-los, como apresentá-los para o usuário.

O que se espera como resultado das pesquisas em bibliotecas digitais é que se possa estabelecer uma infraestrutura conceitual, tecnológica, metodológica, de serviços e gerencial coerente que permita: que sejam sempre possível buscas simultâneas em coleções múltiplas, heterogêneas e operadas por sistemas tecnologicamente diferentes; que essas coleções possam estar custodiadas por organizações distintas; que possam estar geograficamente espalhadas por todo o mundo; que as buscas possam ser realizadas pelo usuário final de maneira fácil e transparente e utilizando ferramentas comuns de acesso à Internet, como são os navegadores Web. Além do mais, essa infraestrutura deve ser tal que possa oferecer ainda ferramentas para gestão de direitos e propriedade intelectual, e finalmente, considerar a identidade do usuário e seus direitos individuais e institucionais de acesso.

3 O que é interoperabilidade

O *Online Dictionary for Library and Information Science* (ODLIS), define o termo interoperabilidade como:

> A capacidade de um sistema de *hardware* ou de *software* de se comunicar e trabalhar efetivamente no intercâmbio de dados com um outro sistema, geralmente de tipo diferente, projetado e produzido por um fornecedor diferente (ODLIS, 2004).

Para a área de tecnologia da informação, há um consenso geral de que interoperabilidade é algo como a capacidade de computadores e programas de fabricantes diferentes trocarem informações. No contexto das bibliotecas, porém, o conceito de interoperabilidade não está circunscrito somente a uma questão de comunicação entre componentes de um sistema de computadores. Mais especificamente no âmbito das bibliotecas digitais, o conceito de interoperabilidade é complexo e estratificado, refletindo a diversidade de visões, o número de variáveis envolvidas e a interdisciplinaridade que está subjacente a ele. Por exemplo, no contexto do

importante trabalho desenvolvido em colaboração pelo *Digital Library Research Group* da Cornell University e a *Corporation for National Research Initiative* (CNRI), interoperabilidade é definida como "a capacidade de componentes ou serviços de bibliotecas digitais serem funcionalmente e logicamente intercambiáveis em virtude de deles terem sido implementados de acordo com um conjunto de interfaces bem definidas e publicamente conhecidas" (PAYETTE *et al.,* 1999, p.2). No modelo implementado, tendo como referência essa definição, serviços e componentes distintos podem comunicar-se mutuamente por meio de interfaces abertas, e os usuários podem interagir com eles de maneira equivalente.

Na visão da Ukoln (2005), expressa também por Miller (2000), a interoperabilidade pode ser considerada como o processo contínuo de assegurar que sistemas, procedimentos e cultura de uma organização sejam gerenciados de tal forma que possibilitem a maximização das oportunidades para intercâmbio e reuso de informação.

A partir dessas definições, fica claro que a interoperabilidade está longe de depender somente de requisitos técnicos – como, por exemplo, o uso de programas e computadores compatíveis –, embora possa ser importante em algumas situações. Assegurar a plena interoperabilidade exige frequentemente uma mudança profunda na forma pela qual uma biblioteca digital trabalha, relaciona-se com as organizações parceiras, usuários e fornecedores e, especialmente, sua atitude diante dos problemas relacionados à informação. O desafio de projetar serviços coerentes para uma diversidade de usuários a partir de componentes que são tecnicamente diferentes e gerenciados por diferentes organizações exige um sofisticado grau de cooperação que pode ser diferenciado em pelo menos três instâncias (ARMS, 2000; ARMS *et al*., 2002):

a) acordos técnicos - cobrem formatos, protocolos, sistemas de segurança de forma que mensagens possam ser trocadas;
b) acordos sobre conteúdos – cobrem dados e metadados e incluem acordos semânticos sobre interpretação das mensagens;
c) acordos organizacionais – cobrem as regras básicas para acesso, para mudanças nas coleções e serviços, pagamento, autenticação, etc.

Soma-se ainda uma instância de acordos políticos, onde são estabelecidos os fóruns necessários e definidas as diretrizes e as políticas concernentes, incluindo as questões de financiamento.

4 As muitas faces da interoperabilidade

A face mais visível da interoperabilidade é certamente a interoperabilidade técnica, por ser, talvez, o aspecto mais perceptivelmente responsável por manter

os sistemas de informação interoperáveis. A interoperabilidade tem muitas outras faces, entretanto, cada uma delas revelando um aspecto da complexidade de efetivamente se ter bibliotecas digitais funcionalmente integradas. Portanto outros aspectos não menos relevantes devem ser também considerados (MILLER, 2000; UKOLN, 2005):

a) **Interoperabilidade técnica** – as considerações sobre os aspectos técnicos incluem assegurar envolvimento de um conjunto de organizações no contínuo desenvolvimento de padrões de comunicação, transporte, armazenamento e representação de informações, tais como são o Z39.50[3], *Search Retrieval Web Service* (SRW)[4], ISO-ILL e o XML[5]. Inclui também os esforços cooperativos para assegurar que padrões individuais evoluam em benefício da comunidade envolvida e para facilitar, onde for possível, convergência desses padrões, de forma que seja possível que os sistemas possam ter como base mais de um conjunto de padrões.
b) **Interoperabilidade semântica** – está relacionada com o significado ou semântica das informações originadas de diferentes recursos e é solucionada pela adoção de ferramentas comuns ou/e mapeáveis de representação da informação, como esquemas de metadados, classificações, tesauros e mais recentemente, ontologias; um exemplo de questão endereçada por essa faceta da interoperabilidade pode ser o seguinte: o que significa "autor" para um recurso informacional? Será a mesma coisa que "criador" para um outro recurso?
c) **Interoperabilidade política/humana** – independente das questões relacionadas à maneira pela qual a informação é descrita e disseminada, a decisão de tornar os recursos informacionais mais amplamente disponíveis e interoperáveis tem implicações para a organização, para as equipes envolvidas e para os usuários em termos comportamentais, de recursos e de treinamento. A ênfase dada por parte de alguns setores governamentais aos problemas de democratização do acesso, da exclusão digital e da federação de fontes de informação voltadas para a educação a distância, tem impacto nas políticas públicas para a área, e estão enquadrados neste item.
d) **Interoperabilidade intercomunitária** – enfoca a necessidade, cada vez mais urgente, impulsionada pela crescente interdisciplinaridade, princi-

3 Disponível em: http://www.loc.gov/z3950/agency/. Acesso em: 23 ago. 2021.
4 Disponível em: http://www.loc.gov/standards/sru/srw/index.html. Acesso em: 1 set. 2008.
5 Disponível em: http://www.w3.org/XML/. Acesso em: 23 ago. 2021.

palmente nas áreas de pesquisa, de acesso a informações provenientes de um espectro amplo de fontes distribuídas por organizações, áreas de conhecimento e comunidades de natureza distintas. Geralmente exige o estabelecimento de fóruns para discussão e consenso em torno de práticas e procedimentos comuns.

e) **Interoperabilidade legal** – considera as exigências e as implicações legais de tornar livremente disponíveis itens de informação;
f) **Interoperabilidade internacional** – quando se atua em escala internacional é necessário contornar a diversidade de padrões e normas, os problemas de comunicação, as barreiras linguísticas, as diferenças no estilo de comunicação e na falta de uma fundamentação comum.

5 Níveis de interoperabilidade

O nível de interoperabilidade é referido aqui como o grau de compromisso ou acoplamento entre sistemas (instituições, bibliotecas digitais) para torná-los interoperáveis e seria uma medida do esforço para tornar-se interoperável. Arms *et al*. (2002) relacionam as funcionalidades ou facilidades oferecidas aos usuários resultantes de um alto nível de interoperabilidade entre diversas bibliotecas digitais versus o custo de adesão a esse nível de funcionalidade por parte de novos parceiros: quanto maior o nível de interoperabilidade, maior o custo ou esforço para que novos parceiros adiram à iniciativa. Esses autores, no contexto do desenvolvimento da NSDL (*National SMETE Digital Library*), identificam três níveis de interoperabilidade aplicáveis ao domínio das bibliotecas digitais: federação, *harvesting* (colheita automática de metadados) e *gathering* (agregação automática de informação).

O nível mais alto, a federação, corresponde à mais robusta forma de interoperabilidade; em contrapartida, é a que exige maior esforço dos participantes. Para efetivar-se, ela exige que um grupo de organizações concorde que seus serviços estejam em conformidade com um conjunto de especificações, geralmente selecionadas a partir de padrões formalizados. Os termos "heterogêneo" e "federado" são frequentemente usados para descrever sistemas cooperativos nos quais componentes individuais são projetados ou operados de forma autônoma. Esse tipo de cooperação está em contraste com o termo geral "sistema distribuído", que também inclui coleções de componentes desenvolvidos em diferentes sites, mas que são cuidadosamente projetados para trabalhar em conjunto (PAEPCKE *et al*., 1998).

O principal desafio que se coloca na formação de federações é o esforço despendido por cada organização em implementar e manter atualizados todos os níveis dos acordos. As bibliotecas que compartilham registros de catálogos online

usando o protocolo z39.50, trabalham segundo o nível de federação. O ansi/niso z39.50 (iso 23950) é um protocolo de comunicação entre computadores que pode ser implementado sobre qualquer plataforma. Ele tem como propósito a pesquisa e a recuperação de informações. A implementação do protocolo permite que, por meio de uma única interface, seja possível o acesso uniforme a uma diversidade de fontes de informações heterogêneas de modo síncrono e transparente para o usuário-final (niso, 2002). Outro padrão nesta área é o *Search Retrieval Web Service* (swr)[6], um protocolo que se propõe a ser o sucessor do z39.50.

As dificuldades de se criar grandes federações, porém, foi a principal motivação para busca de soluções menos onerosas para o estabelecimento de interoperabilidade entre bibliotecas digitais. Ideia subjacente é que os participantes concordem em despender um pequeno esforço que possibilite o compartilhamento de alguns serviços básicos, sem que seja necessário o enquadramento a um conjunto completo de acordos. Nessa situação enquadra-se o conceito de colheita automática de metadados (*metadata harvesting*), estabelecido pelo protocolo oai-pmh (*Open Archive Intiative Protocol of Metadata Harvesting*). Enquanto os serviços baseados em havesting são assíncronos e menos sofisticados do que os providos pelas federações, a sobrecarga sobre os participantes é consideravelmente menor. Como resultado, muito mais organizações, especialmente as surgidas no seio da academia, estão optando por esse tipo de interação, o que é provado pela rápida aceitação do oai-pmh como um protocolo essencial nas transformações que vêm ocorrendo nos padrões de comunicação científica (marcondes; sayão, 2001).

Ainda que um determinado grupo de organizações não estabeleça nenhum grau formal de cooperação, um nível básico de interoperabilidade é ainda possível por meio de agregação automática de informações disponíveis publicamente, utilizando-se metabuscadores, robôs, máquinas de busca e ainda por meio dos protocolos que suportam *web services* e outros padrões da indústria de ti. A agregação requer essencialmente pouco ou nenhum esforço por parte dos participantes, entretanto oferece um grau baixo de interoperabilidade (arms, 2002).

Um nível ainda mais formalizado do que a federação pode ser ainda postulado: trata-se da padronização. Nesse nível, cada aspecto da interoperabilidade é formalmente definido e cada organização tem o rígido compromisso de seguir exatamente o conjunto de padrões e procedimentos convencionados. Na prática, isso pode determinar o uso da mesma plataforma computacional – *hardware*, aplicativos e sistema operacional – e das mesmas condicionantes administrativas, reduzindo drasticamente a autonomia dos componentes individuais. É esse nível que geralmente se estabe-

6 Disponível em: http://www.loc.gov/standards/sru/srw/index.html. Acesso em: 1 set. 2008.

lece no âmbito das redes cooperativas de bibliotecas que utilizam formatos de intercâmbio padronizados como Lilacs e MARC. É uma solução pouco provável além das fronteiras corporativas e de redes e sistemas altamente formalizados (ARMS, 2002).

Na sua fase inicial, os projetos de bibliotecas digitais adotavam um único mecanismo de interoperabilidade, como por exemplo, o protocolo OAI-PMH, ou o protocolo Z39.50. É importante observar que as estratégias adotadas por sistemas importantes e de espectro amplo como o NSDL e como era o projeto original da Biblioteca Digital Brasileira[7] (MARCONDES; SAYÃO, 2001), apontam para soluções em diversos níveis simultâneos, em que o grau de formalização das parcerias e o nível de interoperabilidade vão depender da importância das coleções, dos tipos de serviços e do grau de esforço necessário para operacionalizá-los. É possível, portanto, no contexto de um único sistema, estabelecerem-se federações distintas em torno de coleções especiais e de acordos técnicos e de conteúdo além de acordos organizacionais específicos (por exemplo, coleção de material didático), onde a interoperabilidade menos formal – *harvesting* e *gathering* – também é considerada importante forma de cooperação. Ressalta-se, por último, que o termo federação, apesar de indicar um nível específico de interoperabilidade, tem sido muito frequentemente usado para indicar genericamente a integração e a interoperabilidade entre repositórios digitais em diferentes níveis e operando simultaneamente, principalmente por autores mais próximos das áreas de TI.

6 Tecnologias de informação e bibliotecas digitais

Ao contrário das metodologias usadas, num passado não muito distante, para automação de bibliotecas (SAYÃO; MARCONDES, 1989), que tipicamente utilizavam tecnologias, ferramentas e padrões específicos da área bibliográfica, as bibliotecas digitais, numa visão otimista, podem incorporar quase que imediatamente e com poucos limites os progressos das Tecnologias de Informação e Comunicação (TICs), que fazem avançar quase que cotidianamente outras áreas importantes, como comércio eletrônico e telefonia móvel. Numa avaliação rápida sobre as áreas de pesquisas em bibliotecas digitais, principalmente no que concerne a tecnologias e arquiteturas, pode-se afirmar, sem muito risco, que elas são compostas da união de subáreas provenientes de vários domínios. Grande parte dos problemas específicos relacionados à construção de bibliotecas digitais têm sido enfocados como variações de problemas pertinentes a outros campos, como veremos nesta seção.

O acesso a recursos da Internet qualificados como distribuídos e heterogêneos é seguramente, como já enfatizado, um dos problemas críticos que devem ser su-

7 Disponível em: http://bdtd.ibict.br/. Acesso em: 23 ago. 2021.

perados na pesquisa e desenvolvimento da próxima geração de bibliotecas digitais. Os repositórios digitais disponíveis em rede, via de regra, variam em termos de interfaces, de representação de dados e, em geral, esses repositórios têm esquemas de metadados diferentes em termos de número de campos, de sintaxe e de semântica, e ainda são acessados por meio de protocolos incompatíveis e permeados por esquemas de proteção de propriedade intelectual diversificados. Tudo isso coloca o problema de desenvolver arquiteturas para federação de recursos heterogêneos, tendo como foco a tentativa de estabelecer pontes entre diferentes sistemas de informação e de representação, um dos principais objetos de pesquisa na área. O objetivo de uma arquitetura de federação de serviços é oferecer uma interface uniforme para os recursos individuais, bem como uma visão integrada sobre os dados. Portanto a arquitetura deve ser concebida de forma a aceitar consultas sobre uma visão global das informações, baseada num modelo de dados uniforme.

O projeto de arquitetura pode ser enfocado de diversas formas, porém, de maneira genérica, a arquitetura de federação de serviços pode ser estruturada em três camadas distintas: 1) camada de repositórios digitais, onde as informações estão armazenadas com autonomia de representação e de interfaces de acesso; 2) camada de adaptação, que provê acesso uniforme às informações ocultando as diferenças de modelos de dados e de interfaces de consulta. Nessa camada, adaptadores especiais ou mediadores – por exemplo, *harvesters* - têm que ser implementados para transformar os modelos específicos das fontes de dados em um modelo global do sistema federado. O mapeamento de esquemas particulares de metadados usados por cada repositório em um padrão comum, por exemplo, Dublin Core, serve como ilustração dessa camada; e 3) camada de federação, que responde pela integração global dos dados. Essa camada oferece os serviços para definição de uma visão integrada dos dados e consultas. É nessa instância que se pode dispor de bases de dados para descrever, por meio de metadados, os diferentes recursos disponíveis.

Muitos avanços importantes têm sido alcançados nos últimos anos, especialmente no desenvolvimento de mediadores que acessam as informações de fontes múltiplas. Um mediador tipicamente recebe uma requisição, por exemplo, uma consulta, e submete uma versão traduzida dessa requisição às várias fontes de informação, recupera e integra as respostas e apresenta ao usuário (melnik; garcia-molina; paepcke, 2000).

Na implementação de arquiteturas de federação, várias metodologias comuns à área de ti estão sendo utilizadas isoladamente ou em conjunto, ou são partes integrantes de projetos de pesquisa importantes que estão em andamento. A seguir, relacionam-se algumas dessas metodologias:

Web Semântica - é uma extensão da Web atual -, que é formada por documentos compreensíveis unicamente por pessoas - para uma Web em que documentos seriam auto descritíveis, de forma que seu conteúdo possa ser "compreendido" por programas especiais, os agentes inteligentes de *software* (descrito mais adiante), que assim poderiam "raciocinar" e fazer inferências sobre o conteúdo de documentos, ajudando as pessoas em diferentes tarefas de recuperação e compartilhamento de informações que exijam raciocínio, decisões, inferência de conclusões a partir de informações não explicitamente disponíveis ou de informações contextuais. Nas palavras de Berners-Lee e seus colaboradores, a Web Semântica "é uma extensão da Web atual, na qual é dada à informação um significado bem definido, permitindo que computadores e pessoas trabalhem em cooperação" (BERNERS-LEE; HENDLER; LASSILA, 2001, tradução nossa). Há uma grande expectativa em torno das tecnologias da Web Semântica e a suas implicações para o futuro da cultura humana, que são destacadas pelos seus próprios criadores:

> A Web Semântica não é meramente a ferramenta para administrar tarefas individuais que temos discutido até agora. Além do mais, se adequadamente projetada, a Web Semântica pode ajudar a evolução do conhecimento humano como um todo (BERNERS-LEE; HENDLER; LASSILA, 2001, tradução nossa).

Espera-se que tecnologias dotadas de habilidades semânticas possam ajudar os usuários de bibliotecas digitais de várias formas. As pesquisas nessa área, de forma geral, investigam como as tecnologias da Web Semântica podem potencializar as funcionalidades das bibliotecas digitais, especialmente na descoberta de recursos com mais eficiência e efetividade, e no compartilhamento de conhecimento no escopo da comunidade do usuário circunscrita pela biblioteca digital. Cenários típicos de uso de tecnologias semânticas em bibliotecas digitais incluem ainda, entre muitos outros usos, interfaces e interação homem-computador, exibição de informação, permissão para visualização e navegação em grandes coleções, construção de perfis de usuário, personalização e interação de usuários.

É consenso que uma das maiores fragilidades da Internet atual é que ela, apesar de ser adequada para a compreensão humana, não favorece a compreensão por máquina, de forma que a interpretação da informação contida em um dado documento necessita sempre da interferência humana. Para superar esses problemas, as ontologias recentemente tornaram-se um foco de intenso interesse da Ciência da Computação. As ontologias estabelecem uma compreensão compartilhada de um domínio de interesse para apoiar a comunicação entre seres humanos e

agentes computacionais. As ontologias são representadas caracteristicamente por uma linguagem de representação processada por computador, sendo considerada uma tecnologia-chave para o desenvolvimento da Web Semântica (SURE; STUDER, 2005). Dessa forma, por meio do estabelecimento de esquemas comuns na forma de ontologias, as tecnologias semânticas permitem a descrição de objetos e repositórios digitais com o objetivo principal de capacitar a interoperabilidade, como será visto a seguir.

- **Ontologias** – a nova geração de sistemas de informação deverá ser capaz de resolver o problema da interoperabilidade semântica. Esses sistemas deverão ter habilidade para processar o modelo que o usuário faz do mundo e seus significados e processar também os modelos que há por trás das fontes de informação; para tal, o projeto de ontologias tem um papel de fundamental importância (FONSECA; ENGENHOFER; BORGES, 2000). Ontologia é uma disciplina antiga – que vem desde Aristóteles – que estuda o ser e as suas propriedades. No domínio da Ciência da Informação e da Inteligência Artificial, uma das definições mais citadas diz que "ontologia é uma especificação formal e explícita de uma conceitualização compartilhada" (GRUBER, 1996, tradução nossa). A conceitualização é uma visão abstrata e simplificada de um domínio específico da realidade. Dessa forma, as ontologias são projetadas para possibilitar que o conhecimento seja compartilhado e reusado; elas explicam como um indivíduo, grupo, linguagem ou ciência entende um determinado domínio. Ontologia é um componente importante no processamento de conhecimento por computador e, consequentemente, para a solução de interoperabilidade semântica entre repositórios digitais federados, principalmente no mapeamento entre esquemas de representação distintos. A DAML+OIL e a *Web Ontology Language* (OWL), recentemente publicada pela W3C, são linguagens usadas para aplicações que necessitam compreender o conteúdo da informação, que são correntemente usadas na implementação de arquiteturas federadas (MARTINEZ, 2006; DOERR, 2006). Experiências sobre o uso de ontologias para os problemas das bibliotecas digitais podem ser vistos em Weinstein (1998);
- **Agentes inteligentes** – o conceito de agente constitui uma poderosa ferramenta de abstração no desenvolvimento de *softwares* que facilitem a construção de sistemas robustos, inteligentes e distribuídos. Um agente inteligente[8] é um programa de computador especial capaz de ações flexíveis e autônomas

8 Disponível em: http://en.wikipedia.org/wiki/Software_agents. Acesso em: 23 ago. 2021.

num determinado ambiente, ou seja, ações reativas, proativas e sociais. A solução de um problema complexo, como a federação de bibliotecas digitais, pode ser contemplada por um conjunto de agentes que interagem e cooperam entre si para alcançar os objetivos do sistema. Agentes inteligentes informacionais podem ser definidos como sistemas de *software* que acessam fontes de informação múltiplas, heterogêneas e geograficamente distribuídas no sentido de assistir ao usuário no processo de buscar informações relevantes (MARTINEZ, 2006). As exigências mínimas que se colocam para os agentes informacionais são: saber onde encontrar os repositórios digitais; ter a capacidade de consultar esses repositórios de forma apropriada, interpretando, analisando e traduzindo a solicitação do usuário para cada fonte de informação; possuir métodos para processar os resultados da pesquisa, integrando os resultados em um formato comum e apresentando-os aos usuários (BIRMINGHAM, 1995; TENNANT, 1998);

- ***Digital Library Definition Language* (DLDL)** – linguagem de especificação de bibliotecas digitais baseada em XML capaz de descrever API´s para um grande número de repositórios digitais. A especificação é dividida em três seções: 1) as informações que as bibliotecas digitais contêm; 2) os métodos de acesso das bibliotecas digitais; 3) as informações a serem recuperadas das bibliotecas digitais (ZUBAIR, 2000);
- **Encapsuladores, Tradutores, Adptadores (*Wrapers*)** – são dispositivos de *software* que ocultam parte da heterogeneidade técnica e de modelo de dados de componentes nativos de sistemas de informação, tornando o acesso transparente para o sistema de mediação. De forma geral, os *wrappers* convertem dados provenientes de uma fonte de informação para um modelo de dados comum, e convertem consultas de aplicações em consultas específicas da fonte de informação correspondente. Por exemplo: um *wrapper* Z39.50 desempenha o papel de *gateway* entre o sistema mediador e um servidor Z39.50, com o qual se comunica via o protocolo nativo; nesse caso, o próprio Z39.50 (MELNIK; GARCIA-MOLINA; PAEPCKE, 2000; BARTELT *et al*., 2001; BARU *et al*., 1999);
- **CORBA, sigla para *Common Object Request Broker*** – especificação padrão criada pelo OMG (*Object Management Group*), que propõe uma arquitetura de *software* para suportar a distribuição e garantir a interoperabilidade entre diferentes plataformas de *hardware* e sistemas operacionais, visando estabelecer e simplificar a troca de dados entre sistemas distribuídos e heterogêneos. A CORBA atua de forma que os objetos (componentes de *software*) possam comunicar-se de forma transparente em relação ao usuário, mesmo

que seja necessário interoperar com outro *software* em outro diferente ambiente operaciona. A especificação define um módulo intermediário entre o cliente e o objeto, o ORB (*Object Request Broker*). Trata-se de um *middleware* responsável em aceitar a requisição do cliente, localizar o objeto e passar os parâmetros necessários a esse objeto, fazer as chamadas dos métodos e entregar a resposta ao cliente. Dessa maneira, o usuário não precisa preocupar-se em saber onde esse objeto está localizado, em que sistema operacional ele roda ou qual programa foi usado para desenvolvê-lo. Esse enfoque de objetos distribuídos tem sido adotado como solução em algumas experiências importantes, como, por exemplo, no trabalho cooperativo entre CNRI e a *Cornell University* (PAYETTE *et al.,* 1999; PAEPCKE *et al*., 1998);

- **Web services**[9] – é uma tecnologia utilizada na integração de sistemas e na comunicação entre aplicações diferentes. Essa comunicação efetiva-se de forma padronizada, possibilitando a independência de plataforma e de linguagem de programação. As bases para a construção de um *web service* são a linguagem XML e o protocolo SOAP (*Simple Object Acess Protocol*)[10], definido pela W3C (*World Wide Web Consortium*). O transporte de dados é realizado via o protocolo HTTP (*HyperText Transfer Protocol*). Dessa forma, os dados são transferidos no formato XML e encapsulados pelo protocolo SOAP. Para ser utilizado por outras aplicações e para que tenha também seu funcionamento entendido pelos interessados, um *web service* deve ser publicado e disponibilizar uma definição de como ele é, como deve ser acessado e que valores retornará. Essas definições são descritas em um arquivo de acesso público em XML, de acordo com a padronização *Web Service Description Language* (WSDL) (PAMPLONA, 2004). *Web services* são blocos para construir aplicações, a partir dos quais se pode vislumbrar uma infraestrutura para bibliotecas digitais distribuídas que disponibilize uma gama de funcionalidades, como, por exemplo, buscas distribuídas, perfis de usuários, serviços de autenticação/autorização, diretório de coleções. Isso é possível porque a arquitetura de bibliotecas digitais é muito similar à arquitetura exigida para o estabelecimento de *web services*:

> De fato, a infraestrutura de *web services* pode ser pensada como uma biblioteca digital distribuída voltada para serviços ao invés de infor-

9 Disponível em: http://www.w3.org/2002/ws/. Acesso em: 23 ago. 2021.

10 Disponível em: http://www.w3.org/TR/2003/REC-soap12-part0-20030624/. Acesso em: 23 ago. 2021.

> mação. Isto significa que muitas das questões em que a comunidade de bibliotecas digitais está envolvida, tais como metadados para descoberta de recursos, autenticação e autorização e modelos de negócio para o acesso a propriedade intelectual, são também aplicáveis a *web service* e devem ser resolvidas dentro de um contexto *web services* (GARDNER, 2001, p.1).

- É importante notar que já existem pacotes de *software* para repositórios digitais que implementam interfaces como *web services*, como por exemplo o Fedora[11]; e o protocolo SRW (*Search Retrieval Web Service*)[12], que tem como proposta ser aderente ao padrão de interoperabilidade estabelecido pelo *Web Services* Interoperability;
- ***Arquitetura Peer-to-Peer*** **(P2P)** – é uma arquitetura de sistemas distribuídos, na qual cada componente, chamado nó, é autônomo e tanto presta como fornece serviços aos outros nós. Essa arquitetura tornou-se muito comum com a popularização dos sistemas de troca de arquivos MP3, como era originalmente o Napster e são, atualmente, o Kazaa e o Emule. Aplicada a uma federação de bibliotecas digitais essa arquitetura, dada uma solicitação de serviço por determinado nó, uma busca, por exemplo, o nó ou nós capazes de melhor atender a busca são identificados segundo diferentes critérios e se encarrega de prestar o serviço. Bibliotecas digitais como BRICKS - *Building Resources for Integrated Cultural Knowledge Services*[13] funcionam segundo essa arquitetura.
- **Arquitetura multicamadas (*muti-tier architecture*)** - é uma nova dimensão para a arquitetura cliente-servidor, na qual uma aplicação é executada por mais de um diferente agente de *software*. Por exemplo, uma aplicação que usa *middleware* para um serviço de requisição de dados entre um usuário e uma base de dados emprega uma arquitetura multicamadas em três módulos, ou seja, uma arquitetura de três camadas - camada de apresentação, camada lógica e camada de dados. Um número considerável de novos métodos e tecnologias provenientes da área de federação de bases de dados e sistemas de mediação estão sendo aplicados na integração de bibliotecas digitais. Essas experiências têm em comum uma arquitetura multicamadas que provê adaptadores e/ou *wrappers* para que fontes de dados distribuídas possam

11 Disponível em: http://www.fedora.info. Acesso em: 23 ago. 2021.

12 Disponível em: http://www.loc.gov/standards/sru/srw/index.html. Acesso em: 1 set. 2008.

13 Disponível em: http:www.brickscommunity.org/. Acesso em: 23 ago. 2021.

lidar com o problema de homogeneização de interface e representação de dados (SCHALLEHN; ENDING, 2000).

7 Direitos autorais

"Copyright é a mais irritante barreira no desenvolvimento das bibliotecas digitais" (CHEPESUIK, 1997, p.49).

Considerando que, de forma geral, uma federação de bibliotecas digitais oferece acesso a diversos repositórios digitais, um passo importante na consolidação de modelos de federação é a definição dos papéis dos usuários e dos seus direitos de acesso. Há, entretanto, um consenso absoluto por parte de toda a comunidade envolvida de que a gestão de direitos é um dos mais complexos e desafiadores problemas que a área de bibliotecas digitais tem que enfrentar. Discutir direitos conduz forçosamente para o território legal e on de negócios, os quais as bibliotecas, cuidadosamente, procuraram evitar no passado (COYLE, 2004a). As questões de direitos autorais (*copyright*) e propriedade intelectual foram sempre um problema difícil para as bibliotecas. Elas sempre tiveram que se equilibrar entre as leis de *copyright*, o conceito de fair use, ou seja, de uso razoável ou uso aceitável, e a doutrina da primeira alienação (*first sale doctrine*). O *fair use*, numa definição bem simples, permite que obras sejam copiadas para propósitos educacionais e empréstimo entre bibliotecas; por sua vez, a doutrina da primeira alienação tornou possível a existência das bibliotecas públicas de pesquisa modernas (I-DLR, 2003). Com o surgimento das bibliotecas digitais e do *e-commerce* e das novas configurações do mercado de conteúdo, essa questão torna-se crítica, pois se constata que o conceito tradicional de direito autoral não se ajustou no ambiente digital, dado que o controle de cópias, de integridade e acesso foi perdido: os objetos digitais são menos fixáveis, facilmente copiados e remotamente acessíveis por múltiplos usuários simultaneamente.

A maioria dos idealizadores de projetos importantes de bibliotecas digitais está consciente de que existem problemas de propriedade intelectual que devem ser resolvidos para que as bibliotecas digitais desenvolvam suas potencialidades plenamente. Algumas propostas expressam a intenção de resolver essas questões como parte do planejamento geral dos seus sistemas, embora sem muita especificidade de como isso possa ser praticado efetivamente (SAMUELSON, 1995).

Os sistemas de bibliotecas e arquivos digitais atuais têm servido bem o seu público dentro dos limites da mídia impressa. É necessário, agora, estender esses modelos aonde for possível, e inventar novos modelos onde for necessário, para que se possa oferecer acesso aos artefatos digitais. De forma geral, as bibliotecas

digitais necessitam de licenças flexíveis e inovadoras que lhes permitam, de forma legal, criar arquivos e coleções, gerar serviços compatíveis com as necessidades atuais e futuras de seus usuários e praticar estratégias apropriadas de preservação digital. O desafio atual é estabelecer os papéis, os direitos e as responsabilidades das bibliotecas e arquivos no que concerne à disponibilização do acesso público à informação digital. Pesquisas contínuas são necessárias para dar às bibliotecas a capacidade de gerenciar propriedade intelectual e proteger esses direitos sem inibir o legítimo acesso dos usuários aos materiais qualificados com esses direitos.

Embora seja comum para os utópicos do mundo tecnológico declarar a Internet, e por analogia todas as mídias digitais, uma zona livre de *copyright*, para o bem ou para o mal, as leis de direitos autorais aplicam-se a todas as formas de propriedade intelectual, estejamos ou não de acordo com elas (COYLE, 2004a). O problema para as bibliotecas é que, diferentemente de um negócio privado ou de uma editora, na maioria das vezes, elas são simplesmente custodiantes da informação – as bibliotecas não detêm os direitos sobre o material de seu acervo. É improvável que bibliotecas possam livremente digitalizar e prover acesso a materiais detentores de *copyright* da sua coleção. Ao invés disso, terão que desenvolver mecanismos para gerenciar esses direitos, procedimentos que permitam que elas disponibilizem informação sem violar as regras do direito autoral e da propriedade intelectual - tais procedimentos são chamados coletivamente de gestão de direitos autorais. A indústria de conteúdos, especialmente a de entretenimento, vem desenvolvendo tecnologias que permitam um controle mais severo no ambiente digital. É necessário, entretanto, observar que as soluções para a mídia de entretenimento terão grande efeito sobre bibliotecas. Esses efeitos vão impactar as bibliotecas de duas formas: como consumidoras de produtos de informação comercialmente produzidos e como distribuidoras de recursos para o público em geral (COYLE, 2004b).

DRM é a sigla para *Digital Right Management* ou Gerenciamento de Direitos no Ambiente Digital (tradução dos autores), termo usado para as tecnologias que controlam como os conteúdos digitais são acessados e usados. Enquanto os detentores de direitos autorais têm o direito exclusivo sobre uma obra – tais como o direito de produzir uma cópia ou de distribuí-la para o público – na maioria das vezes eles não têm o direito de controlar como a obra será usada (por exemplo, o direito de ver uma obra ou de ler um trabalho). Os proprietários de conteúdos vislumbram as tecnologias de DRM como um meio de controlar o uso dos seus conteúdos; por outro lado, os usuários e os bibliotecários as avaliam como uma ameaça ao conceito de *fair use* (ALA, 2003).

No contexto de mudanças, é necessário observar o evidente interesse da academia por modelos abertos de publicação, como o *Open Archive Initiative* (OAI)

e *o Budapest Open Access Initiative* (BOAI), como resposta ao custo crescente dos periódicos comerciais e práticas restritivas de publicação. Os modelos de publicações abertas certamente exigem novas estratégias de DRM que enfatizem o *fair use*, a proteção da propriedade intelectual de usos inadequados e os modelos de subscrição múltiplas, que incluam acessos taxados e não taxados. Além do mais, a comunidade envolvida na área de Pesquisa e Ensino (P&E) considera que as soluções comerciais de DRM afetam o frágil equilíbrio entre os que controlam os direitos autorais e os usuários, em favor dos primeiros, quando questionam os conceitos de *fair use* e *first sale* – dois princípios críticos e preciosos no contexto da P&E. Existe ainda a preocupação – bastante pertinente – de que algumas implementações de DRM comprometam a privacidade dos usuários (MARTIN *et al.*, 2002).

Em face dessa realidade, há um esforço em andamento, no âmbito das comunidades de pesquisa em rede e bibliotecas digitais, para desenvolver soluções inovadoras de DRM a fim de dar apoio às atividades de ensino e pesquisa e aos novos modelos de publicações acadêmicas. Martin e seus colaboradores relatam (MARTIN *et al.*, 2002) um projeto que tem como foco específico a apresentação de uma arquitetura de referência voltada para a implementação de um sistema de DRM Federado (FDRM). O objetivo desse projeto é dar apoio ao compartilhamento de recursos em nível local e interinstitucional de uma forma discricionária, segura e com privacidade, enquanto se esforça para manter o equilíbrio entre os direitos do usuário e os dos proprietários dos conteúdos. A solução deve estar em consonância com as exigências das áreas de P&E, e devem incluir: acomodação dos aspectos intensamente colaborativos e distribuídos da maioria das atividades do mundo acadêmico; apoio ao *fair use* de materiais protegidos por *copyright* destinados a propósitos educacionais; suporte ao acesso diferenciado e granular aos recursos; prevenção ao mau uso dos recursos; garantia da integridade do recurso; e interoperabilidade com as infraestruturas existentes e as emergentes.

Os requisitos preconizados pelos acadêmicos revelam uma distinção em relação ao modelo convencional de DRM – o modelo voltado para o comércio eletrônico, cuja função é primordialmente proteger os direitos do proprietário –, indicando que um conceito mais abrangente possa emergir incorporando gestão de acesso, bem como gestão dos direitos de propriedade intelectual e a preocupação com os direitos do usuário, assim como o direto dos proprietários. Dada essa diferença na interpretação, é possível verificar que o termo DRM, embora tenha sido adequadamente apropriado pela indústria de publicações, talvez não seja adequado para expressar os objetivos da comunidade acadêmica que poderiam ser mais bem traduzidos por outro termo (MARTIN *et al.*, 2002). Novas formas de direito de uso também veem sendo criadas e utilizadas em diversas instâncias e devem

ser seriamente consideradas como alternativas aos modelos tradicionais, como a conhecida licença *Creative Commons*[14], e mais recentemente a *Science Commons*[15]. *Creative Commons* é um tipo de licença flexível de *copyright* para obras intelectuais surgidas com a Internet. Ela abre a possibilidade de publicar e disponibilizar na rede os mais diferentes tipos de trabalho intelectual, de modo a permitir a cópia e reuso desses conteúdos por terceiros sob determinadas condições. Ao invés da tradicional enunciação de *copyright* e todos os direitos reservados, a licença *Creative Commons* pretende permitir a cópia ampla, reuso, modificação, desenvolvimento e ampliação do trabalho intelectual original, desde que sejam garantidos alguns direitos. A partir disso, existem gradações, incluindo a redistribuição, alteração, uso comercial, entre outros. Há derivações da licença *Creative Commons* para trabalhos científicos, o *Science Commons* e para recursos educacionais, o *Open Educational Resources*[16], que permitem acesso aberto aos recursos correspondentes (KORN; OPPENHEIN, 2006).

8 Parâmetros de avaliação

Um dos principais problemas em relação à interoperabilidade é que comparar soluções é uma tarefa muito difícil, posto que diferentes abordagens operam baseadas em concepções distintas e com objetivos algumas vezes conflitantes. Mesmo assim, é importante estabelecer alguns critérios para avaliação de soluções de interoperabilidade e compreender os *trade-offs* entre eles; no mínimo, isso nos ajuda a compreender melhor o problema de bibliotecas digitais distribuídas e interoperáveis. Paepcke *et al*., (1998) destacaram, entre tantos, seis critérios, cujas interpretações são as seguintes:

a) **Grau de autonomia** – é um dos critérios mais críticos, pois tem forte impacto nos processos internos das organizações. Refere-se à quantificação da conformidade com as regras globais, que é exigida de cada componente participante do sistema, por exemplo, um serviço. Um alto grau de autonomia é desejável, pois pressupõe maior controle local sobre a implementação e a operação do componente e torna mais fácil a inclusão de sistemas legados como componente participante. Entretanto, a autonomia total pode gerar um descomprometimento com as regras globais – interfaces, protocolos de interação e formatos de dados não padronizados que podem ser

14 Disponível em: http://creativecommons.org/. Acesso em: 23 ago. 2021.
15 Disponível em: http://sciencecommons.org/. Acesso em: 23 ago. 2021.
16 Disponível em: http://oercommons.org/. Acesso em: 23 ago. 2021.

arbitrariamente mudados. No outro extremo, os participantes do sistema têm que obrigatoriamente engajar-se nos procedimentos globais do sistema. Sistemas interoperáveis que funcionam na prática recaem sobre esses dois extremos;

b) **Custo da infraestrutura** – corresponde ao custo da infraestrutura necessária para dar sustentação a uma solução, e para que um componente de serviço possa incorporar-se ao sistema. Corresponde ao custo de infraestrutura necessário para dar sustentação ao sistema e para a incorporação de novos participantes;
c) **Facilidade de incorporação de componentes** – esse critério, diferentemente do anterior, refere-se ao custo incremental de tornar interoperável um novo componente de serviço. Esse custo incremental pode envolver investimentos em *hardware*, ou pode estar expresso na forma de complexidade do *software* necessário para assegurar a interoperabilidade;
d) **Facilidade de uso** – esse critério refere-se a dois itens: a complexidade de se criar um *software* cliente para um componente de serviço, ou seja, uma interface, e a complexidade de interagir com o componente quando em execução;
e) **Dimensão da complexidade das tarefas suportadas** – está relacionado à capacidade do sistema de desenvolver, gerenciar, tornar interoperáveis e operar componentes que tenham como substrato tecnologias e padrões de alto nível de sofisticação;
f) **Escalabilidade** – mede a capacidade do sistema de absorver crescentemente novos componentes de serviço.

Dependendo da dimensão que se deseja mensurar, outros itens podem ser incorporados na matriz de avaliação, incluindo parâmetros que avaliem impactos sociais, como nível de inclusão e impactos nos resultados de pesquisa; entretanto será sempre uma tarefa difícil de cumprir com precisão.

9 Considerações finais

O estudo e o desenvolvimento de bibliotecas digitais, especialmente da sua vertente considerada por muitos especialistas como a mais sofisticada, que é a interoperabilidade, colocam em evidência, trazem para a prática e para o convívio com as tecnologias atuais quase todos os fundamentos da Ciência da Informação. A pesquisa na área de bibliotecas digitais, pelo seu domínio amplo e pela sua complexidade, coloca-nos face a face com problemas que quase sempre só tivemos oportunidade de lidar no plano das ideias, e nos desafia a todo instante a recondicioná-los para uma nova realidade. Mas, tal qual a esfinge, as questões da área de bibliotecas

digitais desafiam-nos com seu olhar de pedra, e são ainda muitos os enigmas que teremos que decifrar. Se não, vejamos alguns deles:

- **Federação**

A próxima geração de bibliotecas digitais deverá ser formada por sistemas de menor dimensão, de natureza diversificada (arquivos, museus, galerias, etc.), autônomos e independentes administrativamente, cada qual oferecendo – por vários meios e tecnologias - funcionalidades diferentes e acesso a diferentes conteúdos. A federação deve trabalhar com diferentes níveis de interoperabilidade, operando simultaneamente, como o caso de uso discutido no início deste artigo, que citava o problema de integrar conteúdos em cultura brasileira e língua portuguesa sobre "Brasil colônia e a influência da cultura negra". A federação dessas bibliotecas digitais preencherá as necessidades de espaços informacionais semanticamente ricos e interconectados, dando margem ao surgimento de serviços de informação inovadores.

- **Interface única e transparência para o usuário**

Requer uma interface unificada que oculte toda a complexidade intrínseca à federação de bibliotecas digitais e proporcione um mecanismo universal para mapear e integrar cada repositório digital nessa interface, dando ao usuário a impressão de uma biblioteca única; a localização geográfica de cada fonte de informação deve ser indiferente para o usuário.

- **Gestão de direitos**

Para tornar viável o acesso universal aos estoques informacionais das bibliotecas digitais acadêmicas, é necessário ampliar: os modelos atuais de gestão de direitos, quando for possível, e desenvolver novos modelos que preservem os conceitos de fair use e da *first sale doctrine*, essenciais para o desenvolvimento científico em escala global; desenvolver modelos automatizados de DRM que considerem, além dos direitos dos proprietários de materiais protegidos por *copyright*, os direitos de acesso individuais e institucionais dos usuários, preservando sua privacidade.

- **Expansibilidade**

As bibliotecas digitais deverão ser sistemas sempre em expansão; para tanto é necessário criar arquiteturas abertas e flexíveis que possibilitem a criação e a in-

corporação contínua de novos componentes de serviços. Todas as funcionalidades das bibliotecas deverão ser particionadas em um conjunto de serviços bem definidos, autodescritos, autoregistráveis e autoconfiguráveis, de forma que possam ser (semi-) automaticamente registrados e configurados ao serem "ligados" ao sistema na forma *Plug-and-Play* (MARTINEZ, 2006). A integração das bibliotecas deve ser suficientemente flexível para permitir que cada biblioteca possa, individualmente, adicionar e/ou modificar características e componentes de serviço;

- **Interoperabilidade semântica e sintática**

Dada a natureza heterogênea da nova geração de bibliotecas digitais, a interoperabilidade deve ser a preocupação central dos requisitos dos sistemas. Os metadados e os *softwares* de interfaces devem ser automaticamente mapeados, de forma que a heterogeneidade sintática e a semântica possam ser resolvidas. As ontologias têm um papel-chave na resolução da interoperabilidade semântica e as pesquisas nessa área vão proporcionar a infraestrutura necessária para o atendimento a essa necessidade, especialmente no que tange à heterogeneidade semântica (MARTINEZ, 2006).

- **Arquitetura e sistemas de mediação**

É de fundamental importância para as pesquisas em interoperabilidade de repositórios digitais explorar concepções novas de arquiteturas para sistemas de mediação para federação, bem como tecnologias da área de T.I. Agentes inteligentes, arquitetura multicamadas (*multi-tier architecture*), CORBA, *web services*, *wrappers* são tecnologias potencialmente importantes para o desenvolvimento de *middleware* para federação de bibliotecas digitais.

- **Aspectos sociais, políticos e culturais**

Por último, mas não menos importante, estão as pesquisas que investigam os impactos da federação de acervos digitais na educação e na cultura, mas precisamente como ferramenta básica nas áreas de educação e treinamento a distância; no apoio aos programas de inserção digital e de cidadania; e como meio para potencializar a visibilidade digital de manifestações culturais e artísticas.

Referências

ALA. **Digital right management and libraries**. Washington, DC: ALA, 2003. Disponível em: http://www.ala.org/ala/washoff/WOissues/copyrightb/digitalrights/digitalrightsmanagement.htm. Acesso em: 6 ago. 2021.

ARMS, W.Y. **Thoughts about interoperability in the NSDL**: draft for discussions. August 2000. Disponível em: http://www.cs.cornell.edu/wya/papers/NSDL-Interop.doc. Acesso em: 6 ago. 2021.

ARMS, W.Y. *et a*l. A spectrum of interoperability: the site for science for prototype for the NSDL. **D-Lib Magazine**, v.8, n.1, 2002. Disponível em: http://www.dlib.org/dlib/january02/arms/01arms.html. Acesso em: 6 ago. 2021.

BARTELT, A. *et a*l. Buiding infrastructure for digital library. In: DELOS NETWORK OF EXCELLENCE WORKSHOP ON INTEROPERABILITY AND MEDIATION IN HETEROGENEOUS
DIGITAL LIBRARIES, 3., 2001, Darmstadt. **Proceedings**... European Research Consortium For Informatics And Mathematics, 2001. 5 p. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.7359&rep=rep1&type=pdf. Acesso em: 6 ago. 2021.

BARU, C. *et al*. XML-based information mediation for digital libraries. In: ACM CONFERENCE ON DIGITAL LIBRARIES, 4., 1999, Berkeley. **Proceedings**... ACM, 1999. p.214-215. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.452.122&rep=rep1&type=pdf. Acesso em: 6 ago. 2021.

BERNERS-LEE, T., HENDLER, J., LASSILA, O. **The semantic web**. Scientific American, May, 2001. Disponível em: https://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf. Acesso em: 6 ago. 2021.

BIRMINGHAM, W. An agent-based architecture for digital libraries. **D-Lib Magazine**, July 1995. Disponível em: http://www.dlib.org/dlib/July95/07birmingham.html. Acesso em: 6 ago. 2021.

CHEPESUIK, R. The future is here: America's Libraries go Digital. **American Libraries**, v.2, n.1, p.47-49, 1997.

COYLE, K. The "Rights" in the digital rights management. **D-Lib Magazine**, v.10, n.9, 2004a. Disponível em: http://www.dlib.org/dlib/september04/coyle/09coyle.html. Acesso em: 6 ago. 2021.

COYLE, K. Rights Management and digital library requirements. **Ariadne**, n.40, 2004b. Disponível em: http://www.ariadne.ac.uk/issue40/coyle. Acesso em: 6 ago. 2021.

DOERR, M. **Increasing the Power of Semantic Interoperability for the European Library**. Ercim News Online Edition, 2006. Disponível em: http://

www.ercim.org/publication/Ercim_News/enw66/doerr.html. Acesso em: 6 ago. 2021.
FONSECA, F.; ENGENHOFER, M.; BORGES, K. Ontologias e interoperabilidade semântica entre SIG´s. *In*: GEOINFO 2000 - WORKSHOP BRASILEIRO DE GEOINFORMÁTICA, 2., 2000,
São Paulo. **Anais**... Disponível em: http://mtc-m16c.sid.inpe.br/col/dpi.inpe.br/vagner/2000/07.04.15.32/doc/011.pdf. Acesso em: 6 ago. 2021.
GARDNER, T. An introduction to web services. **Ariadne**, v.29, 2001. Disponível em: http://www.ariadne.ac.uk/issue29/gardner/. Acesso em: 6 ago. 2021.
GRUBER, T. **What is an ontology**: knowledge system aI laboratory, 1996. Disponível em: http://www-ksl.stanford.edu/kst/what-is-an-ontology.html. Acesso em: 6 ago. 2021.
HATALA, M. *et al*. The interoperability of learning object repositories and services: standards, implementations and lessons learned. *In*: INTERNATIONAL WORLD WIDE WEB CONFERENCE, 13., 2004, New York. **Proceedings**...
New York, ACM, 2004. p. 19 - 27. Disponível em: https://www.researchgate.net/publication/221023617_The_interoperability_of_learning_object_repositories_and_services_Standards_implementations_and_lessons_learned. Acesso em: 6 ago. 2021.
I-DLR – Interactive Digital Library Resource Information System. **Copyright issues and intellectual properties rights in digital libraries**. 2003. Disponível em: http://www.coe.missouri.edu/~DL/iDLR/viewpaper.php?pid=20. Acesso em: 15 mar. 2007.
KORN, N.; OPPENHEIN, C. Creative Commons Licences in Higher and Further Education. **Ariadne**, n.46, 2006. Disponível em: http://www.ariadne.ac.uk/issue/49/korn-oppenheim/. Acesso em: 6 ago. 2021.
MARCONDES, C.H.; SAYÃO, L.F. Integração e interoperabilidade no acesso a recursos informacionais eletrônicos em C&T: a proposta da Biblioteca Digital Brasileira. **Ciência da Informação**, v.30, n.3, p.24-33, 2001. Disponível em: http://www.scielo.br/pdf/ci/v30n3/7283.pdf. Acesso em: 6 ago. 2021.
MARTIN, M. *et al*. Federated Digital Rights Management. **D-Lib Magazine**, v.8, n.7/8, 2002. Disponível em: http://www.dlib.org/dlib/july02/martin/07martin.html. Acesso em: 6 ago. 2021.
MARTINEZ, R.J.F. Agents and ontologies working together to federate digital libraries. **TCDL Bulletin**, v.2, n.2, 2006. Disponível em: https://bulletin.jcdl.org/Bulletin/v2n2/rodriguez-martinez/rodriguez-martinez.html. Acesso em: 6 ago. 2021.
MELNIK, S.; GARCIA-MOLINA, H.; PAEPCKE, A. A Mediation infraestructure

for digital library services. *In*: ACM CONFERENCE ON DIGITAL LIBRARIES, 5., 2000, San Antonio. **Proceedings**... ACM, 2000. p.123-132. Disponível em: http://infolab.stanford.edu/~melnik/pub/dlo0.pdf. Acesso em: 6 ago. 2021.
MILLER, P. Interoperability. What is it and why should I want it? **Ariadne**, n.24, 2000. Disponível em: http://www.ariadne.ac.uk/issue24/interoperability/. Acesso em: 18 mar. 2007.
NISO**. Z39.50**: A primer on the protocol. Bathesda, MD: NISO Press, 2002. Disponível em: http://www.niso.org/standards/resources/Z3950_primer.pdf. Acesso em: 6 ago. 2021.
ODLIS. 2004. Disponível em: https://products.abc-clio.com/ODLIS/odlis_g.aspx. Acesso em: 6 ago. 2021.
PAEPCKE, A. *et al*. Interoperability for digital libraries worldwide. **Communication of the ACM**, v.41, n.4, p.33-43, 1998. Disponível em: http://eolo.cps.unizar.es/docencia/doctorado/Articulos/DiL/CACM-Abril1998/p33-paepcke.pdf. Acesso em: 6 ago. 2021.
PAMPLONA, V.F. **Web Services**: construindo, disponibilizando e acessando web services via J2SE e J2ME. Javafree.org, 2004. Disponível em: http://www.deinf.ufma.br/~mario/pos/corba/tutorial_axis_ws.pdf. Acesso em: 6 ago. 2021.
PAYETTE, S. *et al*. Interoperability for digital objects and repositories: the Cornell/CNRI Experiments. **D-Lib Magazine**, v.5, n.5, 1999. Disponível em: http://dlib.org/dlib/may99/payette/05payette.html. Acesso em: 6 ago. 2021.
PIRRI, M.; PETTENATI, M.C.; GIULI, D. **Design of a federation service for digital libraries**: the case of historical archives in the PORTA EUROPA Portal (PEP) Pilot Projec. *In*: DC-2002: METADATA FOR E-COMMUNITIES, Florence. Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002. Florence: Firenze University Press, 2002. p.157-162. Disponível em: https://dcpapers.dublincore.org/pubs/article/view/706. Acesso em: 6 ago. 2021.
SAMUELSON, P. Copyright and digital libraries. **Communication of the ACM**, v.38, n.3, April 1995. Disponível em: http://doi.acm.org/10.1145/205323.205324. Acesso em: 15 mar. 2007.
SAYÃO, L.F.; MARCONDES, C.H. **Guia de software em automação de bibliotecas**. Brasília: MEC/SESu/PNBU, 1989.
SCHALLEHN, E.; ENDING, M. **Using source capability description for the integration of digital libraries**. 2000. Disponível em: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.22.3395&rep=rep1&type=pdf. Acesso em: 6 ago. 2021.
SURE, Y.; STUDER, R. **Semantic web technology for digital libraries**, 2005. Disponível em: http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/2005_

sw_for_dl.pdf. Acesso em: 2 maio 2007.
TENNANT, R. Interoperability: the holy grail. **Library Journal**, 1998. Disponível em: http://www.libraryjournal.com/article/CA156495.html. Acesso em: 5 jan. 2007.
THE OPEN ARCHIVES INITIATIVE PROTOCOL FOR METADATA HARVESTING. Protocol Version 2.0 of 2002-06-14. Disponível em: http://www.openarchives.org/OAI/openarchivesprotocol.html. Acesso em: 22 mar. 2007.
UKOLN. **Interoperability focus**: looking at interoperability. 2005. Disponível em: http://www.ukoln.ac.uk/interop-focus/about/leaflet.html. Acesso em: 6 ago. 2021.
WARREN, P.; THURLOW, I.; ALSMEYER, D. Applying semantic technology to a digital library. *In*: DAVIS, J.; STUDER, R.; WARREN, P. **Semantic Web tecnologies**: trends and research in ontology- based systems. Sussex: John Wiley, 2006. p.237-257.
WEINSTEIN, P.C. Ontology-based metadata: transforming the MARC legacy. *In*: ACM conference on Digital Libraries 3., Pittsburgh, **Proceedings**… Pittsburgh: ACM, 1998. p.254-263. Disponível em: https://www.researchgate.net/publication/2283709_Ontology-Based_Metadata_Transforming_the_MARC_Legacy. Acesso em: 6 ago. 2021.
ZUBAIR, M. *et al*. Dynamic construction of federated digital libraries. *In*: INTERNATIONAL WORLD WEB CONFERENCE THE WEB: THE NEXT GENERATION, 9., Amsterdam, **Proceedings**... Amsterda, 2000. Disponível em: http://www9.org/final-posters/poster17.html. Acesso em: 2 mar. 2007.

PARTE 3

DADOS NA PESQUISA CIENTÍFICA

Bases de dados: a metáfora da memória científica

Luís Fernando Sayão

1 Introdução

> *"Fazemos apelo aos testemunhos para fortalecer ou debilitar, mas também para completar, o que sabemos de um evento do qual já estamos informados de alguma forma, embora muitas circunstâncias nos permaneçam obscuras".*
> (HALBWACHS, 1990)

É com essas exatas palavras que Maurice Halbwachs (1990) inicia seu livro publicado postumamente, a Memória Coletiva. Por uma coincidência extremamente provocante, é também desta forma, ou melhor, é no estado que poderia ser descrito precisamente por essas mesmas palavras de *Halbwachs*, que um pesquisador, um cientista interroga um banco de dados à procura de informações que insiram seu trabalho de pesquisa na ciência feita pelo seu grupo. Isto é, ele procura um azimute, um quadro de referências que faça com que ele possa reconstruir seu conhecimento sob a luz dos testemunhos dos seus pares e orientar o seu trabalho no sentido estabelecido pela comunidade científica ou acadêmica em que ele está ou deseja estar inserido.

Este estado transitório, no qual se encontra este pesquisador, é chamado por alguns autores da área de ciência da informação de "estado anômalo de conhecimento" (belkin, 1980). Ele é caracterizado por um alto grau de indefinição em relação ao assunto sobre o qual o pesquisador procura informações. O seu próprio desejo de informação é absolutamente nebuloso, fazendo com que suas interrogações só consigam se realizar durante o ato da busca. O processo de interação com os conhecimentos armazenados na base é que estabelece o foco da questão. A percepção do pesquisador sobre o conhecimento, fatos e dados armazenados nestes meios eletrônicos, articulados com os seus próprios conhecimentos, recria cenários de conhecimentos mais nítidos.

Wittgenstein (1961) diz no seu *Tractatus* Logico-*Philosophicus* (1922) que "a dúvida, pois, só existe onde existe uma questão, uma questão apenas onde existe uma resposta, e esta somente onde algo pode ser dito". Ninguém interroga uma base de dados sobre o que não conhece. Quando um cientista busca informações em uma base de dados, ele só está querendo validar as suas dúvidas, traduzindo-as por questões autenticadas por seus pares, por questões presentes na memória coletiva de sua tribo. A resposta já existe. Fundamentar sua questão sobre o que já foi estabelecido é uma imposição do método científico, da natureza tribal e cumulativa da ciência, sem o que o pesquisador está condenado ao limbo da rejeição e do esquecimento, e o seu saber ao descrédito. O que ele busca é fundamentalmente enquadrar sua contribuição à ciência comum, consensual. Isto não quer dizer em absoluto que os debates e as controvérsias se aniquilam diante dessa imposição; quer dizer sim que toda discussão, por mais dramática e acalorada que seja, está submetida a um ritual, cuja consulta aos antecedentes, à memória coletiva comum, é uma das etapas privilegiadas.

Quando um pesquisador, diante de um microcomputador ligado a um banco de dados que pode estar em qualquer parte do mundo, vasculha suas estantes eletrônicas à procura de informações que definam, completem ou estabeleçam as fronteiras do seu trabalho de pesquisa, ele repete o mesmo gesto de quem mergulha na memória de seu grupo para reconstruir as lembranças comuns e dessa forma manter íntegra a sua comunidade.

Isto nos leva a pensar que as bases de dados, com seus complexos esquemas de representação e de recuperação de informação, que hoje encerram praticamente todos os testemunhos' da ciência moderna, constituem a memória consensual desta ciência; a memória eletrônica de que nenhum cientista pode prescindir para ordenar e reconstruir seus conhecimentos e onde, obrigatoriamente, precisa haver suas contribuições, seus testemunhos inseridos, sob pena de não participar dessa memória coletiva e não ser jamais "lembrado", ou melhor, citado por seus colegas.

Talvez pudéssemos pensar nas bibliotecas especializadas cumprindo este mesmo papel de memória dos saberes científicos. Mas, por mais universal que seja seu acervo, ele não consegue reunir todas as obras de uma área do conhecimento e sua seletividade está baseada em critérios "domésticos", específicos de uma instituição ou um programa. Porém, a diferença mais marcante é que as bases de dados são suportadas por uma tecnologia que permite a simulação, a ilusão de um diálogo, de uma interação em tempo real em uma linguagem que está cada vez mais próxima das linguagens naturais. Existem as interfaces inteligentes, o hipertexto, ajudas sensíveis ao contexto, janelas que criam uma atmosfera conversacional. Esta interação pode ser feita de qualquer lugar onde se tenha um microcomputador e um telefone.

É um gesto solitário do cientista à procura dos seus antecedentes; ao passo que a biblioteca interpõe sempre, entre o pesquisador e o acervo, intermediários, vidraças, catálogos, códigos indecifráveis como são as referências bibliográficas.

É sobre este simulacro da memória coletiva científica engendrado pela sociedade pós-industrial - que são as bases de dados e sua relação com os aspectos cumulativo, social e institucional da ciência - que eu gostaria de discutir rapidamente neste texto. Em especial, no que diz respeito à contribuição da memória virtual ao controle e enquadramento da produção científica, que vão desde os limites impostos pelos mecanismos de representação do conheci- mento até os desvios ideológicos que conduzem deliberadamente à marginalização de autores, conhecimentos, fatos não identificados com os padrões da ciência oficial.

2 A formação da memória eletrônica

O caráter cumulativo da ciência, que se apropria de uma forma rigorosamente seletiva das contribuições de seus pesquisadores, resulta em um corpo de conhecimento baseado no consenso. Este corpo de conhecimento é representado pela literatura técnico-científica, fruto mais óbvio e mais facilmente sujeito à mensuração da atividade científica. São os livros, os artigos de revistas, os trabalhos de congresso, as patentes, portadoras das inovações tecnológicas, os mais autênticos registros da faina diária dos cientistas.

O crescimento vertiginoso da capacidade de armazenamento de dados em meios legíveis por computador- meios magnéticos e, mais recentemente, meios óticos - não foi ainda suficiente para tornar viável o armazenamento generalizado em computador dos conhecimentos gerados pela atividade científica. Esta impossibilidade implica que, para se colocar disponível em sistemas on-line, ou seja, sistemas que permitam um certo grau de interação, de conversação em tempo real, é mandatário que o conhecimento registrado na literatura sofra um processo de tradução, de representação, transformando-se em meta conhecimento. Esta tradução é realizada por intermédio de esquemas simbólicos que descrevem tanto a forma física, quanto o conteúdo informacional das obras que estão sendo registradas. O que vai ser armazenado nas grandes bases de dados é uma metáfora da informação original, é o conhecimento virtual, que só existe em função do seu referente, da sua vinculação remota com algum conhecimento real.

A criação dessas representações é factível via linguagens documentárias, que são linguagens artificiais geralmente derivadas da linguagem natural. Essas linguagens são chamadas artificiais no sentido em que não resultam de processo evolutivo e necessitam de regras explícitas para seu uso. Normalmente as linguagens documentárias estabelecem uma relação unívoca entre o termo e o conceito, isto

é, entre o significante e o significado. Cada termo corresponde a um conceito do sistema de conceitos da área específica com que se está trabalhando. A construção dessas linguagens é um processo complexo e longo (GOMES, 1990).

As linguagens documentárias são, pela sua própria artificialidade, extremamente redutoras de significado e só podem cobrir conceitos de um domínio específico do conhecimento humano, não havendo, portanto, linguagens documentárias gerais. Elas fazem parte intrínseca dos sistemas de informação, afetando e sendo por eles afetada.

Qualquer linguagem, já se sabe, é uma forma de poder, de dominação. As nossas próprias contradições culturais são um exemplo disso. A linguagem documentária não é exceção. O poder avassalador no sentido da ordenação, da organização que ela exerce sobre a produção literária, especialmente a científica, é chamado pelo semiólogo Umberto Eco de a "ditadura dos resumos". Esta característica é de dramática importância na construção da memória eletrônica da ciência, pois o grau de resolução e entendimento dos conhecimentos que ela apropria está limitado pela capacidade de representação do código. Mas muita atenção: a propriedade redutora de significado dessas linguagens, antes de ser uma deficiência, é o sustentáculo da identidade, do poder de ordenação e classificação, do qual a ciência não pode prescindir, e é, sobretudo, o canal de expressão da ideologia que a ciência suporta. É o seu poder uniformizador que elimina as diferenças desagregadoras, que garante a harmonia na formação das diversas memórias possíveis.

Um exemplo corriqueiro pode deixar mais clara essa relação entre poder uniformizador e controlador da linguagem, a ideologia que ela representa e a formação da memória eletrônica. Apenas 5% da produção científica dos países periféricos está presente nas grandes bases de dados internacionais (GAILLARD, 1989). Nestas bases, está representada essencialmente a ciência do Primeiro Mundo e os códigos de representação existentes estão voltados para a problemática desses países. Este fato tem grande impacto no armazenamento e recuperação de informações e conhecimentos que estão fora do domínio da ciência e tecnologia primeiro mundistas. Por exemplo: o desejo de incorporar, em uma base de dados internacional sobre fontes de energia, um artigo importante sobre o uso de óleo de dendê e de jojoba produzidos no interior da Bahia como combustível automotivo em substituição ao óleo diesel poderá esbarrar na falta de termos adequados para a representação correta desses óleos e de sua ambientação, resultando em distorções na representação e consequentes desvios na recuperação. Isto significa que os próprios limites da linguagem documentária farão com que sejam preservados a uniformidade e o caráter primeiro mundista dessa memória eletrônica. Se este artigo chegar um dia a fazer parte dessa base, ele dificilmente será recuperado e, como

desdobra- mento, não será citado por seus pares, ou pelo paralelo que traça este texto, não será "lembrado", pois, insisto, os limites da linguagem determinam o seu esquecimento à medida que o esquema simbólico utilizado é incapaz de expressar com nitidez o conhecimento que ele porta.

Neste ponto, talvez, caiba um parêntese. Aprofundando um grau a mais o paralelo que aqui se traça, quando penso no inverso de tudo isto, ou seja, no acesso a bases de dados internacionais e na recuperação de suas informações, penso sempre que podemos estar, quem sabe, apropriando-nos da memória científica de outras tribos com todas as suas idiossincrasias. Dependendo do uso que se faça das informações contidas nestas bases, esta prática pode constituir um reforço na dependência de paradigmas científicos e tecnológicos, com os quais somos obrigados a conviver por todas as nossas contradições culturais, históricas e políticas. O acesso a estas bases nos deixa sempre a um passo de importação de problemas científicos e tecnológicos estranhos e, muitas vezes, irrelevantes para a nossa realidade terceiro mundista, que pode estender nosso esforço de pesquisa para limites irreais e indesejáveis. Mesmo assim, o acesso a estas informações é de dramática importância para os pesquisadores dos países menos desenvolvidos. Na maioria das vezes, elas são as únicas fontes estruturadas de informações disponíveis, seja pela via on-line ou em bases de dados em CD-ROM, posto que ainda não está plenamente presente, na consciência dos que conduzem a política de informação desses países, a preocupação com a formação de bases de dados nacionais que sigam padrões internacionais, que façam uso de linguagens documentárias adequadas e que sejam a expressão mais completa da ciência e tecnologia praticada por esses países. Essas bases de dados nacionais deveriam reunir os testemunhos da atividade de pesquisa de países ou de regiões, tal como faz a base de dados *Lilacs*, de forma a reconstruir, para quem as consulta, conhecimentos, cenários, ambientes, fatos e dados pertinentes a um universo próprio. Uma outra questão possível é que a criação de bases de dados bem estruturadas e com um nível de padronização satisfatória talvez seja a forma mais conveniente de tornar visíveis para a comunidade científica internacional a atividade de pesquisa de países ou regiões em desenvolvimento e inseri-la com identidade própria e autenticidade na "grande memória eletrônica".

Entretanto, os processos de exclusão dos conhecimentos gerados nos países periféricos da *mainstream science* são ainda mais variados e estão muito além dos limites impostos pelos códigos de representação. A argumentação sobre a qualidade dos trabalhos, sobre os critérios, ou a falta deles, adotados pelas referências das revistas, sobre o caráter regional dessas publicações são ciclicamente utilizados como barreiras para incorporação desses saberes nas memórias eletrônicas. Sem pensar no reducionismo e na ideologia das linguagens de indexação, de uma forma

geral, tudo que é escrito em outro idioma que não o inglês é desfavorecido pelos mecanismos de coleta das bases de dados comerciais. Contudo, muitos estudos põem em evidência que uma proporção significativa da pesquisa científica produzida dentro das fronteiras dos países em desenvolvimento, em domínios específicos, mas pertinentes a problemas universais da ciência moderna, são de grande importância tanto para o conjunto desses países, quanto para a ciência em geral. Temas como energia solar, doenças tropicais, agricultura, pecuária são exemplos relevantes que despertam curiosidade científica em qualquer parte do mundo.

Voltando às linguagens documentárias, elas são, em síntese, metalinguagens derivadas da linguagem natural, com semântica e sintaxe própria. Dessa estrutura de representação simbólica, depende, como já foi enfatizado, a formação da memória eletrônica e também das suas possíveis partições. Quero dizer com isto que um único trabalho científico pode ser incorporado em várias memórias, ou seja, ele pode pertencer a várias bases de dados. Os registros em bases de dados distintas das várias leituras possíveis de um documento são viabilizados por códigos de representação específicos e diferenciados, que interpretam este documento mediante regras internas de um sistema de informação qualquer. Esses códigos possibilitam também que esses trabalhos possam ser "lembrados" por diferentes grupos que os valorizem segundo uma ótica própria. É como se um mesmo fato estivesse em memórias coletivas de vários grupos e fosse lembrado de forma distinta em cada um deles.

Suponhamos que um pesquisador escreva um trabalho sobre compreensão de linguagem natural por robôs. Supondo também que esse trabalho ultrapasse todos os filtros de seleção e garanta a sua homologação pela comunidade, ele poderá ser incorporado em uma base de Inteligência Artificial, onde um vocabulário específico fará a representação do seu conteúdo informacional, enfatizando a sua ligação com os problemas dessa área; ao passo que, por um outro código, os lingüistas registrarão nas suas bases de dados os fatos sobre lingüística computacional presentes no trabalho e, exagerando um pouco, os pesquisadores da área de robótica fariam uma terceira interpretação.

Como vimos, um trabalho individual, que faz parte do *curriculum vitae*, da biografia de uma criatura, é submetido a várias interpretações e absorvido por memórias de vários grupos. Esta partição é viabilizada pela linguagem. Obviamente, os mecanismos de lembrança, isto é, de recuperação e citação, vão estar submetidos às mesmas regras dessa linguagem.

Mas, se este trabalho não está em nenhuma memória, não pertence a nenhum grupo, não tem existência reconhecida pela comunidade, ele na verdade não existe. É só um segredo indecifrável na gaveta e no coração de seu autor.

Halbwachs (1990) diz que a memória individual é um ponto de vista da memória coletiva. Dentro desta mesma perspectiva, talvez não fosse um exagero dizer que um trabalho científico cumpre o mesmo pressuposto. Ainda sobre o exemplo anterior, poderíamos dizer que um lingüista pode recordar esse trabalho apenas pela gramática especial que ele propôs; um pesquisador de inteligência artificial estará atento às formas de representação de conhecimento necessárias à representação das regras da gramática proposta; alguém da área de ergonomia tem sua atenção desvia- da pelos fatores que proporcionam melhor interação homem-máquina; os impactos psicossociais serão recordados por um pesquisador em psicologia.

Na verdade, porém, incorporar-se a uma base de dados - na nossa memória eletrônica - é a última etapa de todas por que deve passar uma contribuição à ciência. O ritual se inicia no momento em que o pesquisador determina o escopo de seu trabalho, cujos graus de liberdade estão determinados por constrangimentos sociais, políticos, econômicos e, como veremos adiante, por aspectos mercadológicos. Ao elaborar um projeto de pesquisa, o pesquisador deve estar sensível ao fato de que algumas áreas são oficialmente apoiadas pelos órgãos de fomento à pesquisa e que essas áreas já foram previamente definidas e somente os projetos que estiverem enquadrados nestes planos receberão aval institucional, recursos e financiamentos dos órgãos de apoio à pesquisa. É o caso do Plano Básico de Desenvolvimento Científico e Tecnológico (PADCT), que define no país o que os pesquisadores têm de pesquisar. Estes planos cumprem um papel importante na estruturação, ordenação e homologação da ciência oficial que estarão refletidas nas diversas memórias eletrônicas.

Os autores que logram publicar seus trabalhos em revistas ou em anais de eventos considerados importantes para sua comunidade são aqueles com possibilidade de ter os seus trabalhos incorporados nas bases de dados. Dentro dessa perspectiva, o conhecimento selecionado, representado e registrado nas grandes bases de dados internacionais constitui a documentação sobre a atividade científica oficialmente aceita pela comunidade que a gerou. Essas contribuições receberam o endosso, a homologação dos pares, e receberam, portanto, o direito de pertencer à memória oficial da ciência. Dessa maneira, as bases de dados se constituem na forma mais fiel dos testemunhos dos cientistas. É a esta memória eletrônica que os pesquisadores se dirigem em busca dos referenciais teóricos para as suas atividades.

Um outro fator que tem um impacto determinante na construção da memória eletrônica é que ela constitui um grande negócio. Um negócio que movimenta anualmente milhões de dólares. A informação deixou de ser um bem puramente cultural e transformou-se em bem econômico. Dessa transformação, que modificou totalmente a percepção do valor da informação, apropriou-se o capitalismo,

engendrando o que chamamos "indústria da informação", que tem, em outro plano, uma trajetória bastante semelhante à da indústria da cultura no que diz respeito à sua incorporação à estrutura capitalista ou pós-capitalista, como querem alguns, e à sua definitiva transformação em bem comercializável.

A construção desta memória, como já foi dito várias vezes, depende fortemente do poder de representação das linguagens documentárias. De acordo com as circunstâncias, elas podem enfatizar fatos e descobertas, como há pouco tempo aconteceu com as descobertas de fatos novos sobre a supercondutividade, ou, mais recentemente, sobre a fecundação de mulheres idosas. Mas também, promover o silêncio, o esquecimento, como vimos no exemplo do óleo de dendê. No entanto, as "zonas de sombras e silêncios", na expressão de Michael Pollak(1989), podem anteceder a linguagem e se instalar como desdobramento de outros problemas, principalmente os ideológicos. Quantos cientistas no mundo inteiro, por suas posições antagônicas aos regimes políticos de suas pátrias, vítimas do totalitarismo, de ditaduras e de intransigências, não tiveram seus trabalhos de pesquisa impedidos de serem divulgados e banidos da memória de sua época?

As bases de dados são, pois, a metáfora da memória da ciência que se pratica hoje. Elas reúnem os testemunhos de pesquisadores com uma linguagem própria, que parece ser mais um instrumento na eterna busca da pedra filosofal da ciência, que é a busca da ordem, do enquadramento, da classificação em um mundo cada vez mais desordenado e mais entrópico.

Referências

HALBWACHS, Maurice. **A memória coletiva**. São Paulo: Vértice, 1990. 189 p.

BELKIN, Niccholas J. Anomalous state of knowledges as a basis for information retrieval. **Canadian Journal of Information** Science, v.5, p.133-140, 1980.

WITTGENSTEIN, L. **Tractatus logico- philosophicus**, 1961.

GOMES, Hagar E**. Manual de elaboração de tesauros monolíngües**. Brasília: PNBU, 1990. 78 p.

GAILLARD, Jacques. La science du tiers monde est-elle visible? **La Recherche**, n.210, p. 636-640, 1989.

POLLAK, M. Memória, esquecimento, silêncio. **Estudos Históricos,** v.2, n.3, p. 3-15, 1989.

Dados abertos de pesquisa: ampliando o conceito de acesso livre

Luís Fernando Sayão e Luana Farias Sales

1 Introdução

A Declaração de Berlin sobre o Acesso Aberto ao Conhecimento em Ciências e Humanidades, publicada em 2003, amplia as fronteiras do movimento de livre acesso ao explicitar o que se compreende por contribuições de acesso livre. O documento declara que essas contribuições incluem "resultados de pesquisas científicas originais, dados brutos [dados não processados] e metadados, fontes originais, representações digitais de materiais pictóricos e gráficos além de material acadêmico multimídia".

A expansão do conceito de acesso livre – um pilar de importância crítica para a prática de uma ciência mais aberta – não está circunscrita somente às publicações acadêmicas tradicionais, como são os artigos de periódicos; suas demandas avançam para outros conteúdos que incluem, de forma privilegiada, a disponibilização aberta e de forma inteligível de dados de pesquisa. Os pressupostos de uma pesquisa aberta incluem também na sua agenda de preocupações itens como nível de abertura das ferramentas, instrumentos, dispositivos laboratoriais, *software* e formatos usados em experimentos científicos, como fatores inibidores da interoperabilidade e do compartilhamento.

De uma maneira direta, essas demandas renovadas da comunidade científica estão localizadas no escopo da chamada "ciência aberta" cuja preocupação primordial é tornar a atividade de pesquisa mais transparente, mais colaborativa e mais eficiente. A ideia de ciência aberta tem muitas faces e muitos significados, porém, o mais eloquente deles é o que reconhece, primordialmente, que o conhecimento científico é um patrimônio da humanidade e, que, portanto, deve estar disponível livremente para que as pessoas – cientistas ou não – possam usá-lo, reusá-lo e distribuí-lo sem constrangimentos tecnológicos, econômicos, sociais ou legais.

O crescimento contínuo da quantidade de dados produzidos pelos diversos segmentos da sociedade – agências governamentais, instituições de pesquisa, in-

dústria – confere a esses recursos a condição de componente fundamental para a pesquisa científica moderna e os identifica também como parte dos fenômenos relacionados ao chamado *big data*, tão comentado ultimamente. As expectativas em torno de um mundo rico em dados são imensas e incluem desde descobertas de novas drogas, passando por um entendimento melhor sobre as mudanças climáticas e sobre a origem do universo, até metodologias mais apuradas para examinar a história e a cultura.

No contexto da pesquisa científica atual, há uma compreensão de que uma nova ordem se sobrepõe ao que se convencionou considerar como output dos processos de investigação científica. "Os editores [científicos] reconhecem que em muitas disciplinas os dados, em várias formas, são agora o principal produto de pesquisa" conforme enfatiza Murray-Rust (2003). Uma sequência genômica, a velocidade de partículas subatômicas, as respostas de levantamento social, a frequência de substantivos num *corpus* de textos, as imagens de satélites de outros planetas, todos esses recursos informacionais são como dados de pesquisa. Todos esses dados, praticamente, podem ser descritos e armazenados em bases de dados e usados para propósitos de pesquisa (2007). Todos eles são resultados que precisam ser considerados como parte da infraestrutura mundial de informação científica.

Quando consideramos os dados de pesquisa, as condições e delineamentos preconizados pelos movimentos em prol da prática de uma ciência aberta se consolidam como consensual entre cientistas, organizações de fomento, editores científicos e outros atores envolvidos no mundo científico. Várias ações e movimentos cultivados no próprio seio das comunidades científicas partem do pressuposto de que esses estoques de informação configuram um recurso imprescindível para o avanço da ciência. O acesso aos dados de pesquisa torna-se, portanto, um imperativo para a ciência com reflexos globais, considerando que os pesquisadores trabalham em cooperação internacional e os dados são criados, compartilhados e acessados em escala planetária; mas que têm, entretanto, um rebatimento nos planos locais e nacionais, visto que esses mesmos pesquisadores estão, tipicamente, inseridos em estruturas organizacionais locais e submetidos a políticas de fomento a pesquisa de âmbito nacional (BRASE; FARQUHAR, 2011).

Como um fenômeno do nosso tempo, entende-se que há um reordenamento nos processos científicos trazido pela gestão e compartilhamento de dados de pesquisa. A prática de boa gestão desses recursos abre a possibilidade de verificação confiável dos resultados dos experimentos e permite pesquisas transversais e inovadoras desenvolvidas sobre informações já existentes. Dessa forma, encurta o ciclo clássico de comunicação científica e abre novas formas de interlocução e

de socialização no mundo científico, além de contribuir para a racionalização dos recursos financeiros públicos aplicados na pesquisa científica.

O presente estudo pretende analisar os diversos aspectos mais intensamente relacionados à gestão, preservação, compartilhamento e acesso aos dados de pesquisa no ambiente de pesquisa identificado como *e-Science* ou quarto paradigma, tendo como pano de fundo o papel desses recursos informacionais para a ciência aberta.

Nesta direção, o artigo se organiza da seguinte maneira: primeiramente, é feita uma contextualização do tema que traz à discussão o que vem a ser uma ciência conduzida por dados e qual a importância dos dados de pesquisa para o que vem sendo chamado atualmente de "ciência aberta"; em seguida, é colocado em pauta o conceito de dados de pesquisa – objeto principal deste artigo - , bem como a classificação de seus tipos, os processos, as técnicas e ferramentas que o envolve; em um terceiro momento, o artigo traz à tona os impactos deste novo ambiente orientado por dados na comunicação científica e as infraestruturas existentes para tratamentos desses dados, chegando finalmente à proposta dos elementos que devem ser considerados para a composição de um modelo de curadoria digital de dados de pesquisa para o país.

2 Uma ciência conduzida por dados

O reconhecimento do potencial informacional dos dados digitais, distribuídos em rede de computadores, para a ciência contemporânea transforma a visão que caracterizava dados de pesquisa, registrados em mídia impressa ou mesmo em formatos digitais, como simples subprodutos dos processos de pesquisa. Nesse contexto, os dados eram considerados somente na sua configuração final, sem considerar os seus ciclos de vida, versões e linhagens e, via de regra, eram descartados ou armazenados em mídias ou servidores sem a devida gestão quando os projetos eram concluídos. Quase sempre eram tragados silenciosamente pelo tempo: pela obsolescência tecnológica, pela efemeridade dos formatos e pela fragilidade das mídias digitais.

As tecnologias digitais aliadas aos tentáculos planetários das redes de computadores têm transformado de maneira vertiginosa a forma como os dados de pesquisa podem ser produzidos, disseminados, gerenciados, compartilhados e usados, tanto na ciência como em outros empreendimentos da sociedade, como nas esferas governamentais e nos negócios. Uma nova geração de sensores, instrumentos científicos avançados, *software* de simulação, colaboratórios, escalas mais precisas produzem em ritmo exponencial quantidades imensas e diversificadas de dados de pesquisa brutos ou não processados.

A relevância dos dados no contexto da "*Big Science*" como o da astronomia, da física e da biologia, somada aos mecanismos de colaboração em escala global, induziram não só ao surgimento de novos modelos de ciência – coletivamente chamados de "quarto paradigma científico" ou "*e-Science*" – mas possibilitaram a emergência de novos campos de estudo como a astroinformática e a bioinformática (BORGMAN, 2010). Existem, hoje, disciplinas científicas que são totalmente – em todos os seus ciclos – orientadas por dados. Nessa direção, pesquisadores em áreas específicas e cientistas da computação trabalham colaborativamente em muitos campos, definindo novos domínios de conhecimento e redesenhando os contornos disciplinares da ciência.

A tecnologia digital, como ferramenta fundamental da *e-Science*, interfere intensamente na forma como os dados se inserem nesses novos processos de geração de conhecimento: muitos tipos de dados científicos devem ser vistos, hoje, como componentes fundamentais da infraestrutura de sistemas modernos de pesquisa, cujo valor é expandido pelo acesso aberto e pela ampliação – via processos de curadoria digital - do seu potencial de reuso. Dessa forma, as coletas de dados de pesquisa podem ter um longo ciclo de vida e se integrar aos sistemas tradicionais de informação para a pesquisa na forma de bases de dados armazenados em repositórios de dados e de vinculações às publicações acadêmicas tradicionais, como os artigos de periódicos e teses.

Esse fenômeno contemporâneo cria oportunidades sem precedentes para acelerar a pesquisa científica e gerar riquezas com base na exploração desse acúmulo de dados; abre a possibilidade de que a imensidão de dados gerados pela pesquisa científica contemporânea possa ser coletada, comparada e analisada, engendrando novos conhecimentos e novas questões de pesquisas. Ferramentas avançadas de *software* e de mineração de dados ajudam a interpretar e transformar os dados brutos em configurações ilimitadas de informação e conhecimento. Perguntas instigantes e recursivas colocadas perante os vários segmentos científicos podem agora ser endereçadas, pela combinação de múltiplas fontes de dados provenientes de domínios diferentes, através da aplicação de modelos complexos e de métodos inéditos de análise. A capacidade dos cientistas de compartilhar e combinar importantes conjuntos de dados é o fundamento a partir do qual novos enfoques para resolução de problemas podem ser desenvolvidos (BERMAN; WILKINSON; WOOD, 2014). O compartilhamento e o intercâmbio permitem descobrir conexões no que estava antes desconectado, concluem *Berman* e seus colaboradores (2014). Dessa forma, como ressaltam Uhlir e Schröder (2007), "a produção de conjuntos de dados constitui o primeiro estágio para aprimorar o conhecimento de partes da natureza e da sociedade, engendrando novas pesquisas e inovação".

Entretanto, uma vez que diferentes áreas científicas possuem padrões, práticas e políticas distintas em relação aos seus dados de pesquisa, torna-se essencial para o efetivo uso e reuso desses recursos o estabelecimento de infraestruturas técnicas, gerenciais e sociais que facilitem a integração dos conjuntos de dados de diferentes domínios e a criação de canais de colaboração entre as diferentes comunidades. Em muitas áreas, como a da física de partículas, ciberinfraestruturas baseadas na computação em grade – em que as tarefas estão divididas em várias máquinas - são implementadas para dar suporte tecnológico à colaboração global.

Os pesquisadores, as instituições acadêmicas e as agências de fomento à pesquisa começam a compreender que esses dados, se devidamente tratados, preservados e gerenciados, podem constituir uma fonte inestimável de recursos informacionais. Os repositórios de dados se incorporam rapidamente à infraestrutura mundial de informação científica e, dessa forma, os acervos de dados podem ser usados, reusados e compartilhados. Potencialmente, esses dados podem capacitar os pesquisadores a formular novos tipos de indagações, hipóteses e a usar métodos analíticos inovadores no estudo de questões críticas para a ciência e para a sociedade (MAYERNIK, 2012).

3 A importância dos dados de pesquisa para uma ciência aberta

Assim como se debate hoje, fortemente, a questão do acesso livre aos periódicos acadêmicos, criando-se novos modelos de disseminação de resultados de pesquisas - mais ágeis e mais dinâmicos e organicamente mais próximos das comunidades científicas -, fica claro que é preciso estender o movimento de livre acesso também aos dados científicos, posto que esses recursos constituem uma parte imprescindível do estoque de conhecimento acumulado pelo trabalho acadêmico e de pesquisa, e que são financiados, na maioria das vezes, pelo dinheiro público. As facilidades propostas pelas organizações que lidam com dados de pesquisas para encontrar, identificar, arquivar, adicionar valor e reusar esses dados criam um novo canal de diálogo entre os acadêmicos e pesquisadores, que se reflete nos modelos de socialização acadêmica e de comunicação científica. Isto porque grande parte da ciência contemporânea é construída com base em dados digitais de pesquisas, num ciclo que inclui a sua coleta, análise, publicação, reanálise, crítica e reuso (MOLLOY, 2011). Os dados e conjuntos de dados de pesquisas providenciam as evidências necessárias para conferir veracidade, autenticidade e capacidade de reprodutibilidade ao corpo de conhecimento publicado nos periódicos, o que parece ser fundamental para o progresso científico. Portanto, quanto maior a capacidade dos sistemas de informação de oferecer dados de pesquisas livremente e que sejam tratados por metadados, de forma que possam ser interpretados e reutilizados pelo maior nú-

mero possível de pesquisadores de diversas áreas, maior será o grau de transparência, de reprodutibilidade e de eficiência do processo de geração de conhecimento científico, e maior será a amplitude de aplicação dos projetos de pesquisa para a sociedade.

É perfeitamente compreensível que o acesso aberto inteligível à coleta de dados de pesquisas seja uma etapa crucial para os pressupostos de ciência aberta. Porém, as suas atribuições vão mais além do que somente a reprodutibilidade e a verificação do que está registrado na literatura acadêmica. O potencial cognitivo dos dados redesenha, através do reuso, os fluxos tradicionais de comunicação científica, estabelecendo novos padrões de socialização e de trabalho cooperativo independente de barreiras geográficas e disciplinares. O valor do dado de pesquisa está diretamente relacionado à possibilidade de uso e ao seu potencial de ser reinterpretado em outras áreas e contextos diferentes da que originalmente o gerou.

Uhlir e Schoröeder (2007) vão mais adiante na análise dos benefícios científicos e socioeconômicos de uma ciência mais aberta, tendo como ponto central o papel dos dados de pesquisas financiadas por recursos públicos. Esses autores alinham algumas das muitas razões para o desenvolvimento de regimes de acesso mais abrangentes nas esferas institucionais, nacionais e internacionais, tendo o acesso livre como uma regra predominante.

- Reforça a pesquisa científica aberta;
- Incentiva a diversidade de análise e de opiniões;
- Promove novos tipos de pesquisa;
- Possibilita a aplicação de ferramentas automatizadas online de descoberta de conhecimento;
- Permite a verificação de resultados prévios;
- Torna possível o teste de hipóteses e de métodos novos ou alternativos de análise;
- Dá suporte a estudos sobre métodos de coleta de dados e de mensuração;
- Facilita a formação de novos pesquisadores;
- Possibilita a exploração, por outros pesquisadores, de tópicos não previstos pelos pesquisadores iniciais;
- Permite a criação de novos conjuntos de dados, de informações e de conhecimentos quando os dados de múltiplas fontes são combinados;
- Ajuda a transferir informação factual para países em desenvolvimento, promovendo a capacitação de pesquisadores nesses países;
- Promove a pesquisa interdisciplinar, intersetorial, interinstitucional e internacional;

Geralmente, ajuda a maximizar o potencial de pesquisa das novas tecnologias digitais e das redes de computadores, proporcionando um retorno maior para os investimentos públicos em pesquisa.

Nessa perspectiva, faz-se necessário compreender melhor o que é dado de pesquisa. A seção seguinte vai nessa direção.

4 Afinal, o que é dado de pesquisa?

O Relatório da OECD – sigla em inglês para Organização para a Cooperação e Desenvolvimento Econômico - descreve a expressão "dados de pesquisas" como "registros factuais usados como fonte primária para a pesquisa científica e que são comumente aceitos pelos pesquisadores como necessários para validar os resultados do trabalho científico" (OECD, 2007). Entretanto, o que se observa é que a amplitude do que se entende por dados de pesquisa sugere um conceito complexo que pode se manifestar numa multiplicidade de formas.

A noção de dados pode variar consideravelmente entre pesquisadores e, ainda mais, entre áreas do conhecimento. A constatação de que os dados são gerados para diferentes propósitos, por diferentes comunidades acadêmicas e científicas e por meio de diferentes processos intensifica ainda mais essa percepção de diversidade. Tipos de dados podem incluir, por exemplo, números, imagens, textos, vídeos, áudio, software, algoritmos, equações, animações, modelos, simulações. Alguns tipos de dados têm valor imediato e duradouro, enquanto outros adquirem valor ao longo do tempo; alguns dados são capturados num momento específico e irrecuperável, enquanto outros são passíveis de se reproduzir (BORGMAN, 2010). Essa heterogeneidade intrínseca aos dados de pesquisa implica que é necessário formular políticas de amplo espectro, que não só identifiquem, mas efetivamente sustentem os vários tipos de dados e a sua natureza díspar. O reconhecimento dessa idiossincrasia torna-se crucial quando se estabelecem as opções gerenciais e tecnológicas para o arquivamento persistente e para a curadoria digital.

O *National Science Board da National Science Foundation* (NSF) adota uma lógica de categorização que considera as seguintes características: a natureza dos dados, sua reprodutibilidade, o nível de processamento ao qual eles foram submetidos. Cada uma dessas diferenças tem implicações importantes na formulação das políticas de gestão de dados digitais de pesquisas e na forma como eles devem ser arquivados e preservados.

Seguindo a categorização proposta pelo *National Science Board* (2005), os dados podem ser distinguidos pela sua natureza ou origem em: observacionais, computacionais e experimentais.

- Dados observacionais – são obtidos por meio de observações diretas, que podem ser associadas a lugares e tempo específicos, como por exemplo, a erupção de determinado vulcão numa data específica, a fotografia de uma supernova, o levantamento das atitudes de uma comunidade. Os dados observacionais – por sua natureza instantânea – guardam uma importância crítica que os qualifica como registros históricos que não podem ser coletados uma segunda vez e, portanto, devem ser submetidos a processos de curadoria que os preserve para sempre.
- Dados computacionais – são resultados da execução de modelos computacionais ou de simulações, seja, por exemplo, no domínio da física ou para a criação de ambientes virtuais culturais ou educacionais. Para esta categoria de dados a preservação por longo prazo pode não ser necessária, posto que os dados podem ser replicados ao longo do tempo. Entretanto, replicar o modelo ou a simulação no futuro pode exigir um grande número de informações que incluem descrição das dependências de *hardware*, *software* e outras dependências técnicas, e ainda os dados de entrada. É preciso notar que algumas vezes é mais conveniente preservar somente os dados de saída.
- Dados experimentais – são provenientes de situações controladas em bancadas de laboratórios, como por exemplo, medidas de uma reação química. Em tese, dados experimentais provenientes "de experimentos que podem ser precisamente reproduzidos não necessitam ser armazenados indefinidamente; porém, na prática, nem sempre é possível reproduzir precisamente todas as condições experimentais, particularmente onde algumas variáveis experimentais não podem ser conhecidas e quando os custos de reprodução do experimento são proibitivos" (NATIONAL SCIENCE BOARD, 2005).

É necessário considerar também os registros do governo, de negócios, da vida pública e privada, entre outros, como fontes de dados úteis para a pesquisa científica, seja qual for a natureza do seu objeto: tecnológico, social ou humano (BORGMAN, 2010).

Como se observa nas definições da NSF, o potencial de replicação das pesquisas é um aspecto fundamental a ser considerado na tomada de decisão sobre como gerenciar os dados de pesquisa. A possibilidade ou não de se obter esses dados novamente ou ainda a possibilidade de reaproveitar esses dados para a realização de uma nova pesquisa agregam à curadoria de dados de pesquisa um valor ainda maior.

A partir daí, pode-se pressupor que o reuso e o compartilhamento de dados e informações, num ambiente de pesquisa caracterizado pela pluralidade de vi-

são sobre esses recursos, abrem a possibilidade de se conceituar formas inéditas de agregações abstratas de produtos de pesquisa que sejam portadores de interpretações específicas, criando, dessa forma, novos constructos intelectuais que possuam os atributos mínimos dos recursos informacionais, ou seja, possam ser identificados e tenham sua autoria reconhecida. Esses novos constructos podem constituir formas de expressão que portem novas unidades de pensamento, conceitos, opiniões etc.

Assim, o reuso e a interpretação de dados de pesquisa em diferentes contextos é um desafio importante na área de curadoria digital de dados de pesquisas e para a *e-Science* que tem que lidar com os enigmas colocados pela grande quantidade de dados produzidos pelas disciplinas científicas que se enquadram no quarto paradigma, constituindo-se para ambas as áreas objetos essenciais de pesquisa. A subseção a seguir discutirá brevemente a questão do reuso de dados de pesquisa.

5 Reuso de dados de pesquisas

> *"Os dados que coletamos hoje podem ser usados no futuro de forma que ainda não conseguimos imaginar. Os exploradores de antigamente que coletavam espécimes de plantas e animais não sabiam nada sobre* DNA *e hoje as amostras são submetidas a esse tipo de investigação. Quando você coleta os seus dados, reúne informações que, no futuro, poderão ser analisadas de formas muito diferentes. São coisas que terão um valor enorme para cientistas que ainda nem nasceram*"
> (POLIAKOFF, 2013).

A ciência como um todo avança com maior qualidade, menor custo e mais eficiência quando abre a possibilidade para que o maior número possível de pesquisadores disponha de vias de acesso aos dados acumulados por seus antecessores e contemporâneos. Isso evita, objetivamente, o custo da duplicação de esforços e permite novas interpretações em diferentes contextos científicos para esses dados e, além do mais, permite que eles sejam integrados e retrabalhados de forma mais criativa, descortinando horizontes para novas pesquisas.

Nessa perspectiva, o conceito de "reuso" torna-se de fundamental importância para a ciência aberta, sendo compreendido de maneira ampla como o uso de dados - normalmente sem explícita permissão - para estudos, previstos ou não pelo autor original dos dados, por outros pesquisadores. O reuso inclui processos de

agregação em base de dados, parâmetros em simulação e combinação de dados de diferentes fontes gerando novos insights (MURRAY-RUST, 2008).

Ainda segundo Murray-Rust (2008), a prática de publicar e reusar dados de pesquisa varia enormemente entre diferentes disciplinas. Algumas áreas, como a de biociências, têm uma longa tradição de exigir que os dados sejam publicados e, a partir desse ponto, de agregá-los em bancos de dados financiados publicamente. As disciplinas classificadas como pertencentes à *Big Science* – como a astronomia e a física de partículas - já estabeleceram uma política bem consolidada de reuso de dados de pesquisas que torna mandatório que dados de telescópios, satélites, aceleradores de partículas, fontes de nêutrons, entre outros, sejam universalmente disponíveis para reuso. Para tal, oferecem ciberinfraestruturas sofisticadas para o compartilhamento dos dados gerados por seus aparatos.

Fica claro, então, que o reuso dos dados de pesquisas está sujeito à sua preservação e gestão, que podem ser feitas por meio das técnicas de curadoria digital de dados de pesquisas que serão comentadas a seguir.

6 Curadoria de dados de pesquisas

Disponibilizar as coletas de dados de pesquisas na *Web* é apenas uma das etapas de um ciclo complexo, e que isoladamente não garante que esses recursos possam ser acessados, reusados, e, sobretudo, ter seus significados e estruturas recompostos agora e no futuro. Nos processos de desenvolvimento de coletas de dados, muitos problemas técnicos e gerenciais se interpõem; porém, o mais relevante deles é assegurar que um conjunto de dados de pesquisas possa manter a sua capacidade de transmitir informação e conhecimento ao longo do tempo e do espaço e que, dessa forma, possa ser reusado enquanto persistir o seu valor informacional.

Entretanto, os *bits* - que compõem a maioria dos dados de pesquisa - não falam por si próprios e não impressionam nossos sentidos. Para que eles possam manter a sua capacidade de ser interpretados em domínios distintos, transversalmente, é necessário que eles estejam suficientemente organizados e documentados. Dessa forma, torna-se imprescindível que informações contextuais – semânticas e estruturais – acompanhem os dados digitais de forma que eles estejam autodescritos. Isto é efetivado por meio de modelos conceituais de informação, expressos na prática por esquemas de metadados, que documentam, por exemplo, os elementos semânticos, as partes dos objetos e suas relações, as dependências técnicas, a proveniência, a identificação persistente, as restrições e os direitos associados aos dados, as possíveis intervenções sofridas e seus efeitos. Ou seja, os metadados devem registrar idealmente tudo o que deve ser de interesse do usuário, incluindo modelos de dados, equipamentos especiais, especificação da instrumentação, linhagem dos

dados e muito mais. Essas informações têm um forte impacto na capacidade dos dados de transmitir conhecimentos e poder ser interpretados e reusados.

Os conhecimentos e as práticas acumulados na última década em preservação e acesso a recursos digitais resultaram num conjunto de estratégias, abordagens tecnológicas e atividades que, agora, são coletivamente conhecidas como "curadoria digital". Ainda que seja um conceito em evolução, já está estabelecido que a curadoria digital envolve a gestão atuante e a preservação de recursos digitais durante todo o ciclo de vida de interesse do mundo acadêmico e científico, tendo como perspectiva o desafio temporal de atender a gerações atuais e futuras de usuários.

Portanto, torna-se claro que, subjacentes às metodologias utilizadas pela curadoria digital, estão os processos de arquivamento digital e de preservação digital; porém, ela inclui também as metodologias necessárias para a criação e gestão de dados de qualidade e a capacidade de adicionar valor a esses dados, no sentido de gerar novas fontes de informações e de conhecimentos (LEE; TIBBO, 2007). As tecnologias e os modelos de gestão para a preservação de longo prazo, definidos pelos repositórios digitais confiáveis, cumprem um papel importante no âmbito da curadoria digital de dados de pesquisas.

O *Data Curator Center* (DCC) – cujo lema é "porque a boa pesquisa precisa de bons dados" - informa em uma página da *Web* que a curadoria digital "envolve a manutenção, a preservação e a agregação de valor a dados de pesquisas digitais durante o seu ciclo de vida"; e que a gestão ativa sobre esses dados reduz as ameaças ao seu valor de longo prazo e minimiza os riscos da obsolescência digital. Além de reduzir a duplicação de esforços na criação de dados de pesquisas, a curadoria reforça o valor de longo prazo dos dados existentes quando os tornam disponíveis para a reutilização em novas pesquisas de qualidade.

Daisy Abbott (2008) amplia um pouco mais a ideia de curadoria digital definindo-a como todas as atividades envolvidas na gestão de dados, desde o planejamento da sua criação – quando os sistemas são projetados -, passando pela definição de boas práticas na digitação, na seleção dos formatos e na documentação, de modo a se tornarem disponíveis e adequados para serem descobertos e reusados no futuro. A curadoria digital também inclui a gestão de grandes conjuntos de dados para uso diário, assegurando, por exemplo, que eles possam ser pesquisados e continuem viáveis, ou seja, capazes de serem lidos e interpretados continuamente. Nessa perspectiva, a ideia de curadoria digital estende-se e vai além do controle do repositório que arquiva os recursos e envolve a atenção do criador do conteúdo e dos usuários futuros.

A importância dos dados para a ciência contemporânea está tornando a curadoria digital, o arquivamento persistente, a preservação digital e o estabelecimento

de modelos de informação para a preservação de registros científicos questões-chave para as áreas de pesquisa. Para atender esta necessidade de estabelecimento de novos modelos de informação que aliem preservação de dados científicos digitais à disseminação e acesso aos registros, novos formatos de publicação vêm surgindo, como é o caso da publicação ampliada. Este novo modelo de publicação em que os dados são ligados aos resultados de pesquisas e publicados em um *e-print* tradicional é possibilitado pelas novas tecnologias de informação e comunicação (TICs) e merece atenção especial.

7 Publicações ampliadas: juntando dados e *e-prints*

Não obstante todas as transformações comportamentais e sociais decorrentes do aparato tecnológico que permeia as atividades de pesquisa, a infraestrutura atual de comunicação científica ainda está fortemente centrada no armazenamento e na disseminação de recursos informacionais individuais. Partindo dos modelos de publicação na *Web* e voltando aos sistemas formais de informação acadêmica, como as bibliotecas de pesquisas, verifica-se que eles entregam ao usuário basicamente um artigo ou uma monografia. O que parece cada vez mais claro é que a heterogeneidade e a complexidade dos registros de resultados de pesquisas não podem mais ser expressas por documentos convencionais únicos, impressos ou mesmo digitais.

Recentemente, vários estudos se concentraram na possibilidade de se entrelaçar produtos de e-pesquisa que se encontram distribuídos, gerando novas modalidades de documentos científicos. Nessa direção, a *Open Archive Initiative* (OAI) define normas para descrição e intercâmbio de agregações de recursos da Web em seu projeto denominado *Object Reuse and Exchange* (OAI-ORE). Conforme explicitado numa página desse projeto na *Web*,

> Essas agregações, algumas vezes chamadas de objetos digitais compostos, podem combinar recursos distribuídos com múltiplos tipos de mídia, incluindo texto, imagens, dados e vídeo. O objetivo dessas normas é expor o rico conteúdo nessas agregações para aplicações que mantêm sistemas de autoria, depósito, intercâmbio, visualização, reuso e preservação (OPEN ARCHIVES INITIATIVE, 2014).

As normas equacionam o problema básico que é a ausência de forma padronizada para descrever os elementos constituintes do objeto digital composto e os limites de uma agregação (LAGOZE; SOMPEL, 2008).

O Projeto DRIVER II – sigla para *Digital Repository Infrastructure Vision for European Reearch* - tem como alvo investigar as formas pelas quais a disponibilidade

de dados de pesquisas podem ser usadas para enriquecer as publicações acadêmicas tradicionais. O documento abstrato que combina *e-prints* e dados de pesquisas - chamado de "publicação ampliada" - emerge da compreensão de que as publicações tradicionais são limitadas na sua capacidade de incorporar os resultados de todo o ciclo de geração de conhecimentos da ciência contemporânea. Isso acontece especialmente quando grandes conjuntos de dados são gerados. Nesse momento, fica evidente que os textos acadêmicos só podem apresentar os dados de pesquisas de forma condensada.

A valorização dos dados de pesquisas como recursos relevantes para uma ciência aberta tem reflexo na implantação de infraestruturas gerenciais e tecnológicas para o arquivamento desses dados. É um fato promissor observar que crescentemente os dados de pesquisas estão sendo armazenados em repositórios de dados confiáveis, onde, gerenciados sob os princípios da curadoria digital são preservados e mantêm a sua capacidade de reuso. Entretanto, na atual infraestrutura de comunicação científica, esses conjuntos de dados não estão conectados às publicações onde são discutidos e analisados. A ideia que está por trás das publicações ampliadas é precisamente criar pontes que liguem os conteúdos dos repositórios institucionais, ou seja, que liguem publicações científicas e conteúdos dos repositórios de dados (VERHAAR, 2008).

Dessa forma, a publicação ampliada é pensada como uma forma de objeto digital complexo que combina vários recursos heterogêneos que, porém, são relacionados. A base para esse tipo de objeto ainda é a publicação acadêmica tradicional, por exemplo, uma tese e o conjunto de dados que dá sustentação às suas análises e argumentações, somada também com os metadados necessários para manter a semântica, estrutura e gestão dessa nova publicação. Naturalmente, um artigo de periódico oferece uma "visão" dos significados e interpretação dos dados – e apresentações de congressos e trocas informais podem oferecer outras "visões" – mas o dado em si é cada vez mais um importante recurso para a comunidade científica e essa é a principal justificativa para acoplar ao artigo de periódico os dados da pesquisa que o embasa, conforme enfatiza Verhaar (2008).

Nesta perspectiva, a publicação ampliada é uma ferramenta útil para a abertura e disseminação dos dados de pesquisas de forma integrada, garantindo aos dados sua significação original e a identificação de sua autoria. Além disso, ao unir os dados de uma pesquisa ao seu resultado final publicado em um *e-print*, a publicação ampliada permite a preservação da memória da pesquisa científica realizada, consentindo a replicação da pesquisa para fins de validação ou ainda para agregar valor a uma nova pesquisa. Sendo assim, a publicação ampliada pode ser considerada um veículo de comunicação científica de grande importância para a comunidade de pesquisadores.

É interessante observar que a abertura dos dados, sua curadoria, bem como os novos veículos de comunicação que utilizam a publicação ampliada podem impactar fortemente o processo de comunicação científica, alterando o seu ciclo uma vez que uma nova relação se estabelece entre os pesquisadores quando um pesquisador, para desenvolver seus projetos, passa a depositar toda a confiança nos dados levantados por outro, distante no tempo e no espaço. A próxima seção comentará esses impactos.

8 Comunicação científica num ambiente orientado por dados

De uma forma definitiva, a ciência orientada por dados e pelas tecnologias digitais criam um ponto de inflexão no ciclo tradicional da comunicação científica. Disciplinas como física das partículas, química, astronomia, geologia dependem de forma absoluta do uso intensivo de ambientes de rede altamente distribuídos, instrumentos automatizados, técnicas de captura de imagens e programas de simulação. Esse aparato tecnológico tem impactado ampla e profundamente a forma como os cientistas podem conduzir e disseminar as suas pesquisas, desenhando novos fluxos de cooperação e compartilhamento e definindo conceitos inéditos para a comunicação e para o registro científico.

Tomando como referência os princípios da curadoria digital, são inúmeras as reflexões que se podem fazer face aos impactos do reuso de dados de pesquisas, da publicação e da citação de coletas de dados, e a partir do estabelecimento de novos conceitos de publicações acadêmicas - mais complexas e mais heterogêneas - sobre o ritual de comunicação científica. De uma forma geral, a curadoria de dados científicos adiciona velocidade ao ciclo da comunicação científica na medida em que oferece aos pesquisadores dados prontos para o reuso, ou seja, dados tratados, acompanhados por metadados semânticos e estruturais – que asseguram a fidedignidade de seu significado e a reconstrução correta de sua apresentação, somados a metadados que asseguram a integridade, precisão e autenticidade. Dessa forma, novas pesquisas de qualidade podem ser desenvolvidas, com a segurança necessária, a partir desses dados, que estão instrumentalizados para serem transportados para novos domínios e reusados sob novos propósitos.

No novo ambiente de pesquisa redesenhado pelas práticas da *e-Science*, o ciclo de vida da curadoria digital incorpora-se como uma peça-chave no fluxo tradicional de comunicação científica baseado tradicionalmente em artigos de periódicos. A curadoria digital, no momento em que gerencia e preserva os dados de pesquisa para que sejam acessados e compreendidos por outros pesquisadores estabelecendo um diálogo com o futuro, cria a possibilidade de se criar conceitos inovadores de documentos de registros de pesquisas, rompendo com o paradigma unidimensional e absoluto do artigo de periódico.

O acesso efetivo a dados de pesquisas, de uma forma responsável e eficiente, consubstanciado por tecnologias de informação e comunicação, se torna uma condição crítica para as políticas nacionais de ciência e tecnologia. O Relatório da Organização para Cooperação e Desenvolvimento Econômico - OCDE (2007) enfatiza essa condição, alinhando, entre tantas outras possibilidades, algumas situações em que os dados de pesquisas se tornam um fator imprescindível: na cadeia de inovação, na cooperação internacional, na promoção de novas pesquisas e testes de hipóteses novas ou alternativas, na diversidade de estudos e opiniões; na formação de novos pesquisadores, na exploração de tópicos não idealizados originalmente, na geração de novos conjuntos de dados a partir de dados de múltiplas fontes e, sobretudo, na promoção de uma atividade científica mais aberta e mais transparente, que tenha como princípio produzir conhecimento publicamente acessível.

Nessa direção, infraestruturas para gerenciamento de dados de pesquisa vêm sendo criadas mundialmente com a finalidade de reunir, preservar, dar acesso e auxiliar os pesquisadores na gestão de seus dados de pesquisas. A seção a seguir apresenta um conceito de infraestrutura de tratamento de dados de pesquisas de origem europeia e o padrão adotado pela comunidade para tornar suas informações interoperáveis.

9 CRIS (*Current Research Information Systems*) e CERIF (*Common European Research Information Format*): infraestruturas de tratamento de dados de pesquisa

A crescente complexidade das atividades de pesquisa, a imensa geração de dados e informações e a necessidade de gerenciar processos propiciou o surgimento de infraestruturas tecnológicas com vistas ao tratamento e à recuperação dessas informações. Essas infraestruturas vêm sendo criadas não apenas para o armazenamento de dados, mas principalmente para gerenciar os processos e as etapas das atividades de pesquisa. Os benefícios são vistos não apenas pelos pesquisadores, mas pelos gestores, pelas agências de fomento, pelas empresas, bem como pelo público em geral. Essas infraestruturas permitem a contextualização das atividades científicas, otimizam os fluxos de trabalho, tornando a produção mais transparente, além de padronizá-las e permitir sua avaliação e reavaliação para o bom andamento das pesquisas, bem como para o reuso de dados e para a viabilização de novas descobertas.

Um exemplo de infraestrutura nesses moldes é o *Current Research Information System* (CRIS), que consiste em um modelo de dados que descreve um conjunto de objetos de interesse para as atividades de pesquisa e uma série de ferramentas que possibilitam ao usuário (pesquisador, gestor etc.) a gestão de seus dados de pesquisa em todos os processos, incluindo alocação de recursos, avaliação de projetos,

identificação de novos mercados para produtos de pesquisa, análise de tendências entre outros serviços.

Em geral, o CRIS é construído para uma dada comunidade, como por exemplo, o *United States Data Agriculture* (USDACRIS), que fornece documentação e relatórios para as atividades agrícolas, ciência dos alimentos, nutrição humana, e silvicultura.

No entanto, a ideia do CRIS não é nova. Há aproximadamente 40 anos, diversos sistemas nos moldes do padrão CRIS vêm sendo desenvolvidos pelo mundo, muitas vezes com outros nomes, mas sempre como mecanismo de apoio à organização e recuperação de informações relevantes para a comunidade científica.

Normalmente, o CRIS tem informações sobre os projetos, pessoas, unidades organizacionais, programas de financiamento, resultados de pesquisas (produtos, patentes e publicações), instalações e equipamentos, e eventos, ou seja, todo tipo de informação que, de alguma forma, pode dar apoio às atividades de Pesquisa & Desenvolvimento (P&D) seja para um financiador, para uma instituição de pesquisa, para o pesquisador, para o público ou para os meios de comunicação.

São exemplos de informações constantes nos CRIS o currículo dos pesquisadores e suas páginas, portfólios de projetos de pesquisa, bibliografias, instituições com pesquisas correlatas, informações sobre oportunidades de inovação, informações sobre instalações e equipamentos, eventos etc.

O sucesso do *CRIS*, a riqueza informacional da *Web* e a proliferação de uma grande variedade de sistemas voltados para as comunidades científicas tornaram as informações para a pesquisa heterogêneas e distribuídas. Como consequência, a busca por esse tipo de informação transformou-se numa tarefa árdua para os usuários. Dito de outra maneira, a informação agora armazenada e tratada estava distribuída em sistemas diversos fazendo com que o usuário gastasse muito tempo navegando separadamente em cada um deles.

Lopatenko (2001) mostra esse problema no seu artigo sobre recuperação de informações no CRIS. Segundo ele, normalmente pesquisadores ou gestores de informações em políticas de pesquisa não se limitam apenas à informação armazenada em um dos sistemas existentes. Ao contrário, informações de pesquisas em qualquer área da ciência e tecnologia estão espalhadas por uma variedade de sistemas de informações heterogêneos e, por isso, há uma forte necessidade de reunir todas as informações possíveis ou, de pelo menos, o sistema indicar onde essas informações podem ser encontradas. *Lopatenko* enfatiza a importância de saber se as informações reunidas na pesquisa são efetivas e completas. No entanto, segundo ele, pesquisas anteriores revelaram que a integração de dados de instituições de pesquisas não resolve o problema, especialmente se as instituições forem regidas por órgãos diferentes ou se não usufruírem de benefícios diretos de participação

em tais redes de informações. Assim, o autor reafirma a necessidade de encontrar uma solução para o problema de integração dos dados, solução esta que será o compartilhamento de um padrão com três características essenciais: 1) fácil de implementar para qualquer participante, 2) flexível o suficiente para abraçar a diversidade, a estrutura e o significado dos dados em diferentes estados, organizações, ou áreas da ciência e 3) poderoso para fornecer serviços sofisticados de recuperação de informações. Para isso, sugere o uso de ontologia e de padrões sugeridos pelo *w3c Consortium*.

Nessa direção, a Comunidade Europeia criou o *European* CRIS (EUROCRIS), uma organização sem fins lucrativos, voltada para o desenvolvimento de sistemas de informações de pesquisas e para a interoperabilidade entre esses sistemas.

A ideia de fazer esses sistemas interoperarem é permitir que o usuário final possa acessar as informações, disponibilizadas no CRIS distribuídos e heterogêneos, bem como em repositórios, em um local único. Para isso, o *EuroCRIS* vem adotando uma série de estratégias, como: troca de experiência entre os membros em geral; criação do DRIS (diretório de CRIS); estudo e desenvolvimento de atividades conjuntas de P&D; conferência bienal sobre CRIS; reuniões semestrais com os membros, seminário estratégico anual, workshops, ligações com parceiros estratégicos, desenvolvimento de estratégia e infraestrutura, e o mais importante deles, o desenvolvimento do *Common European Research Information Format* (CERIF), um padrão recomendado aos estados-membros da comunidade europeia, inicialmente com a finalidade de facilitar o intercâmbio de informações entre bases de dados de projetos de pesquisa.

Criado em 1991, o CERIF, com o passar do tempo, precisou ser revisto e, assim, foi também estendido a outros tipos de informações, além daquelas dos projetos de pesquisa. Nessa direção, a versão CERIF *2000* apresentou diretrizes para um modelo de dados CRIS mais completo e um núcleo base que permitia a troca de informações de maneira flexível, possibilitando que a maioria dos CRIS existentes pudessem manter suas características próprias, e ainda assim interoperar com os demais CRIS existentes na comunidade.

O *CERIF2008* – última versão disponível – descreve um modelo de dados formal – que permite a interoperabilidade entre os sistemas de gestão da investigação, a partir de informações sobre pessoas, projetos, organizações, publicações, patentes, eventos, prêmios, equipamentos etc., um modelo de dados físico (JORG, 2009) e um formato de troca de dados em XML (JORG, 2009).

Além disso, de acordo com Ivanovic, Surla e Rackovic (2011), o modelo de dados CERIF tem uma camada semântica que permite a classificação de entidades e suas relações de acordo com algum esquema de classificação. Outras "entidades"

do modelo de dados CERIF estão ligadas à camada semântica através da "entidade" <cfClass>, que descreve o papel da pessoa na criação do resultado (autor da publicação, editor da publicação, presidente do conselho de eventos, gerente de projetos etc), a classificação do resultado da pessoa (ex: monografia, revista de papel etc), a classificação das publicações em que o resultado é publicado (ex: principal revista de importância internacional, revista nacional etc), a classificação do evento onde o resultado é apresentado (conferência de importância internacional, conferência de importância nacional etc) e a classificação do prêmio que é dado à pessoa (excelente prêmio internacional, prêmio internacional, prêmio nacional etc.).

Complementarmente, de acordo com a página mantida pelo grupo gestor, essa versão incluiu a recomendação de um tesauro multilíngue denominado *Ortelius*, que padronizou a indexação de assunto e os códigos utilizados para as áreas de atividades econômicas e produtos e ainda uma lista controlada de valores e atributos de determinados elementos (por exemplo, o papel de uma pessoa no projeto).

Em suma, a inovação apresentada pelo CERIF está na sua estrutura de dados formais, garantindo a integridade dos dados e evitando múltiplas instâncias dos mesmos valores de atributos; no uso de relações n:n permitindo declarar o papel e a duração temporal dos projetos; na preservação das características individuais de cada sistema e em sua essência multilíngue. Interessante observar que, assim como essa pesquisa, o modelo CERIF está preocupado não apenas em identificar as "entidades" a serem descritas, mas também as relações que elas possuem umas com as outras, o que propicia a formação de uma rede interligada de informações.

No Brasil, as iniciativas semelhantes ao CRIS são raras, e o que se encontrou mais próximo foi a Plataforma Lattes. Entretanto, o sistema CRIS conforme concebido na Europa considera não apenas informações sobre pessoas e instituições, como é o caso do Lattes, mas seu primeiro e principal objeto são os projetos de pesquisa e, mais recentemente, os dados não processados gerados por esses projetos, o que não se encontra em nenhuma das agências brasileiras de financiamento, que seriam as principais interessadas. O que se observa, portanto, é que no Brasil ainda não há um sistema avançado de gerenciamento, acesso e compartilhamento da produção científica nacional, como é o *EuroCRIS*.

Considerando que as atividades de pesquisa atuais geram grande quantidade de dados de pesquisa, e que esses dados devem ser preservados e compartilhados para novos usos e reusos - principalmente porque grande parte dessas atividades é financiada com verba pública e porque é preciso conferir agilidade ao desenvolvimento e à geração de novos resultados – verifica-se que o desafio está no estabelecimento de uma política nacional que possa ser apoiada pelas instituições de pesquisa. Assim, como fruto de investigações já realizadas nesta direção, a seção a

seguir apresenta uma proposta de modelo de curadoria digital de dados de pesquisa para o Brasil.

10 Subsídios para um modelo de curadoria digital

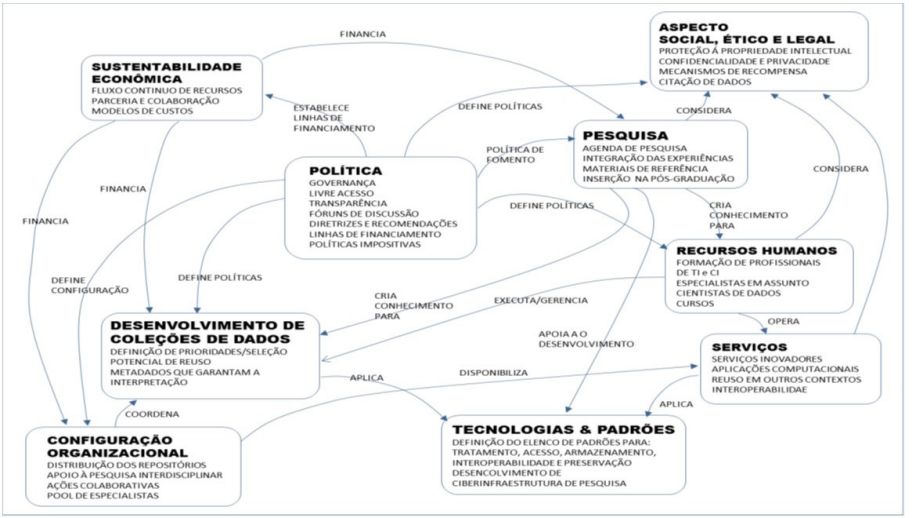
Não obstante as tecnologias de informação e comunicação terem se tornado elementos essenciais para a grande maioria das disciplinas científicas, é necessário considerar, ainda, que o progresso científico não depende unicamente de tecnologias. Políticas voltadas para a pesquisa, fóruns apropriados, legislação específica, fundos para financiamento, valores culturais, ou seja, um espectro multidimensional de fatores afeta profundamente a natureza de novas descobertas, a velocidade com que elas são desenvolvidas e a sua capacidade de se tornarem acessíveis e utilizadas efetivamente (OECD, 2007).

Em relação aos dados de pesquisas, há um consenso nítido entre gestores de C&T, pesquisadores e profissionais das áreas de ciência da informação e de tecnologia da informação de que coletas de dados de pesquisas – principalmente tendo em vista a natureza complexa, heterogênea, distribuída e a fragilidade intrínseca desses artefatos digitais – só podem ser preservados e gerenciados ao longo do tempo, para acesso e reuso, por meio de compromissos sustentáveis e duradouros.

A gestão dinâmica de dados de pesquisas, voltada para uma ciência mais aberta, tem muitas faces e muitos atores, porém, nenhum deles isoladamente é capaz de garantir a capacidade dos dados transmitirem informação e conhecimento aos pesquisadores de hoje e do futuro. Portanto, um modelo de curadoria digital de dados de pesquisa de âmbito nacional deve alinhar as várias dimensões do problema e definir as interlocuções necessárias para a composição de serviços sustentáveis de curadoria digital de amplo alcance e cujas ações se desenrolem em ambientes de e-pesquisa.

Nessa direção, Sayão e Sales (2013) propõem um modelo no qual são consideradas as seguintes instâncias: política, organizacional, desenvolvimento de coleções de dados, pesquisa, infraestrutura tecnológica e de padronização, formação de recursos humanos, sustentabilidade econômica, serviços e implicações sociais, legais e éticas. Essas instâncias são resumidas a seguir, e as relações entre elas são representadas na Figura 1.

Figura 1 - Elementos para composição de um modelo de curadoria digital e suas relações



- Instância política – define políticas, diretrizes, recomendações e estratégias, além de financiamento contínuo, para o desenvolvimento de uma ciberinfraestrutura nacional voltada para o arquivamento, acesso e reuso de dados de pesquisa.
- Instância organizacional – estabelece as configurações organizacionais necessárias para a implantação de repositórios digitais de dados de pesquisa no país.
- Desenvolvimento de coleções de dados – cria os critérios de seleção e as métricas para a avaliação de qualidade, alcance e potencial de reuso dos dados, além dos parâmetros de tratamento técnicos, sobretudo em relação aos metadados, a que os dados devem ser submetidos.
- Instância de pesquisa – preocupa-se com a inserção dos conhecimentos de curadoria digital na agenda de pesquisa de áreas de conhecimento, como a de ciência da informação e ciência da computação, no sentido de se criar um corpo consolidado de conhecimento que possa ser rebatido em todas as áreas que lidam com intensidade com informações e dados digitais.
- Instância de infraestrutura tecnológica e de padronização – estabelece a ciberinfraestrutura necessária para o armazenamento seguro, a recuperação, o acesso a coleções de dados de pesquisas, o planejamento de serviços inovadores; estabelece também as normas e os protocolos que permeiam as ações de preservação e de curadoria digital e os vários níveis de interoperabilidade entre repositórios de dados e informações de pesquisa.

- Instância de formação de recursos humanos – trata da sustentabilidade humana crítica para assegurar continuidade e consistência, ao longo do tempo, de serviços de curadoria de dados de pesquisas, cujas considerações se aplicam a quem financia, produz, gerencia e usa dados de pesquisas.
- Instância de sustentabilidade econômica – define modelos que garantam a sustentabilidade econômica das estruturas de curadoria, posto que a facilitação do acesso, a gestão e a preservação desses dados requerem planejamentos orçamentários específicos e um suporte financeiro apropriado; essa constatação tem origem na própria natureza da curadoria digital, que é um processo que se desenrola indefinidamente no tempo e no espaço; isto implica que o fluxo de fundos para a curadoria deve ser compatibilizado com o ritmo dessa continuidade.
- Instância social, legal e ética – preocupa-se com as barreiras sociais, éticas e legais interpostas entre as comunidades interessadas e o pleno acesso aos dados de pesquisas, tendo em vista o quadro deficiente de proteção ao direito de propriedade intelectual, a dificuldade de documentar os dados para reuso e os problemas associados com a proteção da confidencialidade e privacidade.
- Instância de serviços – delineia o acesso às coletas de dados de pesquisas, na forma de serviços convencionais e inovadores, dirigidos a segmentos variados de usuários; além das facilidades tradicionais – como busca avançada, disseminação seletiva e browsing – os dados devem estar preparados para serem capturados por aplicações computacionais, como data mining, que proporcionem novas análises, estatísticas, indicadores e sirvam também de input para, por exemplo, sistemas de apoio à decisão e sistemas educacionais.

11 À guisa de conclusão

Parece não haver dúvidas de que o chamado dilúvio de dados que caracteriza o ambiente da *e-Science* terá um profundo efeito sobre a atual infraestrutura de pesquisa mundial. Esses efeitos já estão presentes nos novos ambientes de gestão de pesquisa, como o definido pelos padrões e recomendações CRIS e nas superestruturas de cooperação e compartilhamento providas pela computação em grade.

Os próprios sistemas de informação para a pesquisa terão que sofrer mudanças profundas em algumas dos seus fundamentos mais tradicionais, como é, por exemplo, o periódico científico, que entrega aos usuários-pesquisadores no final de projeto de pesquisa, um objeto textual impresso ou digital único que está longe de poder conter a riqueza, diversidade e complexidade dos reais produtos de pesquisa da ciência contemporânea.

Portanto, desafio que se interpõe para os profissionais de informação e de computação é integrar os sistemas e serviços de informação orientados para documentos, como são os catálogos online (OPACs) das bibliotecas de pesquisas e os repositórios institucionais e temáticos de hoje, com os sistemas de informação orientados para dados, como são, por exemplo, os repositórios de dados de pesquisas e os bancos de dados científicos.

Mas é muito importante que esse novo regime de informação definido pela *e-Science*, enquanto uma expressão da ciência aberta, possa se disseminar para todos os segmentos da sociedade. Posto que, na maior parte das vezes, a discussão que se instala sobre as questões críticas que permeiam as utopias de uma ciência aberta transcorrem todas em função da própria ciência e de sua execução transparente para os próprios pesquisadores.

Considerando essa reflexão, um novo desafio se coloca entre a geração e o uso da informação científica e que também tem desdobramentos sobre os sistemas de disseminação de informações: como tornar a informação científica mais aberta, mais transparente e mais próxima de outros segmentos sociais? É preciso estar claro que há uma demanda perceptível por dados científicos decodificados e recondicionados para os "não-pesquisadores": legisladores, formadores de opinião, políticos, professores e o cidadão comum, que precisam conhecer os enigmas científicos do nosso tempo, como as expectativas em torno da célula-tronco, dos alimentos transgênicos, e das mudanças climáticas e de outros problemas científicos que mobilizam a opinião pública, para tomar decisões, emitir sentenças, elaborar leis, transmitir para seus alunos, ou mesmo só para entender o que se faz nos laboratórios com o dinheiro público. Uma ciência aberta pode ser também uma ciência inteligível por todos.

Referências

ABBOTT, D. **What is digital curation**? Edinburgh, UK: Digital Curation Centre, 2008. Disponível em: <http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/what-digital-curation>. Acesso em: 20 mar. 2014.

AULETE, C. **Dicionário escolar da língua portuguesa**. Rio de Janeiro: Lexicon; 2012.

BERLIN DECLARATION ON OPEN ACCESS TO KNOWLEDGE IN THE SCIENCES AND HUMANITIES. Berlin; 2003. Disponível em: <http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf>. Acesso em: 20 dez. 2011.

BERMAN, F.; WILKINSON, R.; WOOD, J. Buiding Global Infrastructure for data sharing and exchange through the Research Data Alliance. **D-Lib Magazine**, n. 20 (1/2). 2014. Disponível em: <http://www.dlib.org/dlib/january14/01guest_

editorial.html>. Acesso em: 04 abr. 2014.
BORGMAN, C. Research data: who will share what, with whom, when, and why? In: **China-North American Library Conference**, .Beijing, 17 aug. 2010. Disponível em: <http://works.bepress.com/borgman/238/>. Acesso em: 21 mar. 2014.
BRASE, J; FARQUHAR, A. Access to research data. **D-Lib Magazine**, n. 17(1/2). 2011. Disponível em: <http://www.dlib.org/dlib/january11/brase/01brase.html>. Acesso em: 30 mar. 2014.
JORG, B.; et al. **CERIF 2008** - 1.0 Full Data Model (FDM). Introduction and Specification. 2009a. 43p. Disponível em: <http://www.eurocris.org/Uploads/Web%20pages/CERIF2008/CERIF2008_1.0_FDM.pdf> Acesso em: 04 abr. 2014.
JORG, B; et al. **CERIF 2008**—1.0 XML. Data Exchange Format Specification. 33p. 2009b. Disponível em: <http://www.eurocris.org/Uploads/Web%20pages/CERIF2008/CERIF2008_1.0_XML.pdf>. Acesso em: 16 fev. 2010.
LAGOZE, C.; SOMPEL, HV. ORE user guide – primer. **Open Archive Initiative**, 2008. Disponível em: <http://www.openarchives.org/ore/1.0/primer.html>. Acesso em: 30 mar. 2014.
LEE, C; TIBBO, H. Digital curation and trusted respositories: steps toward success**. Journal of Digital Information**, n. 8(2). 2007. Disponível em:<http://journals.tdl.org/jodi/index.php/jodi/article/view/229/18>. Acesso em: 20 mar. 2014.
LOPATENKO, AS. **Information retrieval in current research information systems.** arXiv preprint cs/0110026, 2001. Disponível em: <http://arxiv.org/ftp/cs/papers/0110/0110026.pdf>. Acesso em: 30 mar. 2014.
MAYERNIK, M; Et al. The data conservancy instance infrastructure and organization service for research data curation. **D-Lib Magazine**, n.18(9/10), Sep/Oct. 2012. Disponível em: <http://www.dlib.org/dlib/september12/mayernik/09mayernik.html>. Acesso em: 01 fev. 2014.
Molloy J. The Open Knowledge Foundation: open data means better Science. **PLoS Biology**, Dec 2011; n. 9(12). 2011. Disponível em: <http://www.plosbiology.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pbio.1001195&representation=PDF>. Acesso em 01 fev. 2014.
MURRAY-RUST, P. Open data in science. **Serials Review**, 2008; 34(1):p. 52-64.
NATIONAL SCIENCE BOARD. Long-lived digital data collections: enabling research and education in the 21st century. **National Science Foundation**, Sept. 2005. Disponível em: <http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf>. Acesso em: 01 fev. 2014.
OECD PRINCIPLES AND GUIDELINES FOR ACCESS TO RESEARCH

DATA FROM PUBLIC FUNDING. Paris: Organization for Economic Co-operation and Development, 2007. Disponível em: <http://www.oecd.org/sti/sci-tech/38500813.pdf>. Acesso em: 31 mar. 2014.

OPEN ARCHIVES INITIATIVE. **Objective Reuse and Exchenge.** Disponível em: < http://www.openarchives.org/ore >. Acesso em: 21 mar. 2014.

POLIAKOFF, M. [Depoimento]. In: Jones F. Editor-chefe da Nature fala sobre a abertura da ciência. Agência FAPESP, São Paulo, 06 mar. 2013. Disponível em: <http://agencia.fapesp.br/16919>. Acesso em: 01mar. 2014.

SAYÃO, L. F; SALES, L. F. Dados de Pesquisa: contribuição para o estabelecimento de um modelo de curadoria digital para o país. **Tendências da Pesquisa Brasileira em Ciência da Informação**, n.6(1). 2013. Disponível em: <http://inseer.ibict.br/ancib/index.php/tpbci/article/view/102/146>. Acesso em: 04 abril 2014.

UHLIR, P ; SCHRÖDER, P. Open Data for Global Science. **Data Science Journal**, n. 6 (Open Data Issue). 2007. Disponível em: <http://www.spatial.maine.edu/icfs/Uhlir-SchroederPaper.pdf>. Acesso em: 04 abril 2014.

VANOVIĆ, D.; SURLA, D.;RACKOVIĆ, M. A CERIF data model extension for evaluation and quantitative expression of scientific research results. **Scientometrics**, 2011; 86(1): 155-172.

VERHAAR, P. **Report on object models and functionalities.** DRIVER II, 2008. Disponível em: <https://openaccess.leidenuniv.nl/bitstream/handle/1887/16018/Report_on_Object_Models_and_Functionalities.pdf?sequence=2 >.

Curadoria digital: um novo patamar para preservação de dados digitais de pesquisa

Luis Fernando Sayão e Luana Farias Sales

1 Introdução

No período de 1918 a 1919 a gripe espanhola se espalhou pelo mundo inteiro, matando de 20 a 80 milhões de pessoas. De origem viral, não havia tratamento conhecido. Como veio, se extinguiu. Com o intuito de pesquisar meios de evitar uma nova catástrofe, a comunidade internacional das áreas médica e de saúde pública procurou por décadas algum vestígio biológico do vírus causador dessa enfermidade. Só depois de muito tempo, foi encontrada uma amostra de tecido humano infectado pelo vírus num hospital militar da Inglaterra. A partir desses vestígios estão sendo desenvolvidas pesquisas para se descobrir vacinas e meios de tratamento da gripe espanhola. As pesquisas em torno da amostra só se tornaram possíveis graças à preservação dos arquivos científicos, datados de 1916, daquele hospital militar (DITADI, 2003).

Diante do fato de que alguns dados de pesquisa são únicos e não podem ser substituídos se forem destruídos ou perdidos, a questão crucial que se coloca é a seguinte: será que os atuais registros médicos e os demais registros de pesquisa que agora estão sendo documentados de forma digital ou já são gerados em formatos digitais estarão disponíveis para o acesso e para a reutilização em novas pesquisas daqui a alguns anos? Essa questão tem implicações mais amplas, posto que o volume de dados de pesquisa disponibilizados digitalmente está crescendo numa velocidade vertiginosa, engendrando concepções novas de documentos e redesenhando o ciclo tradicional de comunicação científica. É necessário ainda observar que, além de gerar novos dados digitais, os pesquisadores e os acadêmicos, já há algum tempo, começaram a creditar toda a confiança nos conteúdos digitais criados por outros cientistas para dar prosseguimento aos seus empreendimentos (ABBOUT, 2008), inaugurando um novo patamar de compartilhamento de dados e um diálogo transversal ao tempo e ao espaço.

O ato cotidiano das instituições de pesquisa de registrar nos sistemas formais de informação - tais como arquivos, bibliotecas, repositórios, bases de dados - os

resultados de suas pesquisas na forma de documentos, parece não ser suficiente para salvaguardar os dados obtidos ao longo do trabalho de pesquisa. Quando, por exemplo, um estudante de doutorado conclui a sua pesquisa e esta é registrada na forma de um documento que conhecemos por tese, teremos aí somente um retrato parcial dos conteúdos intelectuais gerados no desenrolar de anos de trabalho. Geralmente os dados de pesquisa - que dão sustentação à tese e que serão analisados e discutidos pelo autor - adormecerão armazenados em computadores e mídias pessoais que inexoravelmente serão tragados pela obsolescência tecnológica, pela fragilidade das mídias e, sobretudo, pela falta de intencionalidade de preservá-los adequadamente de forma que sirvam de ponto de partida para novas pesquisas. Isto porque os objetos digitais nunca sobrevivem inercialmente como os seus equivalentes impressos.

O fato determinante é que as atividades pesquisa – como de resto, a maioria dos empreendimentos humanos – estão crescentemente dependentes de materiais digitais. Para que haja avanço do conhecimento científico com um nível mais aceitável de duplicação de esforços, é necessário o estabelecimento de metodologias e compromissos de longo prazo que garantam a capacidade dos dados em formatos digitais, que estão sendo gerados agora, de serem acessados, interpretados e reutilizados com a tecnologia corrente à época do acesso. Portanto, o arquivamento persistente, a preservação digital e o estabelecimento de modelos de informação para a preservação de registros científicos estão se tornando questões-chave para as áreas de pesquisa.

Dados e informações digitais gerados pelas atividades de pesquisa necessitam de cuidados específicos, tornando-se necessário a criação de novos modelos de custódia e de gestão de conteúdos científicos digitais que incluam ações de arquivamento seguro, preservação, formas de acrescentar valor a esses conteúdos e de otimização da sua capacidade de reuso. No intuito de por em prática soluções para o problema, observa-se, no âmbito de várias disciplinas, um esforço em torno do desenvolvimento de repositórios digitais orientados especialmente para uma gestão ativa de dados de pesquisa. É nesse ambiente que surge o conceito de curadoria digital de dados científicos, cujo principal desafio recai na necessidade de se preservar não somente o conjunto de dados, mas de preservar, sobretudo, a capacidade que ele possui de transmitir conhecimento para uso futuro das comunidades interessadas. Isto significa que os ativos genuínos da pesquisa cientifica devem permitir que futuros usuários reanalisem os dados dentro de novos contextos. Porém, para que ocorra um processo de preservação em que os significados dos dados possam atravessar a barreira do tempo, é necessário assegurar que os usuários no futuro estejam instrumentados com as informações essenciais para o efetivo reuso

dos dados (CONWAY, 2011). É de se esperar, portanto, que essas informações estejam estruturadas por modelos de informação e traduzidas por esquemas de metadados.

O objetivo desse estudo é analisar as questões envolvidas na curadoria digital de dados de pesquisa, para tal, discutiremos brevemente a importância dos dados científicos nos padrões atuais de pesquisa; o conceito de curadoria digital e o seu ciclo de vida; a ideia de documentos ampliados que podem vincular publicações acadêmicas, como são as teses, a dados científicos; e para finalizar, uma pequena reflexão a cerca dos impactos da curadoria digital sobre o ciclo tradicional de comunicação científica; à guisa de conclusão ousamos sugerir novas questões para serem investigadas no domínio da ciência da informação.

2 A ciência orientada por dados

A Declaração de Berlin sobre o Acesso Aberto ao Conhecimento em Ciências e Humanidades, publicada em 2003, amplia o escopo do que se entende por acesso livre ao definir que as "contribuições de acesso livre incluem resultados de pesquisas científicas originais, dados não processados e metadados, fontes originais, representações digitais de materiais pictóricos e gráficos e materiais acadêmicos multimídia" .

A expansão do conceito de acesso livre, incorporando agora coleções de dados de pesquisa, vem se consolidando amparada por várias ações cultivadas no próprio seio das comunidades científicas, que reconhecem esses estoques de informação como uma parte do patrimônio da ciência universal e um pilar imprescindível para o seu avanço. O acesso aos dados de pesquisa torna-se, portanto, um imperativo para a ciência com reflexos globais, dado que os pesquisadores trabalham em cooperação internacional e os dados são criados, compartilhados e acessados globalmente; mas que têm um rebatimento nos planos locais e nacionais visto que esses mesmos pesquisadores estão, tipicamente, inseridos em estruturas de financiamento de pesquisa, políticas e organizações acadêmicas de âmbito nacional (BRASE; FARQUHAR, 2011).

"É extremamente raro que novas abordagens fundamentais para pesquisa e educação surjam. A Tecnologia da Informação abriu caminho para essas mudanças cruciais e as coleções digitais estão no cerne dessas transformações". Assim começa o relatório publicado nos Estados Unidos pela *National Science Board* (2005, p.9), cujo titulo "*Long- Lived Digital Data Collections: Enabling Research and Education in the 21st Century*", expressa o reconhecimento da *National Science Foudation* (NSF), para: a importância crescente das coleções digitais para a pesquisa e para a educação; a rápida multiplicação de coleções de dados com o potencial de serem gerenciadas por década através de processos de curadoria; e para necessidade de

investimentos na criação e manutenção de coleções de dados. O Relatório justifica a importância desses itens reafirmando que os dados de pesquisa viabilizam análises em níveis sem precedentes de precisão e sofisticação e oferecem novos conhecimentos por meio da integração inovadora de informações. Pelo grande volume e pela complexidade, as coleções digitais proporcionam novos fenômenos para estudo, ao mesmo tempo em que são forças poderosas para diversos segmentos da educação (NATIONAL SCIENCE BOARD, 2005).

Entretanto, o excesso de dados gerados e armazenados pela pesquisa moderna, numa escala sem precedentes, está muito além da capacidade humana de análise. (CESAR JÚNIOR, 2011). Esse verdadeiro dilúvio de dados vem sendo desencadeado principalmente pelo avanço extraordinário de instrumentos, sensores e escalas, que aumentaram exponencialmente a capacidade de obtenção de dados pela realização de observações e medições de fenômenos, somados às informações geradas artificialmente por simulações e por software.

Esse fato coloca na agenda crítica da ciência um problema novo que é a gestão de dados de pesquisa num mundo digital interligado por redes de computadores, onde há um fluxo intenso de dados, proveniente de diferentes fontes, sendo gerados, processados e compartilhados em ambientes multidisciplinares. A partir desse ponto, instala-se, então, um desafio importante do nosso tempo, que é ao mesmo tempo uma oportunidade extraordinária e absolutamente essencial para se conduzir a pesquisa científica neste século que velozmente se inicia (LANNOM, 2011). Esse desafio traz em si uma questão essencial para os atuais pesquisadores, que é quase o enigma da esfinge: como traduzir em significado e conhecimento a torrente de dados que caracteriza certos domínios da ciência contemporânea?

A resposta a essa questão passa pelo delineamento de um novo paradigma para a ciência, sobre o qual o fazer científico é reordenado pela intensificação do uso de redes e de computadores e pelo uso sem precedentes de conjuntos de dados distribuídos. É o quarto paradigma, ciência orientada por dados ou a *eScience*. Nesse ambiente, novos diálogos se estabelecem entre cientistas da computação e da informação e especialistas de diferentes domínios para o desenvolvimento de novos conceitos e teorias a partir da grande quantidade de dados disponibilizados por diferentes tecnologias (CESAR JÚNIOR, 2011).

Mas quais foram os paradigmas anteriores? No começo da longa aventura da evolução da ciência havia apenas a ciência experimental, em seguida veio a ciência teórica – que pode bem ser ilustrada pelas Leis de Kepler, as Leis de Newton e as equações de Maxwell, entre outras. Depois, como nos relata Gray (HEY, 2009 p. xviii), "os modelos teóricos tornaram-se muito complicados para serem resolvidos analiticamente e as pessoas começaram a simular. Essas simulações nos acompa-

nharam durante a maior parte da última metade do último milênio". No quarto paradigma temos a ciência unificando experimentos, teorias e simulações, através do uso intensivo de dados capturados por instrumentos cada vez mais sofisticados ou gerados por simulação, processados por software e armazenados em computadores na forma de bases de dados. Com a finalidade de se extrair entendimento e inferir significado, a partir desse último ponto os dados podem ser analisados por meio de metodologias de gerenciamento de dados, de soluções estatísticas e também por meio do uso de ferramentas de representação do conhecimento, como ontologias.

É nítida, portanto, a linha que separa os dados dos seus significados. Bell (2011) nos relembra que Keppler (1571-1630) – assistente do astrônomo dinamarquês Tycho Brahe (1546- 1601) – foi quem pegou o caderno de observações astronômicas sistemáticas de Brahe e a partir daí formulou as leis do movimento planetário. Este fato estabeleceu uma divisão clara entre a mineração e a análise de dados experimentais. Por um lado, temos os dados coletados e cuidadosamente arquivados; por outro, a criação de teorias. Esta divisão é um dos aspectos determinante do quarto paradigma.

Nesse contexto de grandes mudanças, novos papeis e responsabilidades emergem como críticos para a gestão de conjuntos de dados de pesquisa, dentre eles está o "cientista de dados" que podem ser cientistas da computação ou cientistas da informação, engenheiros de software e de base de dados, especialistas em disciplinas, entre outros. Apesar de não ser ainda uma carreira de contornos bem definidos e de reconhecimento óbvio, a sua contribuição é fundamental para um diálogo bem sucedido entre todas as partes envolvidas.

Explicitada rapidamente a importância das coleções de dados de pesquisa para o avanço da ciência moderna, concluímos esta seção constatando que a ciência com uso intensivo de dados consiste de três atividades essenciais: captura, curadoria e análise. Tendo em vista esse fluxo, Bell (2011) argumenta que é preciso investir na criação de um conjunto de ferramentas genéricas que cubram todo o espectro de atividades – da captura e validação dos dados à curadoria, analise e, finalmente, arquivamento permanente. Em todo esse ciclo se interpõe o desafio de manter a capacidade de interpretação dos dados e o seu potencial de reuso em vários outros contextos. Prosseguindo no nosso estudo, convidamos o leitor a concentrar a atenção nas soluções e modelos propostos para enfrentar esses desafios.

3 A importância da gestão dos dados de pesquisa

Compreendendo a importância da gestão ativa de coleções de dados para a pesquisa do século XXI a *D-Lib Magazine* (BELL, 2011) – o periódico mais importante no universo das pesquisas em bibliotecas digitais – publicou no inicio do ano

de 2011 um número especial sobre esse assunto. Nessa publicação estão endereçadas questões como acesso livre, curadoria digital, aquisição e gestão, qualidade e confiabilidade e as possíveis conexões entre dados de pesquisa e as publicações acadêmicas tradicionais, que oferecem oportunidades para o surgimento de concepções surpreendentes de documentos mais apropriados ao paradigma da ciência computacional e orientada por dados.

O problema da gestão de dados de pesquisa tem muitas faces que vão se revelando à medida que avançamos. No plano econômico, o custo-benefício de se manter o acesso e a capacidade de reuso aos dados de pesquisa é extremamente difícil de ser mensurado. O valor de um registro pode estar relacionado à possibilidade da reprodutibilidade de um determinado experimento aonde ele foi gerado ou capturado. Algumas pesquisas podem ser fáceis e baratas de se replicar; outras podem ser literalmente impossíveis de se reproduzir – como é a mensuração das características de uma particular erupção vulcânica – ou são repetíveis somente a custos e esforços inaceitáveis (JANSEN, 2006), como uma incursão na atmosfera de Marte. Nessa direção, o arquivamento eletrônico de dados começa a ser estimulado ativamente pelas agências de financiamento de pesquisa, que demandam mais e mais que os projetos científicos contemplem o arquivamento dos dados gerados no decorrer das pesquisas em repositórios de dados confiáveis[3]. O que nos indica que as agências que financiam ou que estabelecem as diretrizes para o setor de pesquisa começam a delinear políticas, estratégias e prioridades que considerem os dados de pesquisa de longa duração como um investimento importante que precisa ser protegido como tal.

O Relatório do Projeto *Digital Repository Infrastructure Vision for European Research II* (Driver II), desenvolvido sob os auspícios da Comunidade Europeia, justifica a preocupação das agências de fomento enfatizando que o acesso a dados de pesquisa proporciona uma série de vantagens, especialmente quando esses dados estão associados a manuscritos acadêmicos disponíveis online. Por exemplo: quando um pesquisador deposita seus dados brutos, ele abre a possibilidade dos seus pares replicá-los e, dessa forma, verificar o que está sendo defendido na publicação científica; isto possibilita também que outros pesquisadores reusem os dados, os comparem e os combinem com outros dados, de forma que novas pesquisas podem ser geradas. Outro benefício apontado pelo Relatório é que a curadoria dos dados torna possível traçar a linhagem dos vários produtos dos projetos de *eScience*, dado que esses projetos se desenvolvem por vários estágios, tais como captura de dados, processamento, modelagem e interpretação. "Se fosse possível destacar as inúmeras conexões entre os recursos que são produzidos durante os vários estágios do processo científico, isto poderia ser de grande utilidade" (VERHAAR, 2008, p.14), enfatiza o autor do Relatório.

Entretanto, para muitas comunidades acadêmicas a gestão e o acesso continuado a esta vasta quantidade de dados ainda é um problema distante de ser superado. Lamentavelmente, muitos dos dados que são produzidos, frequentemente a um custo alto para a sociedade como um todo, são irremediavelmente perdidos.

No curto período do que se convencionou chamar de era digital, algumas instituições científicas se comprometeram no desenvolvimento de atividades que pudessem salvaguardas os dados científicos digitais. Porém as poucas instituições engajadas nesse processo ainda não estabeleceram práticas e não garantiram os fluxos de recursos que assegurem o completo sucesso da gestão desses dados. O que se observa é que ainda persistem lacunas críticas e questões de pesquisas em aberto (LEE; TIBBO, 2007).

Mesmo assim, várias iniciativas importantes, lideradas pelas próprias comunidades científicas, já cumprem papel vital na garantia do acesso livre aos dados de pesquisa e no que se convencionou chamar de curadoria digital, como veremos a seguir.

Ancorado no lema "ajudando você a encontrar, acessar e reusar dados", foi fundada em Londres no ano de 2009 uma organização sem fins lucrativos, chamada de *DataCite* (BRASE; FARQUHAR, 2011), cujos objetivos essenciais, desde então, são: estabelecer bases para o acesso mais fácil a dados de pesquisa na internet; aumentar o grau de aceitação dos dados de pesquisa como contribuições legítimas passíveis de serem citadas nos registros acadêmicos; dar sustentação ao arquivamento de dados de pesquisa de forma que seja possível que os resultados possam ser verificados e readaptados para futuros estudos.

A ideia central que alimenta as ações do *DataCite* é a citação de dados, significando que os dados de pesquisa devem ser citados da mesma forma como são citadas outras fontes de informação, tais como artigos e livros. O *DataCite* preconiza que a citação de dados permite o reuso e a verificação dos dados mais facilmente, possibilitando que o impacto dos dados possam ser rastreados, e que uma estrutura acadêmica que reconheça e recompense os produtores de dados possa ser, finalmente, criada.

Para cumprir seus objetivos o *DataCite* procura juntar as comunidades que lidam com conjunto de dados de pesquisa para que, de forma colaborativa, equacionem o desafio de tornar os dados de pesquisa visíveis e possíveis de serem acessados. Uma das iniciativas importantes nesse processo é o apoio aos centros de dado no assinalamento de identificadores persistentes e na definição de padrões para a publicação de dados; destaca-se também apoio aos editores científicos no sentido de os capacitarem a estabelecer links entre artigos e os dados subjacentes e eles. Para o usuário pesquisador, o *DataCite* oferece recursos e serviços que o ajudam a encontrar, identificar e citar conjunto de dados de forma confiável.

Temos que considerar também o *OpenAIRE* (césar júnior, 2011) – *Open Access Infrastructure for Research in Europe* – que é um projeto de duração de três anos, iniciado em dezembro de 2009, cujo objetivo é apoiar a implementação do acesso aberto na Europa. Para isso vem estabelecendo uma ampla infraestrutura baseada numa rede distribuída de pontos de contato nacionais e regionais nos países europeus, que assegure o apoio localizado aos pesquisadores no seu próprio ambiente. O Projeto está focado em três principais objetivos, dentre eles está a gestão de dados científicos e a sua vinculação com publicações científicas, como explicitado na sua página web: "trabalhar com várias comunidades temáticas para explorar os requisitos, práticas, incentivos, fluxos de trabalho, modelos de dados e tecnologias para depósito, acesso e manipulação de conjunto de dados de pesquisa de várias formas em combinação com publicações de investigação científica."

Outra iniciativa essencial, porém com uma perspectiva voltada para a preservação e reuso de dados de pesquisa, é o *Digital Curation Centre* (dcc)(conway, 2011), que é um ponto de disseminação de práticas e conhecimentos na área de curadoria digital. O lema que está estampado na sua *home page*, resume e justifica a importância das suas atividades: "porque boa pesquisa precisa de bons dados". O modelo de curadoria digital de dados científicos proposta pelo dcc será tratado com um grau a mais de detalhe nas seções seguintes.

4 Afinal, o que é curadoria digital?

Os conhecimentos e as práticas acumulados na última década em preservação e acesso a recursos digitais resultaram num conjunto de estratégias, abordagens tecnológicas e atividades que agora são coletivamente conhecidas como "curadoria digital". Ainda que seja um conceito em evolução, já está estabelecido que a curadoria digital envolve a gestão atuante e a preservação de recursos digitais durante todo o ciclo de vida de interesse do mundo acadêmico e científico, tendo como perspectiva o desafio temporal de atender a gerações atuais e futuras de usuários Torna-se claro, portanto, que subjacente às metodologias utilizadas pela curadoria digital estão os processos de arquivamento digital e de preservação digital; porém, inclui também as metodologias necessárias para a criação e gestão de dados de qualidade e a capacidade de adicionar valor a esses dados no sentido de gerar novas fontes de informação e de conhecimento (lee; tibbo, 2007).

O dcc, na sua visão fundacional, nos informa, na sua página web, que a curadoria digital "envolve a manutenção, a preservação e a agregação de valor a dados de pesquisa durante o seu ciclo de vida"; e que a gestão ativa sobre esses dados reduz as ameaças ao seu valor de longo prazo e minimiza os riscos da obsolescência digital. Além de reduzir a duplicação de esforços na criação de dados de pesquisa,

a curadoria reforça o valor de longo prazo dos dados existentes quando os tornam disponíveis para a reutilização em novas pesquisas de qualidade.

Abbott (2008) amplia um pouco mais a ideia de curadoria digital definindo-a como todas as atividades envolvidas na gestão de dados, desde o planejamento da sua criação – quando os sistemas são projetados -, passando pelas boas práticas na digitação, na seleção dos formatos e na documentação, e na garantia dele estar disponível e adequado para ser descoberto e reusado no futuro. A curadoria digital também inclui a gestão de grandes conjuntos de dados para uso diário, assegurando, por exemplo, que eles possam ser pesquisados e continuem viáveis, ou seja, capazes de serem lidos e interpretados continuamente. Nessa perspectiva, a ideia de curadoria digital estende-se além do controle do repositório que arquiva os recursos e envolve a atenção do criador do conteúdo e dos usuários futuros.

Portanto, verifica-se um deslocamento no padrão de arquivamento estático e inacessível promovido pelos *dark archives*, repositórios de acesso restrito voltados para garantir integridade e autenticidade. O foco da curadoria digital está na gestão por todo o ciclo de vida do material digital, de forma que ela permaneça continuamente acessível e possa ser recuperado por quem dele precise. Ampliando a capacidade dos dados serem recuperados e acessados estão os modelos de informação, expressos por metadados; além do mais, os metadados são também ferramentas importantes para os procedimentos de controle de autenticação (HIGGINS, 2011).

A curadoria digital, em resumo, assegura a sustentabilidade dos dados para o futuro, não deixando, entretanto, de conferir valor imediato a eles para os seus criadores e para os seus usuários. Os recursos estratégicos, metodológicos e as tecnologias envolvidas nas práticas da curadoria digital facilitam o acesso persistente a dados digitais confiáveis por meio da melhoria da qualidade desses dados, do seu contexto de pesquisa e da checagem de autenticidade. Dessa forma, a curadoria contribui para assegurar a esses dados validade como registros arquivísticos, significando que eles podem ser usados no futuro como evidência legal. O uso de padrões comuns entre diferentes conjuntos de dados, proporcionado pela curadoria digital, cria mais oportunidades de buscas transversais e de colaboração. Na ótica financeira, o compartilhamento, o reuso dos dados e as oportunidades de novas análises, além de outros benefícios, valorizam e protegem o investimento inicial na obtenção dos dados.

A curadoria digital emerge como uma nova área de práticas e de pesquisa de espectro amplo que dialoga com várias disciplinas e muitos gêneros de profissionais.

5 Ciclo de vida da curadoria de dados científicos

O DCC oferece um Modelo do Ciclo de Vida da Curadoria, expresso graficamente (DITADI, 2003), que reflete uma visão de alto nível dos estágios necessários

para o sucesso do processo de curadoria e de preservação de dados de pesquisa, que se inicia no estágio de conceitualização ou de recebimento do dado no repositório. O modelo proposto pelo DCC está orientado para o planejamento das atividades de curadoria nas organizações ou consórcios ajudando a garantir que todos os passos do ciclo serão cumpridos. Entretanto, isto não implica que todas as organizações devam cumprir o ciclo desde o primeiro estágio, na realidade, a operacionalização dos estágios dependerá das necessidades reais de cada organização.

Os elementos-chaves do modelo são: dado, objetos digitais e bases de dados. No centro do ciclo de vida da curadoria está o dado, que é qualquer informação codificada em formato binário. A ideia de dado inclui: os objetos digitais simples, que são aqueles compostos por um único arquivo, identificador e metadados, e os objetos digitais complexos, que por sua vez são formados pela combinação de outros objetos digitais formando uma unidade discreta, como é, por exemplo, uma página *web*. Nesse contexto, base de dados é definida como coleções estruturadas de registros ou de dados armazenados em sistemas de computadores.

O outro elemento básico do modelo são as ações que devem ser tomadas no decorrer do processo de curadoria. O modelo do *DCC* classifica as ações em três tipos: ações para todo o ciclo de vida; ações sequenciais e ações ocasionais.

As ações para todo o ciclo de vida são assim chamadas por compreenderem atividades que permeiam todo o ciclo de vida da curadoria digital. Para transmitir essa ideia de presença contínua, essas ações estão representadas graficamente como anéis concêntricos envolvendo os objetos de dados que estão no centro do modelo. As ações são as seguintes:

- **Descrição e a representação da informação** - é efetivada pela atribuição de metadados administrativos, técnicos, estruturais e de representação de acordo com os padrões apropriados; visa assegurar a descrição adequada e o controle de longo prazo; compreende também a coleta e a atribuição de informações de representação necessárias para o entendimento do dado e para a sua apresentação (ou renderização).
- **Planejamento da Preservação** – é necessária a definição de um plano de preservação cujo espectro englobe todo o ciclo de vida da curadoria do material digital, incluindo gestão, administração, politicas, e tecnologias.
- **Participação e monitoramento** – enfatiza a necessidade de atenção para as atividades que se desenrolam no âmbito das comunidades envolvidas com o problema de curadoria, bem como a necessidade de participação no desenvolvimento de padrões, de ferramentas e de software adequados ao problema e que possam também serem compartilhados;

- **Curadoria e preservação** - estar continuamente alerta e empreender as ações administrativas e gerenciais planejadas para a curadoria e preservação por todo o ciclo de vida da curadoria.

As Ações Sequenciais, por sua vez, são etapas que devem ser cumpridas repetidamente para assegurar que o dado permaneça em contínuo processo de curadoria de acordo com as melhores práticas. Essa sequência não é para ser cumprida meramente uma vez do começo ao fim; na realidade ela forma as bases da cadeia de curadoria e continua ciclicamente todo o tempo que o dado estiver sob curadoria.

A sequência de ações do modelo de ciclo de vida da curadoria digital proposto pelo *DCC* tem os seguintes estágios:

- **Conceitualização** – conceber e planejar a criação do dado, incluindo os métodos de captura e as opções de armazenamento; questões tais como propriedade intelectual, embargos e restrições, financiamento, responsabilidades, objetivos específicos da pesquisa, ferramentas de captura e calibração devem ser registradas.
- **Criação e/ou Recebimento** – compreende a criação do dado incluindo o elenco de metadados necessários à sua gestão e compreensão, ou seja, metadados administrativos, descritivos, estruturais e técnicos (os metadados de preservação também podem ser incluídos no momento da criação do dado). Nem sempre os dados são arquivados por quem os gerou, dessa forma, esse estágio inclui também a recepção dos dados segundo políticas bem documentadas, sejam dos seus criadores, de outros arquivos, de repositórios ou centro de dados; quando necessário, assinalar metadados apropriados para a curadoria e a preservação dos dados recepcionados.
- **Avaliação e seleção** – avaliar o dado e selecionar o que será objeto dos processos de curadoria e de preservação por longo prazo; manter-se aderente tanto às boas práticas quanto às políticas pertinentes e também às exigências legais.
- **Arquivamento** – transferir o dado para um arquivo, repositório, centro de dados ou outro custodiante apropriado.
- **Ações de preservação** – promover ações para assegurar a preservação de longo prazo e a retenção do dado de natureza oficial; as ações de preservação devem assegurar que o dado permaneça autêntico, confiável e capaz de ser usado enquanto mantém sua integridade; essas ações de preservação incluem: a limpeza do dado e a sua validação, a adição de metadados de preservação e de informação de representação e a garantia de estruturas de dados ou formatos de arquivos aceitáveis.

- **Armazenamento** – armazenar o dado de forma segura mantendo a aderência aos padrões relevantes.
- **Acesso, uso e reuso** – garantir que o dado possa ser cotidianamente acessado tanto pela sua comunidade-alvo, quanto pelos demais usuários interessados no reuso do dado; isto pode ser realizado por meio de publicação disponível para as comunidades interessadas; porém, controle de acesso e procedimentos de autenticação podem ser aplicados.
- **Transformação** – compreende a criação de novos dados a partir do original, por exemplo, pelo processo de migração para diferentes formatos ou pela criação de subconjuntos - realizada por meio de seleção ou formulação de consultas – derivando novos resultados que podem ser publicados.

Por fim, o Modelo estabelece também os estágios que são aplicados eventualmente, chamados de ações ocasionais. Essas ações interrompem ou reordenam as ações sequenciais como desdobramento de uma decisão. Por exemplo, após uma avaliação pode ser decidido que o dado em questão não se enquadra no escopo de um repositório digital, isso implica que o dado deve ser transferido para outro custodiante. Em outra situação, o dado deve ser destruído, possivelmente por motivações legais.

- **Eliminação** – eliminar os dado que não foram selecionados para curadoria e preservação de longo prazo de acordo com políticas documentadas, diretrizes ou exigências legais.
- **Reavaliação** – retornar ao dado cujos procedimentos de avaliação foram falhos para nova avaliação e possível seleção para curadoria.
- **Migração** – migrar os dados para um formato diferente; isto pode ser feito no sentido de compatibilizá-lo com o ambiente de armazenamento ou para assegurar a imunidade do dado contra a obsolescência de hardware e de software.

O modelo desenhado pelo *DCC* permite uma visão coletiva sobre o conjunto de funções necessárias à curadoria e à preservação de dados de pesquisa. Além de definir papeis, responsabilidades e conceitos, ele explicita a infraestrutura de padronização e as tecnologias que devem ser implementadas.

6 Juntando dados e publicações: documentos ampliados

Não obstante todas as transformações comportamentais e sociais decorrentes do aparato tecnológico que permeia e dinamiza as atividades de pesquisa, a infraestrutura atual de comunicação científica ainda está fortemente centrada no

armazenamento e na disseminação de recursos informacionais individuais. Partindo dos modelos de publicação na *web* e voltando aos sistemas formais de informação acadêmica, como as bibliotecas de pesquisa, verifica-se que eles entregam ao usuário basicamente um artigo ou uma monografia. "Muitos editores acadêmicos não aceitam outro produto de projetos de e-pesquisa, tais como base de dados, gravação de vídeos e serviços web" (VERHAAR, 2008, p.9). O que parece cada vez mais claro é que a heterogeneidade e a complexidade dos registros de resultados de pesquisa não podem mais ser expressas por documentos convencionais únicos, impressos ou mesmo digitais.

Recentemente vários estudos se concentraram na possibilidade de se entrelaçar produtos de e-pesquisa que se encontram distribuídos, gerando novas modalidades de documentos científicos. Por exemplo, Hunter (2007) visualizou um "pacote de publicações científicas" que encapsula e relaciona, na forma de objetos compostos, dados brutos com os seus subprodutos, publicações e metadados contextuais, de proveniência e administrativos. Enquanto Gray (HEY, 2009), no contexto de sua proposta de um método científico transformado, conceitualiza os "documentos sobrepostos" - *overlay documents*, no original em inglês -, que interligam artigos de periódicos revisados por pares, dados, anotações e comentários.

Nessa mesma direção, o *Open Archive Initiative* (OAI) define uma norma para descrição e intercâmbio de agregação de recursos *web* chamada de *Object Reuse and Exchange* (OAI-ORE). "Esta agregação, algumas vezes chamada de objetos digitais compostos, pode combinar recursos distribuídos com tipos múltiplos de mídia, incluindo texto, imagens, dado e vídeo. O objetivo da norma é expor o conteúdo rico dessa agregação para aplicações que suportem sistemas de autoria, depósito, intercâmbio, visualização, reuso e preservação", conforme explicitado na página *web* do *OAI-ORE*. A norma equaciona o problema básico que é a ausência de forma padronizada para descrever os elementos constituintes do objeto digital composto e os limites de uma agregação (LAGOZE; SOMPEL, 2008).

O Projeto *DRIVER II* tem como alvo investigar as formas pela qual a disponibilidade de dados de pesquisa pode ser usada para ampliar as publicações acadêmicas tradicionais. O documento abstrato que combina texto e dados de pesquisa é chamado de *enhanced publication* – termo ainda sem tradução para o português, mas que poderíamos, traduzi-lo por documento ampliado –, emerge da compreensão de que as publicações tradicionais são limitadas na sua capacidade de incorporar resultados de todo o ciclo do processo de investigação científica. Isso acontece especialmente quando grandes conjuntos de dados são gerados. Nesse momento fica evidente que os textos acadêmicos só podem apresentar os dados de pesquisa de forma condensada.

É um fato promissor observar que crescentemente os dados de pesquisa estão sendo armazenados em repositórios de dados confiáveis, onde, gerenciados sob os princípios da curadoria digital são preservados e mantêm a sua capacidade de reuso. Entretanto, na atual infraestrutura de comunicação científica estes conjuntos de dados não são conectados às publicações onde eles são discutidos e analisados. A ideia que está por traz das publicações ampliadas é precisamente criar pontes que liguem os conteúdos dos repositórios institucionais, ou seja, publicações científicas, com os conteúdos dos repositórios de dados (VERHAAR, 2008).

Dessa forma, a publicação ampliada ou o documento ampliado é pensado como uma forma de objeto digital complexo que combina vários recursos heterogêneos, que são, porém, relacionados. A base para esse tipo de objeto ainda é a publicação acadêmica tradicional, por exemplo, uma tese e os seus conjuntos de dados gerados, somada também com os metadados necessários.

7 Os dados científicos e a comunicação científica

De uma forma definitiva a ciência orientada por dados cria um ponto de inflexão no ciclo tradicional da comunicação científica. Disciplinas como física das partículas, química, astronomia, geologia, dependem de forma absoluta do uso intensivo de ambientes de rede altamente distribuídos, instrumentos automatizados, técnicas de captura de imagens e programas de simulação. Esse aparato tecnológico tem impactado ampla e profundamente a forma como os cientistas podem conduzir e disseminar as suas pesquisas (VERHAAR, 2008), desenhando novos fluxos de cooperação e compartilhamento e definindo conceitos inéditos para a comunicação e para o registro científico, que merecem estudos partindo de muitos olhares.

No domínio específico da curadoria digital, são inúmeras as reflexões que se podem fazer face aos impactos do reuso de dados de pesquisa, da publicação e da citação de coleções de dados e a partir do estabelecimento de novos conceitos de publicações acadêmicas - mais complexas e mais heterogêneas - sobre o ritual de comunicação científica. De uma forma geral, a curadoria de dados científicos adiciona velocidade ao ciclo da comunicação científica na medida em que oferece aos pesquisadores dados prontos para o reuso, ou seja, dados tratados, acompanhados por metadados semânticos e estruturais – que asseguram a fidedignidade de seu significado e a reconstrução correta de sua apresentação, somados a metadados que asseguram a integridade, precisão e autenticidade. Dessa forma, novas pesquisas de qualidade podem ser desenvolvidas, com a segurança necessária, a partir desses dados, que estão instrumentalizados para serem transportados para novos domínios. Pode-se observar que uma nova relação se estabelece entre os pesquisadores na medida em que um pesquisador, para desenvolver seus proje-

tos, pode depositar toda a confiança nos dados levantados por outro, distante no tempo e no espaço.

Assim como se debate hoje fortemente a questão do acesso livre aos periódicos acadêmicos, criando-se novos modelos de disseminação de resultado de pesquisa - mais ágeis e mais dinâmicos e organicamente mais próximos das comunidades científicas -, hoje fica claro que é preciso estender o movimento de livre acesso também aos dados científicos, posto que esses recursos constituem uma parte imprescindível do estoque de conhecimento acumulado pelo trabalho acadêmico e de pesquisa, e que são financiados, na maioria das vezes, pelo dinheiro público. As facilidades propostas pelas organizações que lidam com dados de pesquisa para encontrar, identificar, arquivar, adicionar valor e reusar esses dados criam um novo canal de diálogo entre os acadêmicos e pesquisadores, que se reflete nos modelos de socialização acadêmica e de comunicação científica.

No novo ambiente de pesquisa redesenhado pelas práticas da *eScience*, o ciclo de vida da curadoria digital incorpora-se como uma peça-chave no fluxo tradicional de comunicação científica baseado tradicionalmente em artigos de periódicos. A curadoria digital, no momento em que gerencia e preserva os dados de pesquisa para que sejam acessados e compreendidos por outros pesquisadores estabelecendo um diálogo com o futuro, cria a possibilidade de se criar conceitos inovadores de documentos de registros de pesquisa, rompendo com o paradigma unidimensional e absoluto do artigo de periódico.

8 À guisa de conclusão

A tecnologia digital nos coloca diante de um dos dilemas mais críticos do nosso tempo: por um lado ela nos permite criar, manipular, armazenar e tornar disponível uma quantidade impressionante de informações; por outro lado, esta mesma tecnologia fugidia coloca em perigo a longevidade dos objetos informacionais por ela engendrada, colocando a humanidade – que depende cada vez mais dos estoques informacionais digitais – face a face com o perigo de uma amnésia digital. Isto porque os objetos digitais requerem metodologias de gestão que são muito diferentes das que são utilizadas no universo da impressão tradicional.

Uma das atividades humanas em que mais se gera e se manipula materiais digitais é precisamente o trabalho de pesquisa científica. Em alguns nichos específicos, a totalidade das atividades que se desenrolam nos laboratórios distribuídos está centrada num intenso fluxo de dados, nos mais diversos formatos digitais. Era de se esperar, portanto, que surgissem iniciativas que pudessem tornar os dados científicos digitais mais visíveis e sempre possíveis de serem acessados, mantendo a sua integridade, fidedignidade e o seu papel de evidência.

Nessa direção, a curadoria digital emerge como uma nova área de práticas e de pesquisa de espectro amplo que dialoga com várias disciplinas e muitos gêneros de profissionais. Ela une as tecnologias e boas práticas do arquivamento e da preservação digital e dos repositórios digitais confiáveis com a gestão dos dados científicos, criando uma nova área de pesquisa cujos desdobramentos, de amplo espectro, ainda são imprevisíveis. Isto porque, como se trata de uma área que só recentemente despontou como crítica para a pesquisa, ainda restam muitas lacunas práticas e teóricas a serem equacionadas, orientadas, preferencialmente, por uma abordagem multidisciplinar.

A Biblioteconomia e a Arquivologia, que se renovam cotidianamente para enfrentar novos problemas, têm muito a contribuir para a curadoria digital com suas experiências em gestão de patrimônios intangíveis. Representação e organização do conhecimento, os novos conceitos de bibliotecas, repositórios e arquivos digitais, a integridade e autenticidade de materiais digitais e a recuperação da informação, para citar alguns itens, são imprescindíveis para a gestão de coleções de dados de pesquisa; a Museologia digital, por sua vez, pode trazer aportes importantes na questão dos objetos digitais complexos e multimidiáticos, cuja presença é comum na curadoria de exposições museológicas virtuais e pode ser interessante para renderização de estruturas cientificas mais sofisticadas.

Porém, para a Ciência da Informação, os impactos nos obrigam a repensar alguns pontos críticos, como no conceito ancestral de documento, no modelo tradicional de disseminação de resultados de pesquisa e na extensão dos formatos de metadados como instrumentos de recomposição de significados e estruturas.

Esses pontos nos inspiram a propor novos itens para uma agenda de pesquisa dentro do domínio interdisciplinar da Ciência da Informação:

a) em primeiro lugar, seria importante avaliar como o ciclo da comunicação científica se altera mediante as novas formas de colaboração, socialização e disseminação proporcionadas pelo reuso de dados científicos, especialmente em áreas de conhecimento com maiores interfaces com a *eScience*;
b) em segundo, seria interessante investigar as novas modalidade de publicação científica, cuja gênese está na vinculação entre as publicações tradicionais depositadas repositórios em digitais temáticos e institucionais com os dados gerenciados pelos centros de dados e de curadoria digital;
c) por fim, em terceiro, mas não menos importante, está a concepção de modelos de informação que possam orientar a definição de conjunto de metadados capazes de garantir significado, estrutura, fidedignidade e autenticidades aos dados de pesquisa — pelo tempo que for necessário.

Referências

ABBOUT, Daisy. **What is digital curation?** Edinburgh, UK : Digital Curation Centre, 2008. Disponível em: <http://www.era.lib.ed.ac.uk/ bitstream/1842/3362/3/ Abbott%20What%20 is%20digital%20curation_%20_%20Digital%20 Curation%20Centre.doc>. Acesso em: 20 dez. 2011.

BELL, Gordon. Prefácio. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Org.). **O quarto paradigma:** descobertas científicas na era da eScience**.** São Paulo : Oficina do Texto, 2011.

BERLIN DECLARATION ON OPEN ACCESS TO KNOWLEDGE IN THE SCIENCES AND HUMANITIES. Berlin, 2003. Disponível em : <http://www. zim. mpg.de/openaccess-berlin/berlin_declaration. pdf>. Acesso em: 20 dez. 2011

BRASE, Jan; FARQUHAR, Adam. Access to research data. **D-Lib Magazine**, v. 17, n. 1/2, Jan. / Feb. 2011.

CESAR JÚNIOR, Roberto Marcondes. Do mundo aos dados e dos dados ao conhecimento. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Org.). **O quarto paradigma**: descobertas científicas na era da eScience. São Paulo : Oficina do Texto, 2011.

CONWAY, Esther et al. Curating scientific research data for the long term: a preservation analysis method in context. **The International Journal of Digital Curation**, n. 2, v.6, 2011.

DITADI, Carlos. **Preservação de documentos eletrônicos.** Rio de Janeiro: Arquivo Nacional/ CTDE, 2003.

HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin. Jim Gray on eScience: A Transformed Scientific Method. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Org.). **The Fourth Paradigm:** Data-Intensive Scientific Discovery, 2009. Disponível em: <http://research.microsoft. com/en-us/collaboration/ fourthparadigm/4th_ paradigm_book_jim_gray_transcript.pdf>. Acesso em: 20 dez. 2011.

HIGGINS, Sarah. Digital curation: the emergence of a new discipline. **The International Journal of Digital Curation**, v.6, n. 2, 2011. Disponíel em: <http:// www.ijdc.net/index.php/ijdc/article/ view/184>. Acesso em: 20 dez. 2011.

HUNTER, Jane. Scientific publication packages - A selective approach to the communication and archival of scientific output. **The International Journal of Digital Curation,** v.1, n.1, 2006. Disponível em: <http://www.ijdc.net/index.php/ ijdc/article/view/8/4>. Acesso em: 13 jan. 2012.

JANSEN, Hans. Permanent access to electronic journals**. Information Services & Use,** v. 26, 2006. Disponível em: <http://iospress.metapress. com/ content/7drby91r8t4gf8ap/fulltext.pdf >. Acesso em: 10 nov. 2010.

LAGOZE, Carl; SOMPEL, Herbert Van de**. Ore user guide – primer.** Open Archive Initiative, 2008. Disponível em: <http://www. openarchives.org/ore/1.0/primer.html>. Acesso em: 13 jan. 2010.

LANNOM, Laurence. Research Data. **D-Lib Magazine,** v. 17, n. 1/2, Jan. / Feb. 2011. Disponível em: <http://www.dlib.org/dlib/ january11/01editorial.html 2011>. Acesso em: 20 dez. 2011.

LEE, Cristopher; TIBBO, Helen. Digital curation and trusted respositories: steps toward success. **Journal of Digital Information**, v. 8, n. 2, 2007. Disponível em: <http://journals.tdl.org/jodi/article/viewArticle/229/183>. Acesso em: 20 dez. 2011.

NATIONAL SCIENCE BOARD. Long-lived digital data collections: enabling research and education in the 21st century. National Science Foundation, sept. 2005. Disponível em: <http:// www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf >. Acesso em: 01 fev. 2012.

VERHAAR, Peter. **Report on object models and functionalities**. DRIVER II, 2008. Disponível em: <https://openaccess.leidenuniv.nl/bitstream/handle/1887/16018/Report_on_Object_Models_ and_Functionalities.pdf?sequence=2>. Acesso em: 20 dez. 2011.

PARTE 4

PRESERVAÇÃO DE OBJETOS DIGITAIS

Repositórios digitais confiáveis para a preservação de periódicos eletrônicos cientificos

Luis Fernando Sayão[1]

1 Introdução

Desde o seu surgimento no século xvii, os periódicos científicos têm exercido um papel central no processo de comunicação científica. Começando com a publicação em 1665 do *Journal des Sçavans e das Philosophical Transactions of the Royal Society* (day, 1999), por mais de três séculos os principais personagens do ciclo de comunicação científica – autores, editores, bibliotecas, usuários - vêm tendo seus papéis estabelecidos e institucionalizados, juntamente com todo o aparato acadêmico, até a configuração atual, instituída nos fins do século xix. O periódico científico é o coroamento desse sistema de comunicação que cria um compromisso explicito entre a qualidade e a visibilidade na geração de novos conhecimentos científicos. (marcondes; sayão, 2002). O desenvolvimento das ciências é marcado pelos artigos publicados nos periódicos acadêmicos revisados por pares; dezenas de milhares de títulos são publicadas e distribuídos em escala planetária, atingindo cifras que alcançam o patamar de milhões de artigos por ano (venkadesan, 2010).

Os periódicos científicos, desde os seus primórdios, vem sendo distribuídos em forma impressa. Porém, na última década o mercado de publicação científica começou a se deslocar na direção da publicação eletrônica num ritmo muito rápido, gerando um período de transições profundas, fértil em possibilidades, mas também em questionamentos, tensões e problemas inéditos para o mundo acadêmico.

A ruptura com o modelo impresso em prol das formulações digitais abriu possibilidades extraordinárias para o mundo da comunicação científica, libertando definitivamente as publicações acadêmicas dos limites bidimensionais e autocontidos do texto, inaugurando novas formulações de apresentação e interoperabilidade, e, sobretudo, estabelecendo novos padrões de cooperação e interatividade em

1 Doutor em Ciência da Informação (ibict-ufrj), Comissão Nacional de Engenharia Nuclear (cnen), lsayao@cnen.gov.br.

favor da geração de novos conhecimentos. As transformações ainda estão em curso e é difícil prever todos os seus desdobramentos e todas as suas potencialidades.

O deslocamento da impressão em papel para a publicação eletrônica é um fenômeno vertiginoso: prevê-se que por volta do ano de 2016 metade de todas as publicações seriadas terão migrado para formatos unicamente eletrônicos, e os títulos das áreas de ciência, tecnologia e medicina serão os primeiros a se fixarem nesse novo patamar (KENNEY *et al.*, 2006, p. 5). São muitas as forças que impulsionam esse movimento: pesquisadores, bibliotecas, editores, movimento de livre acesso e uma nova conformação do mercado editorial científico. Os pesquisadores, professores, estudantes e outros leitores demandam formatos eletrônicos porque eles oferecem um mundo de vantagens em relação às formas impressas, especialmente no que diz respeito à busca, à recuperação, à navegação, à apresentação das informações e à capacidade de interoperarem com outras publicações eletrônicas que estão em rede.

As bibliotecas acadêmicas, por sua vez, estão crescentemente cancelando as subscrições em papel em favor das licenças eletrônicas para satisfazer as demandas dos seus usuários e para evitar os custos associados com a organização, recepção, catalogação, encadernação, armazenamento e circulação de volumes de papel. Em outro plano, reconhecendo as potencialidades do novo mercado que o meio digital oferece, os editores científicos estão tratando as versões eletrônicas como as versões definitivas (WATERS, 2005). Nessa direção, eles estão deslocando rapidamente seu modelo de negócio de acordo com as novas aspirações do mercado. Além do mais, as facilidades de publicar diretamente na web, aliados aos movimentos de livre acesso aos resultados das pesquisas, permitiram também que a comunidade acadêmica viabilizasse a publicação de periódicos eletrônicos autogeridos.

Esse fenômeno moderno, entretanto, confronta o mundo da ciência com um conjunto de problemas e compromissos inéditos que são inerentes à condição digital das informações que necessita e gera nas suas atividades, como por exemplo, a gestão de *copyright*, economia da informação digital e o controle de qualidade. Publicação eletrônica e auto-arquivamento podem levar a uma proliferação de versões de documentos científicos que torna a qualidade, a autenticidade e a integridade difíceis de assegurar. Nesse museu de grandes novidades, cânones como a revisão por pares e o monopólio dos editores científicos têm sido também colocados em cheque a todo momento.

Dentre todos os problemas inerentes à condição digital da informação, o de mais dramática importância - ainda inscrito na agenda crítica da humanidade a espera de uma solução definitiva – é o perigo real de uma amnésia digital. A ameaça de uma era de esquecimento é causada basicamente por dois problemas que atingem fortemente os documentos digitais: obsolescência tecnológica e a fragilidade das mídias.

Isso acontece porque a informação digital depende, na sua mais pura essência, de um aparato tecnológico para ser acessada e, sobretudo, corretamente interpretada. Mas esse aparato tecnológico de intermediação – formado por hardware, software, mídias formatos - está em constante mutação, em ciclos de obsolescência cada vez mais rápidos determinados principalmente pelo dueto inovação e competição. Contribui ainda enormemente para esse problema o fato dos meios de armazenamentos serem muito frágeis e extremamente suscetíveis à degradação física. Não é um exagero afirmar que a informação digital é mais frágil que os papiros encontrados nas tumbas dos faraós.

A revolução digital está continuamente transformando o modo como os acadêmicos criam, comunicam e preservam o conhecimento científico. Os lugares virtuais distribuídos mundialmente são berços tecnológicos que otimizam a geração cooperativa de novos conhecimento, ao mesmo tempo em que recriam formas de publicação e disseminação. Longe, entretanto, da preocupação com a proteção dos conteúdos, no longo prazo, como assinala Dodebei (2010).

O problema da vulnerabilidade dos materiais digitais confronta o mundo da ciência com a necessidade do arquivamento digital persistente como um elemento crítico que preocupa todos os atores envolvidos. "Criar metodologias que garantam a preservação digital dos estoques científicos em formato digital equivale a estabelecer a interoperabilidade com o futuro" (Owens, 2007).

Tal afirmação enfatiza o fato de que preservar publicações eletrônicas tornou-se uma matéria crítica na medida em que a massa de publicações eletrônicas se multiplica e as comunidades de pesquisa dependem delas tão intensamente como dependiam das coleções em papel (Sayão, 2008).

2 As dimensões do problema

A pesquisa e o ensino – como de resto toda a sociedade – estão crescentemente dependentes de informações e de dados gerados por ferramentas baseadas em computador. Para que haja avanço do conhecimento, esses registros requerem o estabelecimento de metodologias e compromissos de longo prazo que garantam a sua capacidade de serem acessados e decodificados com a tecnologia corrente à época do acesso, e que os usuários potenciais dessas informações possam interpretá-las corretamente. Portanto, o arquivamento persistente, a preservação digital e o estabelecimento de modelos de informação para a preservação de registros científicos estão se tornando questões-chave para as áreas de pesquisa.

É necessário enfatizar que o arquivamento persistente e a preservação digital constituem um problema complexo que envolve muitas variáveis, compromissos de longa duração e impõem a necessidade de grandes investimentos. O custo-benefício

de se manter o acesso de longo prazo aos registros científicos é extremamente difícil de se mensurar. O valor de um registro pode estar relacionado à reprodutibilidade da pesquisa onde ele foi gerado ou capturado: algumas pesquisas podem ser fáceis e baratas de se replicar; outras podem ser literalmente impossíveis de se reproduzir, ou são repetíveis somente a custos e a esforços inaceitáveis (JANSEN, 2006).

Neste patamar de complexidade, verifica-se que dados e informações científicos digitais – principalmente os resultados de pesquisa - necessitam de cuidados específicos; torna-se necessário a criação de novos conceitos de custódia e de gestão conteúdos científicos digitais que incluam ações de arquivamento e preservação, formas de acrescentar valor a esses conteúdos e de otimizar a sua capacidade de reuso.. O termos "curadoria digital" e "centros de curadoria digital" são crescentemente usados nesse contexto (ARELLANO, 2008). Mas é preciso reafirmar que a curadoria digital é uma questão desafiadora, considerando que dados de pesquisas em formato digital formam categoria mais complicada de informação digital para se exercer ações de preservação, pois se apresentam em diversas formas e frequentemente incluem objetos digitais complexos, como bases de dados (ANGEVAARE, 2009).

Quando focamos nossa atenção no domínio específico dos periódicos eletrônicos, fica evidente que o instrumental disponibilizado pelas tecnologias – computadores, capacidade de armazenamento, redes, tecnologias de apresentação e pacotes especializados de software –, aliado a fenômenos recentes como o movimento de livre acesso e autopublicação, têm favorecido e acelerado o aparecimento de novos títulos de periódicos exclusivamente eletrônicos, muitos deles geridos pela própria comunidade acadêmica. Nesse movimento avassalador, só agora o mundo da pesquisa tem colocado as questões pertinentes ao armazenamento persistente e à preservação digital nas suas agendas para ações imediatas e futuras. Ações que enfatizem formulações cooperativos de preservação digital e que reúnam todos os atores tocados pelo problema, posto que a preservação de documentos digitais, no seu sentido mais completo, requer a integração de novos métodos, de políticas, de padrões e de tecnologias, e deve ser sustentada por investimentos financeiros vultosos (RAMESH, 2010).

A nova configuração dos sistemas de publicação de informações acadêmicas demanda responsabilidades em muitos níveis. Entretanto, na medida em que a geração e uso da informação digital se aceleram, a responsabilidade de preservação dos estoques informacionais em formato digital se torna bastante difusa; as partes responsáveis – pesquisadores, gestores, bibliotecas e editores – têm sido lentas em identificar e investir na infraestrutura necessária para assegurar que os registros acadêmicos publicados, representados em formatos digitais permaneçam íntegros ao longo do tempo. Essa inércia coloca a porção digital dos registros

acadêmicos – e a habilidade de usá-los em conjunto com outras informações que são necessárias para o avanço do conhecimento – em risco crescente. A solução pode exigir acordos e compromissos amplos e de longa duração no mundo acadêmico para dividir a responsabilidade de preservação (WATERS, 2005), pois há um consenso claro entre as comunidades envolvidas de que a preservação digital antes de ser problema tecnológico é, sobretudo, um problema organizacional, político e de gestão.

As bibliotecas e outras instituições de conhecimento têm tradicionalmente armazenado e preservado periódicos científicos, desenvolvendo coleções massivas que oferecem ao usuário ou originais ou facsímiles dos trabalhos publicados. Contudo, há um alto grau de incerteza em relação à capacidade dos enfoques tradicionais de preservação de serem suficientemente robustos para lidar com os modos inéditos de como os resultados de pesquisa vão ser processados, disseminados, revisados, acessados e mantidos no futuro. Isto porque, arquivar e preservar objetos digitais é fundamentalmente diferente de arquivar e preservar objetos impressos (HOORENS; ROTHENBERG, 2008).

A perplexidade que se instala tem origem no fato de que a tecnologia de informação causou uma mudança significativa nos papéis tradicionais das bibliotecas e dos editores científicos. Uma das transformações mais contundentes foi o deslocamento da responsabilidade de arquivamento das bibliotecas para os editores, no domínio das publicações eletrônicas (MOGHADDAN; MOBALLEGHI, 2009). Nesse novo cenário, porém, há uma ruptura na continuidade do arquivamento: muitas bibliotecas têm a custódia de periódicos impressos desde os primeiros números; por outro lado, poucos editores têm arquivado a coleção completa de seus periódicos eletrônicos para a posteridade.

Na era do papel, sempre existiu uma redundância em larga escala no armazenamento dos periódicos. Muitas e diferentes instituições coletam o mesmo título. As cópias salvas para as futuras gerações são as mesmas cópias lidas pela geração atual de usuários. Muitas das metodologias e técnicas usadas para ajudar a manter os periódicos impressos por longo prazo – encadernação, conservação, controle ambiental, etc. - fazem parte das funções da biblioteca no oferecimento dos serviços a seus usuários. Entretanto, o modelo comum de serviço para os periódicos eletrônicos é dramaticamente diferente do modelo dos periódicos tradicionais. A maioria do acesso aos periódicos eletrônicos é oferecida somente pelo seu editor ou pelo seu agente. Existe um nível baixo de replicação e somente umas poucas instituições mantêm copias de periódicos eletrônicos localmente. As bibliotecas podem cumprir as exigências dos serviços atuais sem se envolver em questões de preservação dos recursos informacionais. No mundo digital, as questões envolvi-

das no dia-a-dia dos serviços de informação são bem diferentes e apartadas das questões envolvidas na preservação de longo prazo.

Na perspectiva das bibliotecas acadêmicas cujas coleções de periódicos estão sendo substituídas por licenças de acesso, a percepção sobre a perda da posse física da publicação torna-se uma preocupação constante. Quando as bibliotecas acadêmicas e de pesquisa subscrevem títulos de periódicos eletrônicos elas não têm a possessão de uma cópia dos exemplares como antes. Elas usam o conteúdo armazenado em sistemas remotos controlados pelos editores. Embora algumas licenças reconheçam que as bibliotecas têm o direito permanente de uso dos conteúdos dos periódicos eletrônicos, esses direitos permanecem em grande parte no plano teórico. Se um editor falha em manter seus arquivos ou se se retira do negócio por qualquer razão e deixa de tornar disponível o título do qual um campo particular de pesquisa depende, não existem meios práticos para substituir o direito permanente de uso da publicação por parte da biblioteca. Dessa forma, os registros ficam expostos ao risco de se perderem.

Muitos dos atores envolvidos com o problema consideram que estabelecimento de um modelo de negócio para o arquivamento persistente e para a preservação digital – ou seja, a definição de quem oferece acesso para quem, em que forma, em que momento e quem paga - como o maior desafio a ser enfrentado pelos editores científicos, pelas bibliotecas de pesquisa e pelos demais protagonistas desse problema. Isto porque a definição de um modelo é ameaçada por incertezas que tornam o desenvolvimento de estratégias de arquivamento extremamente difíceis. Contudo, por trás dessas incertezas, é possível identificar as interrogações que vão ajudar a ordenar o futuro da preservação digital: Como as pesquisas serão comunicadas no futuro? O que cada deve preservar? Quem paga o quê? O que as bibliotecas de pesquisas e universitárias devem demandar dos serviços de arquivamento e preservação? (hoorens; rothenberg, 2008; vekadesan, 2010).

São muitas as dimensões do problema. Jansen (2006) tenta identificar as principais questões segundo a perspectiva dos principais atores: biblioteca, usuários e editores científicos.

Sobre a perspectiva da biblioteca, as questões que estão em pauta são as seguintes:

- Quais são os direitos de acesso permanente ao material já pago, principalmente quando a biblioteca suspende a assinatura do periódico?
- O que acontece quando o editor retira um trabalho eletrônico do acesso on-line, ou se afasta do negócio, ou torna, por outro motivo qualquer, o acesso inviável?
- Quem vai assegurar que os arquivos vão manter a sua usabilidade? E quem vai pagar por isso?

Para o autor cuja primazia do seu trabalho acadêmico está registrada num periódico unicamente eletrônico, que tem a avaliação da sua atividade baseada nas suas publicações e necessita de visibilidade entre seus pares, a preocupação está também em volta do problema recursivo do acesso persistente. Mas as questões sobre a integridade, a autenticidade e a estabilidade dos seus originais, enquanto versão definitiva, se tornam um foco novo de preocupações. As respostas às seguintes questões são de grande importância para ele:

- O meu trabalho vai estar disponível para sempre?
- Como eu posso assegurar que o meu trabalho não será alterado se ele está somente em forma eletrônica?
- Como vão ser controladas as várias versões e manifestações do meu trabalho?

O editor precisa demonstrar aos que licenciam os seus produtos que eles permanecerão estáveis, íntegros e acessíveis sob qualquer circunstância. Suas questões são as seguintes:

- Como assegurar à comunidade acadêmica como um todo que existe um mecanismo confiável de arquivamento digital de longo prazo para as minhas publicações?
- Como assegurar que os *links* e os *links* referenciais das minhas publicações permanecerão estáveis ao longo do tempo, mantendo a integridade das minhas publicações e os seus relacionamentos com outras publicações?

Os periódicos eletrônicos estão integrados à vida acadêmica há mais de uma década. Embora alguns dos grandes editores tenham anunciado que eles estão tomando para si a responsabilidade por manter por longo prazo os seus conteúdos eletrônicos, a maioria dos pequenos editores – incluindo aqueles vinculados à própria comunidade acadêmica – ainda está em dúvida de como preservar as suas publicações eletrônicas, ou, o que ainda é mais grave, não tem a noção exata da dimensão e da complexidade do problema e clareza sobre o seu papel nesse contexto de incertezas.

O levantamento promovido pelo serviço de armazenamento digital Portico[2] e pela Universidade de Ithaca em 2008, que consultou 1.371 diretores de bibliotecas universitárias americanas sobre as perspectivas atuais de preservação de periódicos eletrônicos – em termos de atitudes, prioridades, recursos e ações das bibliotecas.

2 Disponível em: http://www.portico.org. Acesso em: 3 ago. 2021.

-, indica que a maioria das bibliotecas está significativamente incerta sobre as suas opções de preservação de periódicos eletrônicos. Entretanto, há um consenso absoluto entre os diretores de que a perda potencial desses periódicos é inaceitável e que suas instituições têm obrigação de tomar ações que evitem o surgimento de lacunas nos registros de pesquisa (PORTICO, 2008).

A academia não tem ainda um equivalente funcional para periódicos eletrônicos do tipo "possuir uma copia" oferecido pelo padrão impresso. Até que se crie um mecanismo de arquivamento digital permanente, a academia não pode deslocar-se inteiramente para o mundo dos periódicos unicamente eletrônicos, e não pode usufruir inteiramente os benefícios dessa mudança. (WATERS, 2005). Este problema que iremos analisar a seguir.

3 A proposta de repositórios digitais

Para contornar as incertezas provocadas pela fragilidade tecnológica e organizacional dos periódicos eletrônicos, o mundo acadêmico vem, nos últimos anos, estabelecendo pactos que tentam viabilizar trabalhos cooperativos em torno dessa questão. Essas iniciativas têm como objetivo primário equacionar soluções técnicas, gerenciais, organizacionais e normativas para criar mecanismos de preservação dos conteúdos dos periódicos eletrônicos, que representam, em grande parte, o testemunho da geração dos saberes científicos atuais.

No contexto desse movimento, os repositórios para versões impressas estão sendo continuamente desenvolvidos em âmbito nacional, regional e mesmo localmente, na tentativa de assegurar que pelo menos uma cópia em papel permaneça acessível. Entretanto, as instituições crescentemente reconhecem que a forma impressa não é um formato de arquivamento aceitável para conteúdos eletrônicos, dado que isso significa abdicar das funcionalidades conferidas pelo formato digital dos conteúdos e da sua conectividade, ou seja, da sua qualidade de estar em rede e vinculados por *hyperlinks* a outros documentos.

De uma forma geral, cada país tem uma política de depósito legal para publicações impressas e, na maioria dos casos, esses acervos estão sediados nas bibliotecas nacionais. Gradativamente, esses depósitos oficiais estão incorporando repositórios digitais aos seus sistemas de depósito legal, destinados à preservação de longo prazo das publicações eletrônicas produzidas dentro das fronteiras nacionais. Entretanto, esse modelo tradicional de depósito baseado em estados nacionais e fronteiras geográficas, pode ser uma solução parcial para os periódicos eletrônicos publicados num país, mas pode não ser capaz de garantir a permanência e a segurança das publicações científicas internacionais. Isto acontece porque: a) a literatura acadêmica em formato digital é, em muitos casos, desterritorializada e

nem sempre tem um país nativo e, consequentemente, não possui um guardião óbvio; b) a velocidade com que as bibliotecas nacionais podem desenvolver seus repositórios digitais não acompanha o ritmo de multiplicação dos periódicos eletrônicos; e c) não se pode esperar que os editores internacionais depositem seus conteúdos num grande número de repositórios digitais nacionais. Por todas essas razões existe um risco considerável de que, circunscrito ao modelo tradicional, os registros eletrônicos científicos possam não sobreviver ao longo do tempo (JANSEN, 2006). Fica claro, portanto, que é necessário um enfoque mais sistemático e mais específico para o problema.

Existem primariamente duas opções para assegurar acesso contínuo a conteúdos de periódicos eletrônicos licenciados. A primeira delas está baseada inteiramente na confiança de que o editor ou distribuidor irá oferecer acesso permanente aos conteúdos que foram subscritos, mesmo que o editor pare de publicar os títulos ou a biblioteca pare de subscrevê-los. A segunda opção – comum em muitos contratos de licença – está fundamentada na exigência de que o editor repasse à biblioteca cópias dos arquivos que constituem os periódicos eletrônicos que foram subscritos por ela. Ambas as soluções passam necessariamente pelo desenvolvimento de sistemas de arquivamento digital que incorporem soluções tecnológicas e organizacionais que assegurem - no caso de algum evento que impeça o acesso regular ao recurso – algum grau de acesso aos conteúdos dos periódicos licenciados, com um nível de qualidade aceitável.

Angevaare (2009), coordenadora do *Netherlands Coalition for Digital Preservation* (NCDD)[3], argumenta que a maioria das bibliotecas nacionais não tem opção: elas têm o dever legal de atuar como bibliotecas depositárias para materiais impressos e digitais. As bibliotecas de pesquisa, porém, têm outras opções para preservar os dados e informações digitais que elas custodiam: a) simplesmente armazenar as coleções digitais ambientes internos da organização e esperar o melhor do sistema de armazenamento, apesar de confortável, esta é uma opção de risco; b) encontrar um serviço operados por terceiros – governamental ou privado – que abrigue as suas coleções digitais; esta parece ser a opção mais viável para a maioria das bibliotecas de pesquisa, entretanto é necessário encontrar um repositório que seja confiável, e que seja capaz também de integrar o acesso às informações armazenadas no sistema da biblioteca; c) desenvolver o seu próprio arquivo digita; esta é a opção mais ambiciosa.

É preciso salientar que opção de desenvolver repositório próprio tem um grau maior de viabilidade quando há um enfoque coletivo, como veremos mais adiante.

3 Disponível em: http://www.ncdd.nl/en/index.php. Acesso em: 3 ago. 2021.

Os empreendimentos cooperativos tem maior chance de sustentabilidade política, financeira e metodológica. Não obstante, o sonho do repositório próprio tem se tornado real para muitas instituições acadêmicas por meio dos repositórios institucionais.

Nessa direção, em complementação as exigências de direitos de acesso permanente, as bibliotecas e outras instituições de conhecimento estão crescentemente fundando repositórios institucionais usando pacotes livres de softwares – Dspace, Fedora e outros - e pressionando os editores no sentido de garantir aos autores direitos de modalidades viáveis de auto-arquivamento. Isto porque, para que os repositórios institucionais atendam às expectativas do movimento do acesso livre e do arquivamento seguro eles dependem de instrumentos que garantam o depósito, por parte dos pesquisadores, dos seus trabalhos publicados em periódicos revisados por pares. "Em todo mundo, as universidades vêm estabelecendo mandatos (também chamados de políticas) para garantir o povoamento dos repositórios digitais" (KURAMOTO, 2010, p. 207).

No limiar do ano 2000, a preocupação das bibliotecas de pesquisa em relação à perpetuação do acesso aos conteúdos dos periódicos eletrônicos criou um ambiente favorável ao debate em torno da idéia de programas confiáveis de arquivamento para esse gênero de publicação eletrônica. Essa idéia foi se consolidando e, nos dias de hoje, há um consenso quase unânime dos especialistas na área de que a sustentação das pesquisas futuras e do ensino vai depender da fundação de repositórios digitais confiáveis, nos quais as publicações acadêmicas registradas em formato digital possam persistir independentes do controle exclusivo dos editores, independente de esforços individuais de bibliotecas e sob o controle de entidades comprometidas com valores de longo prazo (WATERS, 2005).

4 Adicionando confiabilidade aos repositórios institucionais

A constatação de que a transitoriedade da web não a habilita como memória faz com que tenhamos que construir intencionalmente os espaços de memória que a sociedade necessita para preservar por longo prazo os seus estoques informacionais digitais.

No mundo acadêmico principalmente, a ideia de se criar estes espaços persistentes na web se concretiza com o desenvolvimento dos repositórios digitais confiáveis, que tem como base um documento fundador elaborado por um grupo de trabalho internacional liderado pela *Research Library Group* (RLG) e pela *Online Computer Library Center* (OCLC), cujo título é *Trusted Digital Repositories: Atributes and Responsabilities* (RESEARCH LIBRARY GROUP, 2002).

O relatório enuncia as bases conceituais para os repositórios digitais confiáveis,

estabelecendo uma definição, os atributos e as responsabilidades que devem ser assumidas por eles. Inclui ainda uma discussão importante sobre a como se recria a idéia tradicional de confiança – que é essencial para as instituições de patrimônio - para um ambiente instável como a web.

Na perspectiva do grupo de trabalho RLG/OCLC (2002), um "repositório digital confiável tem como missão oferecer à sua comunidade-alvo acesso confiável e de longo prazo aos recursos digitais por ele gerenciados, agora e no futuro." (RESEARCH LIBRARY GROUP, 2002, p. 5). Para cumprir essa missão, os repositórios digitais devem: aceitar, em nome de seus depositantes, a responsabilidade pela manutenção por longo prazo de recursos digitais; ter um sistema organizacional que apoie não somente a viabilidade de longo prazo do repositório, mas também a informação digital da qual ele tem responsabilidade; demonstrar responsabilidade fiscal e sustentabilidade; projetar seu(s) sistema(s) de acordo com convenções e padrões comumente aceitos no sentido de assegurar a gestão, o acesso e a segurança continua dos materiais depositados; estabelecer metodologias para avaliação dos sistemas que considerem as expectativas de confiabilidade esperadas pela comunidade; considerar, para desempenhar suas responsabilidades de longo prazo, os depositários e os usuários de forma aberta e explicita; ter políticas, práticas e desempenho que possam ser auditáveis e mensuráveis; e por fim, cumprir uma série responsabilidades.

O Relatório do grupo de trabalho RLG/OCLC (2002) identifica ainda as qualidades que devem possuir os repositórios confiáveis. Ele estabelece uma grade de atributos que acomodam diferentes situações e responsabilidades, ao mesmo tempo em que oferece uma base para o que se esperar de um repositório confiável. Dentre os itens, onde se incluem sustentabilidade, segurança de sistemas, adequação tecnológica, responsabilidade administrativa, se destaca a conformidade à infraestrutura estabelecida pelo Modelo de Referencia, *Open Archival Information System* (OAIS) (CONSULTATIVE COMMITEE FOR SPACE DATA SYSTEM, 2002), no que concerne ao modelo funcional e ao modelo de informação. É importante assinalar que esses atributos estão muito mais próximos às questões organizacionais do que as tecnológicas (OWENS, 2007).

As instituições tradicionais de patrimônio – museus, bibliotecas e arquivos – desfrutam de uma justa confiança do seu publico-alvo, dado que têm preservado uma grande quantidade de registros, de toda a natureza, ao longo da história (RESEARCH LIBRARY GROUP, 2002). Os repositórios digitais, porém, devem conquistar confiança de uma forma mais objetiva e mensurável. A informação digital é menos tangível que outros materiais e muito mais instável, isto pode significar que elementos como "confiança" e "credibilidade" possam ser mais difíceis de provar e mensurar.

A norma OAIS oferece uma referência sólida para os termos, conceitos e fluxos de informações que circunscrevem um repositório OAIS, entretanto ele não toca em prescrições de implementação. Como ter certeza de que um repositório digital segue as práticas e procedimentos que vão assegurar a preservação de longo prazo?

Adicionar confiança aos repositórios digitais implica no estabelecimento da presunção de que um dado repositório digital é o que diz ser e que a informação armazenada lá está segura por longo prazo. Isso é conferido, principalmente, pelas ações de certificação que se "torna um componente-chave para repositórios digitais contemporâneos" (THOMAZ, 2007, p. 84).

Um dos documentos essenciais para se estabelecer graus de confiabilidade de repositórios digitais é o *Trustworthy Repository Audit & Certification: Criteria and Checklist* (RESEARCH LIBRARY GROUP, 2007), mais conhecido pela sigla TRAC. Conforme expressa o seu título, o documento apresenta um conjunto de critérios e um *checklist* que são tomados como referência para a certificação de repositórios digitais. Nessa direção, ele oferece ferramentas para auditoria, avaliação e certificação potencial de repositórios; estabelece a documentação exigida para a auditoria; delineia um processo de certificação; e estabelece as metodologias apropriadas para determinar a solidez e a sustentabilidade de repositórios digitais.

Contudo, há uma forte demanda por por parte das comunidades envolvidas por um instrumento de normatização completo e de alcance amplo. Essa demanda, que está expressa no próprio corpo do Modelo de Referência OAIS (CONSULTATIVE COMMITEE FOR SPACE DATA SYSTEM, 2002, p. 1-4), impulsionou a formação, em 2007, de um grupo de trabalho que tem como objetivo elaborar uma norma internacional sobre a qual a auditoria e a certificação plena de repositórios digitais serão baseadas.

O TRAC - juntamente com outras ferramentas importantes como o *Digital Repository Audit Method on Risk Assessement* (DIGITAL CURATION CENTRE, 2007), conhecido pela sigla DRAMBORA e o *Catalogue of Criteria for Trusted Digital Repositories* (NESTOR, 2006) - fixa os fundamentos que orientam o desenvolvimento, ainda em curso, da norma.

A elaboração da norma é capitaneado pelo *Consultative Committe* for *Space Data System*[4] (CCSDS) e segue a mesma metodologia aplicada no desenvolvimento do Modelo de Referência OAIS, com ampla participação da comunidades, discussões e *workshops*. Registros das discussões, esboços e documentos de trabalho *bem como* a minuta do "livro vermelho" *Requirements for bodies providing audit and certification of candidate trustworthy digital repositories* (CONSULTATIVE COMMITEE

4 Disponível em: http://public.ccsds.org/default.htm. Acesso em: 3 ago. 2021.

for space data system, 2010) estão diponíveis para exame e comentários no *website Digital Repository Audit and Certification Wiki*[5].

5 PREMIS: o modelo OAIS em ação

No contexto dos periódicos eletrônicos, a instituição de sistemas de arquivamento digital parte do pressuposto de que a "preservação de periódicos eletrônicos é uma espécie de seguro e não uma forma de acesso" (waters, 2005, p. 2) que tem como foco a gestão de risco contra a perda permanente de conteúdos digitais importantes para a pesquisa e para o ensino; contra a possibilidade de cessar, por falha dos editores, os meios de acesso a esses conteúdos.

Para equacionar esses fatores de risco e estabelecer uma forma de seguro contra perdas, os repositórios qualificados de arquivamento para a preservação devem oferecer um patamar mínimo de serviços bem definidos. Nessa direção, eles devem: 1) receber de uma biblioteca participante ou diretamente do editor os arquivos que constituem um periódico eletrônico em uma forma padronizada; 2) armazenar os arquivos em formatos não proprietários de forma que possam ser facilmente transferidos e usados; 3) usar meios padronizados para verificar a integridade dos arquivos e oferecer mecanismos de verificação continua de integridade dos arquivos armazenados internamente; 4) limitar o processamento dos arquivos recebidos com o propósito de manter baixos os custos operacionais, entretanto, deve oferecer processamento suficiente para que os arquivos possam ser localizados e adequadamente apresentados para bibliotecas participantes nos casos de eventos de perda; 5) restringir o acesso por parte das bibliotecas participantes aos arquivos depositados que estão protegidos por *copyright*, tendo como propósito proteger os interesses comerciais dos editores, porém isso não é válido para os casos em que os editores estão incapacitados de oferecer acesso, ou os conteúdos não estão mais protegidos por *copyright*; 6) oferecer um meio transparente e aberto de auditar as práticas de arquivamento adotadas pelo repositório (waters, 2005).

Passado alguns anos, uma série de avanços no cenário internacional está criando condições para estabilização das condições de acesso aos periódicos eletrônicos. Esses esforços estão começando a render frutos: as bibliotecas acadêmicas estão oferecendo opções viáveis para o arquivamento periódicos eletrônicos; os editores científicos estão colaborando com as organizações de conhecimento oferecendo "repositórios ocultos" – isto é, repositórios que não permitem acesso on-line rotineiro – para os seus *backfiles*. Em muitos países, a legislação sobre

5 Disponível em: http://wiki.digitalrepositoryauditandcertification.org/bin/view. Acesso em: 3 ago. 2021.

depósito legal que orienta o depósito de publicações on- line inclui periódicos eletrônicos; e existe uma vinculação próxima do movimento de livre acesso com a preservação digital.

6 Iniciativas importantes

Estabelecido o consenso em torno da fundação de repositórios digitais confiáveis, um número de experiências importantes começou a ser desenvolvida, sempre almejando a sustentação metodológica e financeira da idéia.

As preocupações e as tensões geradas pelos problemas com a preservação dos periódicos eletrônicos exigiram que editores, bibliotecários e tecnologistas se articulassem em torno de uma solução comum. Essa articulação é marcada primordialmente pelo pacto estabelecido em 1999 por três importantes organizações ligadas à questão da informção digital - *Council on Library and Information Resource* (CLIR), *Digital Library Federation* (DLF) e *Coalition Networked Information* (CNI). Essas organizações convocaram um grupo de editores e bibliotecários para discutir a responsabilidade de arquivamento dos conteúdos dos periódicos eletrônicos. Uma série de reuniões resultou na publicação em maio de 2000 do documento fundamental "*Minimum Criteria for an Archival Repository of Scholarly Jounals*", versão 1.2. (DIGITAL LIBRARY FEDERATION, 2000)

Esse documento, conforme proclama a sua introdução, estabelece os critérios mínimos para um repositório digital que atua para preservar publicações digitais acadêmicas. Ele é baseado muito proximamente ao Modelo de Referência OAIS (CONSULTATIVE COMMITEE FOR SPACE DATA SYSTEM, 2002), modificado para refletir as necessidades específicas das bibliotecas, dos editores e das comunidades científicas. O documento também aponta algumas das pesquisas-chave que deverão ser desenvolvidas para aperfeiçoamento dos repositórios digitais acadêmicos aderentes aos critérios que provavelmente emergirão. Essas pesquisas são divididas em três categorias: as associadas com o depósito de dados, as associadas com a preservação e as pesquisas associadas com o acesso. Os critérios estabelecidos são os seguintes (DIGITAL LIBRARY FEDERATION, 2000):

- **Critério 1** – Um repositório digital que atua para preservar publicações acadêmicas digitais será uma parte confiável que estará em conformidade com os requisitos mínimos acordados entre os editores científicos e as bibliotecas;
- **Critério 2** – Um repositório definirá sua missão considerando as necessidades dos editores científicos e das bibliotecas de pesquisa. Ele também deve explicitar quais publicações acadêmicas ele pretende arquivar e para qual comunidade-alvo a publicação está sendo arquivada;

- **Critério 3** - Um repositório negociará e aceitará depósitos de editores científicos. Esse critério indica que o repositório deverá desenvolver diretrizes sobre quais publicações serão aceitas para arquivamento;
- **Critério 4** – Um repositório deverá obter controle suficiente da informação depositada com o propósito de assegurar a sua preservação de longo prazo;
- **Critério 5** – Um repositório deverá seguir políticas e procedimentos documentados que assegurem que a informação será preservada contra todas as contingências razoáveis;
- **Critério 6** – Um repositório deverá manter preservada a informação disponível para a biblioteca sob condições negociadas com o editor;
- **Critério 7** – Os repositórios deverão funcionar como parte de uma rede.

Ainda nos idos de 2000, logo após a publicação do documento *Minimum Criteria for an Archival Repository of Scholarly Jounals* (DIGITAL LIBRARY FEDERATION, 2000), a Fundação Andrew W. Mellon, trabalhando em conjunto com o CLIR, tomou a iniciativa pioneira de saltar do patamar de trocas de ideias para a experimentação e para a implementação (FLECKER, 2001). Nessa direção, convidou as principais bibliotecas de pesquisa americanas a apresentarem propostas de projetos voltados para a criação e operação experimental de arquivos de periódicos eletrônicos. Como resultado, sete propostas foram contempladas com recursos para desenvolverem projetos no período de 2001-2002. Os projetos tinham perspectivas diferentes. Enquanto os projetos das universidades de Harvard, Pennsylvania e Yale estavam orientados para os editores científicos, que eram parceiros na empreitada, os projetos da universidade de Cornell e da *New York Public Library* estavam orientadas por temas (agricultura e arte performática respectivamente). Por sua vez, o *Massachusetts Institute of Technology* (MIT) propôs investigar os desafios dos "periódicos eletrônicos dinâmicos" – *websites* acadêmicos que não seguem os padrões tradicionais de periódicos. A Universidade de Stanford recebeu fundos para desenvolvimento e testes beta do sistema *Lots of Copies Keep Stuff Safe* – mais conhecido pela sigla LOCKSS -, que tem como objetivo apoiar de forma automática e com baixo custo a replicação em larga escala do conteúdo de periódicos eletrônicos. O trabalho desenvolvido pelas universidades Harvard, Yale e Cornell tem influenciado fortemente o trabalho do JSTOR *Electronic-Archiving Initiative*, agora chamado de Portico, uma referência importante na área de preservação de periódicos eletrônicos (OWENS, 2007).

Mesmo trabalhando com visões distintas do problema, os projetos, de forma geral, compartilham um conjunto de pressupostos básicos (FLECKER, 2001):

- Os repositórios devem ser independentes de editores. Suas necessidades de arquivamento devem ser responsabilidade das instituições para as quais eles prioritariamente se dirigem;
- O arquivamento deve ser baseado em parcerias ativas com os editores. Isso exigirá tipos de licenças diferentes das licenças atuais voltadas para o uso do conteúdo;
- Os repositórios devem considerar o problema da preservação sob o prisma de longuíssimo prazo – cem anos ou mais;
- Os repositórios deverão estar em conformidade com as normas, os protocolos e com as diretrizes de melhores práticas do mundo digital e acompanhar o desenvolvimento desses marcos. Os repositórios devem ainda estar sujeitos à auditoria e à certificação;
- Os repositórios devem ser baseados no Modelo Referencial OAIS, elaborado pela NASA que estabelece as informações e as funções necessárias para a preservação digital de longo prazo.

Considerando a importância dos periódicos eletrônicos como meio primário de disseminação de uma parte significante de nossa herança intelectual, não é surpresa que existam várias iniciativas em andamento por todo o mundo com o objetivo de preservar esses estoques de conhecimento. Esses projetos caracteristicamente expressam a necessidade de uma resposta coletiva ao desafio de se manter o acesso perene aos periódicos eletrônicos, dado a impossibilidade quase absoluta de se ter respostas individuais e específicas a essa questão. Os esforços contínuos de pesquisa e desenvolvimento necessários ao estabelecimento de sistemas de repositórios confiáveis exigem compromissos financeiros, técnicos e comprometimento e expertise das equipes que excedem as possibildades de instituiçoes individuais, resume RAS (2009). A lista abaixo mostra que o cooperativismo parece ser a marca mais óbvia das principais iniciativas.

- *Canada Institute for Scientific and Technical Information - National Science Library Trusted Digital Repository* (TDR)[6];
- LOCKSS *Alliance*[7] e CLOCKSS[8];

6 Disponível em: http://cisti-icist.nrc-cnrc.gc.ca/eng/ibp/cisti/about/overview-initiatives.html. Acesso em: 3 ago. 2021.

7 Disponível em: http://www.lockss.org/lockss/LOCKSS_Alliance. Acesso em: 3 ago. 2021.

8 Disponível em: http://www.clockss.org/clockss/Home. Acesso em: 3 ago. 2021.

- Portico[9];
- *Koninklijke Bibliotheek e-Depot* (KB *e-Depot*)[10];
- *Kooperativer Aufbau eines Langzeitarchivs Digitaler Informationen* (kopal/DDB)[11];
- *Los Alamos National Laboratory Research Library* (LANL-RL)[12];
- *National Library of Austrália* PANDORA (NLA PANDORA)[13];
- OCLC *Electronic Collection Online* (OCLC ECO)[14];
- *OhioLINK Electronic Journal Center* (OhioLINK EJC)[15];
- *Ontario Scholars Portal* [16];
- *PubMed Central*[17].

7 Algumas recomendações à guisa de conclusão

O levantamento coordenado por Anne Kenney e seus colaboradores (KENNEY *et al.*, 2006) alinha um conjunto de recomendações dirigidas às bibliotecas acadêmicas, aos editores e aos programas de arquivamento de periódicos eletrônicos que bem podem servir como fechamento do presente trabalho.

7.1 Recomendações para as bibliotecas e organizações acadêmicas

- As bibliotecas e consórcios devem pressionar os editores científicos para que eles se incorporem em programas confiáveis de arquivamento de periódicos eletrônicos; e que transfiram todos os direitos e responsabilidades necessários ao arquivamento digital como parte da negociação das licenças de subscrição. As bibliotecas de pesquisa devem coletivamente concordar em não assinar novas licenças e renovações para acesso a periódicos eletrônicos se essas condições não forem satisfeitas.
- As bibliotecas devem compartilhar informações com outras bibliotecas sobre as soluções adotadas para o arquivamento de periódicos eletrônicos.

9 Disponível em: http://www.portico.org. Acesso em: 3 ago. 2021.
10 Disponível em: http://www.kb.nl/dnp/e-depot/e-depot-en.html. Acesso em: 3 ago. 2021.
11 Disponível em: http://kopal.langzeitarchivierung.de/index_koLibRI.php.de. Acesso em: 3 ago. 2021.
12 Disponível em: http://library.lanl.gov/ . Acesso em: 3 ago. 2021.
13 Disponível em: http://pandora.nla.gov.au/. Acesso em: 3 ago. 2021.
14 Disponível em: http://www.oclc.org/electroniccollections/. Acesso em: 3 ago. 2021.
15 Disponível em: http://www.ohiolink.edu/. Acesso em: 3 ago. 2021.
16 Disponível em: http://www.scholarsportal.info/index.html. Acesso em: 3 ago. 2021.
17 Disponível em: http://www.pubmedcentral.nih.gov/. Acesso em: 3 ago. 2021.

- As bibliotecas devem tornar-se membro ou participar de pelo menos uma iniciativa de arquivamento mais adequada aos seus propósitos; bem como participar no desenvolvimento de modelos de registros de informações sobre publicações acadêmicas arquivadas.
- As bibliotecas devem ainda influenciar os programas de arquivamento para participarem de redes de compartilhamento de informações, sistematizar melhores práticas e promover redundância num nível suficiente para assegurar a persistência dos conteúdos.

7.2 Recomendações para os editores científicos

- Os editores científicos devem tornar público os seus esforços de arquivamento digital e se incorporarem em alguns dos programas importantes de arquivamento de periódicos eletrônicos;
- Devem também tornar mais liberais as cláusulas sobre direitos de arquivamento nos acordos fechados com os consórcios e com os agregadores de conteúdo, de forma que o arquivamento dos periódicos eletrônicos seja uma responsabilidade compartilhada entre todos os envolvidos;
- Os editores devem ainda oferecer informações suficientes aos programas de arquivamento para que o processo de depósito seja adequadamente registrado.

7.3 Recomendação para os programas de arquivamento de periódicos eletrônicos

- Os programas de arquivamento devem apresentar evidências públicas de que eles oferecem um patamar mínimo de serviços voltados para a manutenção da coleção. Eles devem estar abertos a auditorias e – quando a certificação de repositórios confiáveis estiver disponível – eles devem ser certificados;
- Os programas de arquivamento devem disponibilizar informações sobre editores, títulos, datas e conteúdos incluídos no repositório; essas informações devem ser facilmente obtidas nas páginas web dos programas;
- Uma vez que as informações tenham sido incorporadas ao repositório, elas se tornam sua propriedade e não podem ser removidas ou modificadas pelo editor ou seus sucessores;
- Os contratos que regem a custódia do programa de arquivamento devem ser periodicamente revistos, posto que mudanças em relação ao editor – aquisição, fusão, etc. -, criação do conteúdo, forma de disseminação e tecnologia podem afetar os direitos e responsabilidades sobre o arquivamento. É

necessário considerar também que alguns conteúdos podem eventualmente entrar em domínio público e isso deve ser considerado nos acordos com o editor;

- Finalmente, os programas de arquivamento devem se organizar em rede de apoio e mútua dependência para troca de informações sobre cobertura de conteúdo, tecnologias, melhores práticas e formas de obtenção das condições contratuais necessárias para preservação e, quando for caso, oferecer acesso aos conteúdos.

Referências

ANGEVAARE, Inge. Take care of digital collection and data: "curation" and organizational choices for research libraries. **Liber Quarterly**, [*S. l.*], v. 19, n. 1, p. 1-12, 2009. Disponível em: http://liber.library.uu.nl/publish/articles/000278/article.pdf. Acesso em: 6 ago. 2021.

ARELLANO, Miguel Ángel Márdero. **Critérios para a preservação digital da informação científica**. 2008. 306 f. Tese (Doutorado em Ciência da Informação) - Universidade de Brasília, Departamento de Ciência da Informação, Brasília, DF, 2008. Disponível em: https://core.ac.uk/download/pdf/11884842.pdf. Acesso em: 6 ago. 2021.

CONSULTATIVE COMMITEE FOR SPACE DATA SYSTEM. **Requirements for bodies providing audit and certification of candidate trustworthy digital repositories**. Washington, DC: [*s. n.*], 2010. Disponível em: http://wiki.digitalrepositoryauditandcertification.org/pub/Main/WebHome/Requirement sForBodiesProvidingAuditAndCertification-SecRev1.doc. Acesso em: 10 nov. 2010.

CONSULTATIVE COMMITEE FOR SPACE DATA SYSTEM. **Reference model for open archival information system (OAIS)**: recommendation. Washington, DC: CCSDS, 2002. Disponível em: http://public.ccsds.org/publications/archive/650x0b1.pdf. Acesso em: 10 nov. 2010.

DAY, Michael. The scholarly journal in transition and the PubMed Central proposal. **Ariadne**, Loughborough, v. 21, [p. 1-11], 1999. Disponível em: http://www..ariadne.ac.uk/issue21/pubmed/. Acesso em: 15 jan. 2008.

DIGITAL CURATION CENTER. Digital Preservation Europe. **Digital repository audit method based on risk assessment** (DRAMBORA). Edimburgo: DCC, 2007. Disponível em: http://www.repositoryaudit.eu/download/. Acesso em: 23 julho 2007.

DIGITAL LIBRARY FEDERATION. **Minimum criteria for an archival**

repository of digital scholarly journals. Alexandria: DLF, 2000. Disponível em: https://old.diglib.org/preserve/criteria.htm. Acesso em: 9 ago. 2021.
DODEBEI, Vera. Repositórios institucionais: por uma memória criativa no ciberespaço. *In*: SAYÃO, Luis Fernando *et al.* (org.). **Implantação e gestão de repositórios institucionais**: política, memória, livre acesso e preservação. Salvador: EDUFBA, 2010. p. 83-106. Disponível em: https://repositorio.ufba.br/ri/bitstream/ufba/473/3/implantacao_repositorio_web.pdf. Acesso em: 8 ago. 2021.
FLECKER, Dale. Preserving scholarly e-journals**. D-Lib Magazine**, [*S. l.*], v. 7, n. 9, [p. 1-8], 2001. DOI: 10.1045/september2001-flecker. Disponível em: http://www.dlib.org/dlib/september01/flecker/09flecker.html. Acesso em: 9 ago. 2021.
HOORENS, Stijn; ROTHENBERG, Jeff. **Digital preservation**: the uncertain future of saving the past. Cambridge: RAND Europe, 2008. Disponível em: http://www.rand.org/content/dam/rand/pubs/research_briefs/2008/RAND_RB9331.pdf. Acesso em: 1 dez. 2010.
JANSEN, Hans. Permanent access to electronic journals. **Information Services & Use**, [*S. l.*], v. 26, n. 2, p. 129-134, 2006.
KENNEY, Anne R. **Surveying the e-journal preservation landscape**. ARL 245, Apr. 2005. Disponível em: http://www.arl.org/bm~doc/arlbr245preserv.pdf. Acesso em: 10 nov. 2010.
KENNEY, Anne R. *et al***. E-journal archiving metes and bounds**: a survey of the landscape. Washington, DC: Council on Library and Information Resources, 2006.
KURAMOTO, Helio. Repositórios institucionais: políticas e mandato. *In*: SAYÃO, Luis Fernando et al. (org.). **Implantação e gestão de repositórios institucionais**: política, memória, livre acesso e preservação. Salvador: EDUFBA, 2010. p. 203-218. Disponível em: https://repositorio.ufba.br/ri/bitstream/ufba/473/3/implantacao_repositorio_web.pdf. Acesso em: 6 ago. 2021.
MARCONDES, Carlos Henrique; SAYÃO, Luis Fernando. Documentos digitais e novas formas de cooperação entre sistemas de informação em C&T. **Ciência da Informação**, Brasília, DF, v. 31, n. 3, p. 42-54, 2002. DOI: https://doi.org/10.1590/S0100-19652002000300005. Disponível em: https://www.scielo.br/j/ci/a/NKhjHgVf63bYGmkHJWQkWhB/. Acesso em: 5 ago. 2021.
MOGHADDAN, Golnessa Galyani; MOBALLEGHI, Mostafa. Trends in preserving scholarly electronic journals. *In*: INTERNATIONAL CONFERENCE ON THE FUTURE OF INFORMATION
SCIENCES, 2., 2009, Zagreb. **Proceedings** [...]. Zagreb: INFuture, 2009. p. 15-184. Disponível em: http://eprints.rclis.org/14212/1/trends_in_preserving_scholarly_electronic_journals_infuture_2009_croatia.pdf. Acesso em: 6 ago. 2021.

NESTOR WORKING GROUP ON TRUSTED REPOSITIORIES CERTIFICATION. **Catalogue of criteria for trusted digital repositories**. [*S. l.: s. n.*], 2006. Disponível em: http://edoc.hu- berlin.de/series/nestor-materialien/8en/PDF/8en.pdf. Acessado em: 23 julho 2007.

OWENS, Evans. Digital preservation and electronic journals. **Library and Information Services in Astronomy**, [*S. l.*], v. 377, 2007.

PORTICO. Digital preservation of e-journal in 2008: urgent action revisited. Princeton: Portico, 2008. Disponível em: https://www.portico.org/wp-content/uploads/2017/12/porticosurveyondigitalpreservation.pdf. Acesso em: 6 ago. 2021.

RAS, Marcel. The KB e-Depot: building and managing a safe place for journals. **Liber Quarterly**, [*S. l.*], v. 19, n. 1, p. 44-53, 2009. Disponível em: http://dspace.library.uu.nl/handle/1874/241563. Acesso em: 9 ago. 2021.

RAMESH, Ghandi *et al*. Need of digital preservation strategies, issues and challenges for future. **SRELS JournaL of Information Management**, [*S. l.*], v. 47, n. 3, 2010. DOI: 10.17821/srels/2010/v47i3/44025.

RESEARCH LIBRARY GROUP. National Archives and Records Administration. **Trustworthy repositories audit & certification**. Mountain View: RLG: OCLC, 2007. Disponível em: http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf. Acesso em: 1 dez. 2010.

RESEARCH LIBRARY GROUP. National Archives and Records Administration. **An audit checklist for the certification of trusted digital repositories**: draft for public comment. Mountain View: RLG: OCLC, 2005. http://www.rebiun.org/opencms/opencms/handle404?exporturi=/export/docReb/audit_cheklist.pdf&%5d. Acesso em: 10 nov. 2010.

RESEARCH LIBRARY GROUP. Online Computer Library Center. **Trusted digital repositories**: attributes and responsibilities. Mountain View: RLG: OCLC, 2002. Disponível em: https://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf. Acesso em: 6 ago. 2021.

SAYÃO, Luís Fernando. Preservação de revistas eletrônicas. *In*: FERREIRA, Sueli Maria Soares Pinto; TARGINO, Maria das Graças. **Mais sobre revistas científicas**: em foco a gestão. São Paulo: Editora Senac: Cengage Learning, 2008. p. 167-210.

THOMAZ, Katia. Repositórios digitais confiáveis e certificação. **Arquivistica.net**, [*S. l.*], v. 3. n. 1, p. 80-89, 2007. Disponível em https://brapci.inf.br/_repositorio/2010/05/pdf_fed0720dbb_0010726.pdf. Acesso em: 2 ago. 2021.

VENKADESAN, S. **Digital preservation of electronic resources**. [*S. l.*]:

INDEST, 2010. Disponível em https://library.iitkgp.ac.in/pages/con/INDEST/pdf/Digital%20Preservation%20of%20Electronic%20Resources-INDEST2010-IIT%20Kgp%20-%20S.%20Venkadesan.pdf. Acesso em: 8 ago. 2021.
WATERS, Donald J. **Urgent action needed to preserve scholarly electronic journals**. Virgínia: Digital Library Federation, 2005. Disponível em: https://old.diglib.org/pubs/waters051015.pdf. Acesso em: 5 ago. 2021.
Artigo Originalmente publicado em: SAYÃO, Luis Fernando. Repositórios digitais confiáveis para a preservação de periódicos eletrônicos científicos. **Ponto de Acesso**, Salvador, v. 4, n. 3, p. 68-64, 2010. DOI: https://doi.org/10.9771/1981-6766rpa.v4i3.4709. Disponível em: https://periodicos.ufba.br/index.php/revistaici/article/view/4709. Acesso em: 9 ago. 2021.

Uma outra face dos metadados: informações para a gestão da preservação digital

Luis Fernando Sayão[1]

1 Introdução

Nos dias de hoje, é virtualmente impossível discutir serviços e sistemas de informação sem o envolvimento direto com questões relacionadas aos metadados. Embora o termo "metadados" seja uma invenção relativamente recente – primordialmente ele foi usado no contexto dos sistemas de banco de dados para descrever e controlar a gestão e o uso dos dados - a ideia que ele porta remonta outros tempos, tendo suas raízes na catalogação realizada pelas bibliotecas e organizações similares (day, 2005). Essa noção é determinante, posto que, quando pensamos em metadados, a primeira idéia que nos ocorre é inspirada no seu uso no ambiente da biblioteca; no seu papel de um esquema formal para descrição de todo tipo de objetos informacionais, digitais e não digitais. A catalogação tradicional é uma forma de atribuição de metadados; o *Machine Readable Cataloging* (marc 21)[2] e o conjunto de regras usadas com ele, tais como o Código de Catalogação Anglo-Americano (AACR2)[3], são padrões de metadados (national information standard organization, 2004).

Relacionada à função de catalogação, existe outra importante razão para a criação de metadados: facilitar a descoberta de informações relevantes, seja no ambiente da biblioteca, seja no ambiente *web*. O exemplo mais ilustrativo é o *Dublin Core Metadata Element Set*[4], uma das mais importantes iniciativas na área de metadados, cujo objetivo essencial é apoiar a descoberta de recursos no extenso e fragmentado universo *web*, que apesar da sua riqueza informacional não foi pensado

1 Doutor em Ciência da Informação (ibict-ufrj), Comissão Nacional de Engenharia Nuclear (cnen), lsayao@cnen.gov.br.

2 Disponível em: https://www.loc.gov/marc/. Acesso em: 19 jul. 2021.

3 Disponível em: http://www.aacr2.org/. Acesso em: 19 jul. 2021.

4 Disponível em: http://dublincore.org/. Acesso em: 19 jul. 2021.

especificamente para a recuperação de informação. Porém, quando uma biblioteca assinala metadados descritivos para um livro de sua coleção, ela não precisa se preocupar com a possibilidade dele se dissolver numa série de páginas e figuras desconectadas caso as informações sobre a seqüência das páginas e a estrutura do livro não forem registradas; nenhum pesquisador ficará impossibilitado de avaliar o conteúdo do livro se os dados sobre a máquina *offset* que o imprimiu não forem informados. O mesmo não pode ser dito para a versão digital desse livro (LIBRARY OF CONGRESS, 2009). Quando submergimos no mundo dos documentos digitais, constatamos que outras dimensões dos metadados, que ultrapassam os limites de ferramenta para a descrição e descoberta de recursos, precisam ser reveladas e exploradas. Isto porque os objetos digitais para serem gerenciados e usados requerem processos de maior amplitude, que implica em identificar informações precisas para instruí-los adequadamente.

Na medida em que a idéia de metadados se torna uma parte essencial do mundo digital, eles se mostram conceitualmente mais complexos e mais abrangentes, apoiando um espectro extremamente amplo de atividades. Essas novas dimensões de metadados são vitais para o acesso e para a interpretação dos recursos informacionais digitais; como são importantes também para a estruturação e para os processos de gestão associados a esses recursos, que podem incluir inúmeras funções, tais como: controle dos direitos, intercâmbio, comércio eletrônico, interoperabilidade técnica e semântica, reuso da informação e curadoria digital, para citar alguns. Esse elenco crescente de funções circunscreve conceitos tradicionais e conceitos inéditos que convergem para apoiar a composição de novos ambientes informacionais, como as bibliotecas, os arquivos e os museus digitais.

Esta ampliação do domínio de aplicação faz com que os metadados necessários para a gestão e para o uso de objetos digitais sejam mais diversificados e, na maioria dos casos, diferentes dos metadados usados para gestão de coleções de obras impressas e de outros materiais físicos.

Em outro plano, o acesso e a usabilidade dos recursos informacionais digitais é impactado fortemente pela sua dependência a contextos tecnológicos específicos; esse fato gera uma área de tensão e complexidade na gestão de acervos digitais. A fragilidade estrutural da informação digital configura um dos maiores desafios a ser enfrentado pelos pesquisadores e profissionais das áreas de informação e de tantas outras áreas, neste começo de século. A preservação da informação digital por longo prazo é um problema que envolve um número grande de variáveis, planejamento cuidadoso, tecnologia e orçamentos vultosos, e cuja complexidade tem arrefecido o entusiasmo das bibliotecas digitais e demais organizações de patrimônio informacional em disponibilizar seus estoques digitais para as futuras gerações.

Entretanto, está cada vez mais claro – para a prática e para a teoria – que existe uma parte do problema de preservação digital de longo prazo que só será resolvido a partir da identificação de um conjunto de dados e informações, expressos na forma de metadados, que ancorem os processos de gestão da preservação digital.

Este elenco específico de metadados é chamado de metadados de preservação; é uma nova face para os metadados que vai assegurar que o recurso de valor contínuo sobreviva ao longo do tempo e continue sendo acessível e, não menos importante, que não perca a capacidade de ter seus significados apropriadamente interpretados no tempo que for necessário pelas comunidades para quem a informação, de forma privilegiada, se dirige.

Nessa direção, uma série de especificação de metadados e de infraestruturas físicas e conceituais vem sendo desenvolvida em torno do compromisso da preservação de longo prazo das informações digitais.

É exatamente o papel dos metadados como ferramenta voltada para instruir os processos de preservação de documentos digitais que vamos discutir resumidamente nesse trabalho. Para contextualizar o problema, começamos com uma rápida definição de metadados e seus tipos; passamos pelas estratégias de preservação digital; em seguida, discutimos os metadados de preservação, tomando como referências o modelo conceitual definido pelo *Open Archival Information System* (OAIS), o dicionário de dados do *Preservation Metadata: Implementation Strategies* (PREMIS)[5] e o papel da infraestrutura de empacotamento definida *pelo Metadata Encoding Transmission Protocol* (METS)[6].

2 Uma definição e uma categorização para metadados

Primordialmente, as iniciativas relacionadas à criação de formatos de metadados estavam focadas no desenvolvimento de padrões para organização e para a descoberta de recursos informacionais. Entretanto, novas exigências, impostas principalmente pelos desafios do mundo digital, foram redesenhando a ideia puramente descritiva de metadados, criando expansões para o seu conceito com o intuito de abrigar novos propósitos e funções.

Como desdobramento, a definição minimalista e quase clássica, que enuncia que 'metadados é dados sobre dado', torna-se inexpressiva e rasa diante da complexidade dos papeis atribuídos aos metadados nos diversos contextos correntes da gestão da informação; além do mais, ela não nos ajuda a entender o que é e como os

5 Disponível em: https://www.oclc.org/research/activities/pmwg.html. Acesso em: 19 jul. 2021.
6 Disponível em: http://www.loc.gov/standards/mets/. Acesso em: 19 jul. 2021.

metadados podem ser usados. A NISO[7] – sigla para *National Information Standard Organization* - apresenta uma definição que expande o que se entende por metadados, ampliando o seu domínio de aplicação: "Metadados é a informação estruturada que descreve, explica, localiza, ou possibilita que um recurso informacional seja fácil de recuperar, usar ou gerenciar" (NATIONAL INFORMATION STANDARD ORGANIZATION, 2004, p. 1, tradução nossa).

Não se pode afirmar que haja um consenso, mas uma fração significante dos autores que tratam do assunto concorda que os metadados podem ser divididos em três categorias conceituais: metadados descritivos, metadados estruturais e metadados administrativos. Essa segmentação é útil para uma compreensão mais clara sobre os tipos de informações que eles podem circunscrever, muito embora os seus contornos não possam ser precisamente definidos.

- **Metadados descritivos**: é a face mais conhecida dos metadados, são eles que descrevem um recurso com o propósito de descoberta e identificação; podem incluir elementos tais como título, autor, resumo, palavras-chave e identificador persistente;
- **Metadados estruturais**: são informações que documentam como os recursos complexos, compostos por vários elementos, devem ser recompostos e ordenados. Por exemplo, como as páginas de um livro, digitalizadas separadamente, são vinculadas entre si e ordenadas para formar um capítulo;
- **Metadados administrativos**: fornecem informações que apoiam os processos de gestão do ciclo de vida dos recursos informacionais. Incluem, por exemplo, informações sobre como e quando o recurso foi criado e a razão da sua criação. Nessa categoria, estão metadados técnicos que explicitam as especificidades e dependências técnicas do recurso; inclui também os metadados voltados para apoio à gestão dos direitos relacionados ao recurso.

Um requisito importante para os sistemas de informações atuais é a possibilidade da representação de recursos informacionais em níveis variados de granularidade; isso compreende a capacidade dos metadados de descreverem camadas diferenciadas de agregação dos recursos, por exemplo: descrever uma coleção, um item ou uma parte de um item, como um capítulo, uma fotografia ou um gráfico. Ainda relacionada à amplitude de resolução dos metadados, está a capacidade de descrever uma obra e suas expressões, manifestações e itens particulares (NATIONAL INFORMATION STANDARD ORGANIZATION, 2004).

7 Disponível em: http://www.niso.org/home. Acesso em: 19 jul. 2021.

Metadados são agrupados em estruturas abstratas conhecidas como esquemas ou formatos de metadados, que são conjuntos de elementos criados com fins específicos, por exemplo: descrever um tipo particular de recurso de informação. Muitos e diferentes esquemas de metadados têm sido continuamente desenvolvidos tendo como perspectiva uma grande variedade de usos em contextos variados, porém cada qual é limitado por suas especificidades e pelos seus domínios de aplicação próprios. Os poucos exemplos a seguir nos mostram um pouco dessa diversidade: *Metadata Object Description Schema* (MODS)[8] esquema bibliográfico derivado do MARC 21; *Encoded Archival Description* (EAD)[9] voltado para a área de Arquivologia; *Learning Object Metadata* (LOM)[10] para gerenciar, avaliar e localizar objetos de aprendizagem; *Multimedia Metadata* (MPEG)[11] para representação de objetos multimídiaticos.

O esquema de metadados *Dublin Core*, por sua vez, cria uma situação especial, posto que não está focado em nenhum tipo específico de objeto ou de domínio de assunto; está voltado para descoberta de recursos em domínios transversais; e é minimalista por natureza, sendo composto por poucos elementos essenciais (o *core*), passíveis de serem mapeáveis em outros formatos, constituindo a língua franca dos metadados e uma das chaves para o santo graal da interoperabilidade. Outra característica importante do *Dublin Core* é ser auto-explicativo o suficiente para permitir que o próprio autor – ou melhor, criador – da obra possa descrevê-la e publicá-la na *web*. Não obstante, o esquema possui uma estrutura simples e flexível e pode ser aplicado a recursos complexos; além do mais, pode ser representado através de sintaxes variadas, por exemplo, codificado em HTML ou em XML e estruturado segundo a arquitetura proposta pela *Resource Description Framework* (RDF)[12], facilitando o intercâmbio e o reuso.

É importante notar que os metadados podem estar embutidos num objeto digital inscrito na sua codificação, como é comum nos documentos HTML e XML ou no *header* de arquivos de imagens; ou podem estar armazenados separadamente, estruturados em bases de dados, facilitando a busca e a recuperação, como num catálogo *on-line* no ambiente de biblioteca. No mundo da *web*, os metadados precisam também ser compreendidos por computadores, por meio de robôs e agentes de software, para que possam ser recuperados e tenham sua relevância avaliada

8 Disponível em: https://www.loc.gov/standards/mods/. Acesso em: 19 jul. 2021.
9 Disponível em: http://www.loc.gov/ead/. Acesso em: 19 jul. 2021.
10 Disponível em: http://ltsc.ieee.org/wg12. Acesso em: 19 jul. 2021.
11 Disponível em: https://mpeg.chiariglione.org/standards/mpeg-7. Acesso em: 19 jul. 2021.
12 Disponível em: http://www.w3.org/RDF/. Acesso em: 19 jul. 2021.

e sejam manipulados com maior eficiência. O uso de programas para processar metadados codificados em XML é um dos pilares da iniciativa denominada *web semântica* (MARCONDES, 2005).

3 Preservação digital e o papel dos metadados

O artefato digital traz consigo uma fragilidade estrutural intrínseca que coloca permanentemente em risco a sua longevidade, tornando a preservação dos conteúdos em formatos digitais um dos desafios essenciais do nosso tempo. O problema da instabilidade das informações digitais, que nos ameaça com uma espécie de amnésia digital e uma nova pré-história, está inscrito na agenda crítica da humanidade, acompanhando outros desdobramentos negativos da tecnologia, a espera de uma solução completa e abrangente (CONSELHO NACIONAL DE ARQUIVOS, 2004).

A preservação digital, enquanto um conjunto de atividades voltadas para garantir o acesso aos conteúdos digitais por longo prazo, é, ao mesmo tempo, um desafio técnico e organizacional que se desenrola permanentemente no tempo e no espaço; seus objetivos exigem processos que portem uma intencionalidade contínua, dado que os objetos digitais não sobrevivem inercialmente, como sobrevivem as plaquetas de argila de cinco mil anos encontradas casualmente no deserto. Não existe absolutamente essa possibilidade para os objetos digitais.

As ameaças que cercam os objetos digitais são engendradas pela sua própria condição física, não fixada em suportes e fortemente dependente de contextos tecnológicos específicos e fugazes. Pela primeira vez na história, temos que preservar registros que não estão ao alcance de nenhum dos nossos sentidos, como os papiros egípcios e os pergaminhos romanos, registros cuja materialidade estruturada em átomos e moléculas está mais evidente do que os padrões virtuais - formados por *bits* e *bytes* que estabelecem a fisicalidade dos objetos digitais.

Ao contrário de uma carta ou de um livro impresso, em que a leitura e a interpretação são ações diretas e sem intermediação, entre um objeto digital e seu usuário se interpõe um ambiente tecnológico complexo e específico, formado por camadas de *software* (sistema operacional, aplicativos, etc.), *hardware*, tecnologia de redes e equipamentos especiais. "Por esta razão, não basta simplesmente preservar o objeto digital: os meios de apresentar e de usar o objeto devem também ser preservados" (LAVOIE; GARTNER, 2005, p. 6). Isso implica ter disponível, para acesso aos conteúdos e às funcionalidades do objeto digital, o ambiente correto ou, pelo menos, um substituto tecnologicamente equivalente.

Entretanto, esse ambiente tecnológico, insuflado pela inovação, competitividade e mercados em expansão, tem um ciclo de evolução continuamente mais dinâmico, tornando-se ultrapassado em lapsos de tempo cada vez menores; esse fato

coloca como imprescindível que se documente cuidadosamente o ambiente tecnológico necessário para acesso e uso dos objetos digitais arquivados.

Outra característica crítica dos objetos digitais é que eles são altamente suscetíveis a alterações (intencionais ou não) e à fragilidade das mídias, cuja gradual degradação pode levar a perdas parciais ou totais de informações. A mutabilidade dos objetos digitais tem impacto significante na fixação e na manutenção de sua aparência e da sua usabilidade; mesmo as ações de preservação podem alterar a forma e a função de um objeto digital. Essa transitoriedade dos objetos digitais torna essencial que eles estejam acompanhados de informações que documentem as suas características, sua história, incluindo todas as alterações sofridas por eles.

Por fim, é necessário considerar que operações sobre objetos digitais podem estar limitadas por cláusulas de direitos de propriedade intelectual, que podem impor limitações às ações de preservação digital, posto que, em muitos casos, elas implicam em intervenções sobre o conteúdo, funcionalidades e aparência dos objetos. Por esse motivo, é necessário documentar os direitos associados aos objetos arquivados, para que os processos de preservação estejam coordenados com as restrições impostas aos objetos (LAVOIE; GARTNER, 2005).

Desde os primeiros momentos da criação de dados e informações em meio eletrônico, já se previa que estes problemas seriam os *leviatãs* que ameaçariam o acesso persistente aos conteúdos digitais e trariam a incerteza de que a aparência, as funcionalidades, a autenticidade e a integridade desses conteúdos poderiam não ser recompostas no futuro. Entretanto, essa preocupação vem se tornando dramaticamente mais crítica, na medida em que segmentos importantes da sociedade moderna – a pesquisa científica, o governo, os negócios, a cultura e a educação - dependem mais e mais de informações digitais, na maioria das vezes produzidas por eles mesmos, como elemento essencial para todos os seus empreendimentos; e que patrimônios digitais valiosos já foram perdidos para sempre, por exemplo, parte significativa das informações sobre a exploração do planeta Marte pela sonda americana *Viking* na década de 1970 (BESSER, 2000) e as primeiras mensagens de correio eletrônico trocadas entre os cientistas na década de 1960 (LUKESH, 1999), testemunhos do início de uma época que, ironicamente, não sobreviveram à própria essência desse tempo, a transitoriedade da tecnologia.

3.1 Estratégias de preservação digital

As funções de preservação podem variar de repositório para repositório, mas geralmente circunscrevem ações que asseguram que os objetos digitais permaneçam viáveis, isto é, que possam ser lidos a partir de uma mídia; que possam ser

apresentados, ou seja, possam ser visualizados, executados ou interpretados pelo software de aplicação; e que mantenham sua integridade, significando não serem alterados inadvertidamente e que as mudanças legítimas sofridas tenham sido documentadas (CAPLAN, 2009).

As estratégias de preservação digital que estão sendo praticadas e pesquisadas pelas comunidades envolvidas com o problema de acesso, a longo prazo, a informações digitais são resumidas a seguir:[13]

- **Preservação da tecnologia** – estratégia baseada na criação de museus tecnológicos que mantêm equipamentos e software obsoletos, de forma que os documentos digitais possam ser processados no seu ambiente original. É uma solução de curto prazo;
- **Emulação** – estratégia fundamentada na premissa de que o melhor meio de preservar as funcionalidades e a aparência de um objeto informacional digital é preservá-lo junto ao seu s*oftware* original; dessa forma, o objeto pode ser rodado em plataformas atuais por meio de emuladores, que são programas que criam mímicas do comportamento de *hardware* e sistemas operacionais obsoletos em computadores novos. Essa estratégia tem sido foco de muitas pesquisas e controvérsias;
- **Migração** – tem como fundamento a migração periódica de um patamar tecnológico em vias de se tornar obsoleto e/ou de se degradar fisicamente para outro mais atualizado e íntegro, incluindo mídias, ambientes de software, formatos e computadores; é a estratégia correntemente mais utilizada pelas organizações (SAYÃO, 2007);
- **Encapsulamento** – baseia-se na idéia de que os objetos preservados devem ser autodescritos e encapsulados em estruturas físicas ou lógicas com todas as informações necessárias para que seja decifrado e compreendido no futuro.

3.2 Metadados de Preservação

Todas essas estratégias, para alcançarem seus objetivos, dependem fortemente da captura, criação e manutenção de vários tipos de dados que informem sobre histórico, características técnicas, estruturas, dependências e alterações sofridas pelo objeto digital. São esses dados que viabilizarão o pleno acesso e permitirão a recriação e a interpretação da estrutura e do conteúdo da informação digital ao

13 Para uma análise mais aprofundada, recomenda-se o estado da arte publicado por Lee e seus colaboradores em 2002, mas que permanece atual.

longo do tempo. Para tal, eles são estruturados na forma de metadados, compondo o que chamamos de "metadados de preservação".

Dessa forma, os metadados de preservação constituem uma parte essencial das estratégias de preservação digital. A síntese de sua importância pode ser expressa pelo fato deles permitirem que um objeto digital esteja autodocumentado ao longo do tempo e, portanto, posicionado para a preservação de longo prazo e para o acesso contínuo, apesar da sua propriedade, custódia, tecnologia, restrições legais, e mesmo da sua comunidade de usuários estar continuamente mudando (LAVOIE; GARTNER, 2005).

Os metadados de preservação podem ser definidos, de uma forma simples e direta, como a informação que apoia e documenta a preservação de longo prazo de materiais digitais. Entretanto, com o provável intuito de se alinhar ao consenso de que a preservação digital é um processo de gestão, alguns autores categorizam os metadados de preservação como metadados administrativos. Porém, com um grau a mais de aproximação, verificamos que os esquemas de metadados de preservação incluem elementos que se enquadram em todas as três categorias – descritivos, administrativos e estruturais. Considerando essa maior abrangência, podemos reescrever a definição de metadados de preservação mais precisamente como metadados descritivos, estruturais e administrativos que apoiam e documentam a preservação de longo prazo de materiais digitais (DAY, 2003).

Definidos dessa forma, fica claro que os metadados de preservação são criados para apoiar um grande número de funções diferentes, porém relacionadas. O amplo espectro de funções, que se espera que os metadados de preservação cumpram, sinaliza que a definição de um padrão é uma tarefa difícil e de grande amplitude; a maioria dos esquemas atualmente publicados é extremamente complexa ou somente estabelece infraestruturas básicas que precisam ainda ser implementadas para que possam ser efetivamente utilizadas. Como complicador adicional, observa-se que diferentes estratégias de preservação e diferentes tipos de informação digital exigem tipos distintos de metadados.

4 Quais são as informações necessárias para a preservação digital?

A definição dos tipos e dos contornos das informações necessárias para se instruir corretamente os processos de preservação digital foi objeto de grandes discussões num passado recente. Porém, apesar dos inúmeros pontos de tensões, os debates foram capazes de estabelecer um consenso em torno de cinco grandes categorias de informação. Essas categorias são materializadas por uma descrição aprofundada e ampla dos aspectos técnicos, custodiais e legais dos recursos digitais que devem ser traduzidos por metadados de preservação. Resumidamente, são as

seguintes: 1) *proveniência* – os metadados de preservação devem registrar informações sobre a história do objeto desde sua origem, traçando a sua cadeia de custódia e de propriedade; 2) *autenticidade* – os metadados de preservação devem incluir informações suficientes para validar que o objeto é de fato o que diz ser e que não sofreu alterações – intencionais ou não - não documentadas; 3) *atividades de preservação* – os metadados de preservação devem documentar as ações tomadas ao longo do tempo para preservar o objeto digital e as consequências dessas ações sobre aparência, usabilidade e funcionalidades do objeto; 4) *ambiente técnico* – os metadados de preservação devem descrever as dependências técnicas necessárias para a apresentação e uso dos objetos digitais, tais como *hardware*, sistema operacional e *software* de aplicação; 5) *gestão de direitos* – os metadados de preservação devem registrar todos os itens relacionados às questões de propriedade intelectual que limitem as ações de preservação, de disseminação e uso por parte de usuários de hoje e do futuro (LAVOIE; GARTNER, 2005).

Quando pensamos na estruturação das informações necessárias para preservação digital na forma de esquemas de metadados, muitos fatores devem ser levados em consideração. Entretanto, três deles são particularmente importantes, consideradas as idiossincrasia s da área: *abrangência* – o esquema deve ter uma amplitude tal, em termos de escopo e de profundidade, que considere as necessidades presentes e futuras de preservação do sistema de repositório considerado; *orientação para a implementaçã*o – o esquema deve ser projetado tendo como perspectiva os níveis práticos de implementação e a possibilidade de adaptação a sistemas automatizados voltados para gerir e assinalar metadados; *interoperáveis* – os esquemas devem ser pensados para promover e facilitar as transações entre diversos fatores que envolvam o objeto digital e os seus diversos metadados ao longo do seu ciclo de vida, por exemplo, submissão a um repositório, disseminação para um usuário ou transferência para outro repositório (LAVOIE; GARTNER, 2005).

Mas a aplicação de esquemas de metadados de preservação é uma aposta que tem como referência um cenário postulado para o futuro e torna-se um espaço amplo para incertezas e conjecturas.

Um dos principais desafios no desenvolvimento de esquemas de metadados de preservação é antecipar que informação será realmente necessária para assistir uma atividade específica de preservação digital. A extensão e a profundidade das que são exigidas para apoiar uma determinada atividade de preservação digital é função direta de algumas varáveis importantes, por exemplo: a intensidade de preservação aplicada a um dado objeto digital arquivado, ou seja, o número de características que devem ser preservadas – funcionalidades, usabilidade, aparência, autenticidade, etc.; a duração do arquivamento; a complexidade do objeto digital;

ou mesmo a base de conhecimento da comunidade a quem a informação se dirige. A decisão sobre como será aplicado o esquema define a política de preservação de um dado repositório. Por exemplo, um repositório de teses e dissertações, cujos materiais arquivados são caracteristicamente textos, tem exigências diferentes de um repositório de objetos multimídias.

Uma vez que um esquema de metadados de preservação é desenvolvido e implementado, fica difícil julgar sua efetividade *a priori,* já que uma avaliação só poderá ser realizada no futuro. Ao contrário dos metadados voltados para apoiar a descoberta de recursos, que podem ser prontamente testados e refinados para que melhorem as métricas de relevância e precisão dos resultados de busca, a adequação de um conjunto de elementos de metadados de preservação só pode ser determinada muito tempo depois da sua implementação. Só nesse momento, se pode avaliar se as informações foram excessivas ou – o que pode ser desastroso – insuficientes para garantir a preservação de longo prazo.

5 O modelo de referência OAIS – Open Archival Information System

No movimento entre teoria e prática nos espaços da preservação digital, dois pontos extremos são referenciais e significativos para o desenvolvimento de uma infraestrutura voltada para a implementação de metadados de preservação: no extremo conceitual está o OAIS *Information Model* e no prático, o PREMIS *Data Dictionary*; entre eles há um campo vasto onde várias iniciativas importantes se sucedem e se sobrepõem. Nessa seção trataremos do modelo OAIS, na seguinte, do PREMIS.

O modelo de referência OAIS é uma infraestrutura conceitual que descreve o ambiente, as interfaces externas, os componentes funcionais e os objetos de informação, associados com um sistema responsável pela preservação de longo prazo de materiais digitais. O modelo é uma tentativa de oferecer uma infraestrutura comum que pode ser usada para se compreender melhor os desafios que os repositórios precisam enfrentar; define também uma linguagem comum de alto nível que serve de instrumento para facilitar a discussão entre as diferentes comunidades interessadas no problema de preservação digital (DAY, 2004; SARAMAGO, 2004).

O OAIS foi aprovado como uma norma internacional em 2003[14], porém, antes disso, ele já era amplamente adotado por comunidades importantes na área de preservação digital que definiam seus repositórios como aderentes ao OAIS. A elaboração do Modelo foi coordenada pelo *Consultive Commitee for Space Data Systems*

14 ISO *Standard* 14721:200. No Brasil a norma foi traduzida e publicada pela ABNT como ABNT NBR 15.472:2007 - Sistema Aberto de Arquivamento de informações (SAAI).

(CCSDS)[15], vinculado a NASA[16], como parte de uma iniciativa da *Intenational Organization for Standardization*[17] (ISO) para o desenvolvimento de normas capazes de regular a preservação de longo prazo de dados originados por satélites e missões espaciais. Porém, o OAIS foi desenvolvido como um modelo genérico, aplicável a qualquer contexto de preservação digital. Nessa direção, a norma descreve um enquadramento conceitual para um repositório digital genérico, aberto, interoperável e com garantias de confiabilidade (SARAMAGO, 2004), que se autodefine "uma organização de pessoas e sistemas que aceitaram a responsabilidade de preservar a informação e torná-la disponível para uma comunidade-alvo" (CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEM, 2002, p. 1-11, tradução nossa).

Em primeiro plano, o OAIS define duas infraestruturas abstratas: *um modelo funcional* e um *modelo de informação*. O modelo funcional é compreendido como um conjunto de atividades que devem ser desempenhadas por um repositório OAIS, seja ele digital ou não; a infraestrutura funcional especificada no documento inclui admissão, armazenamento, gestão de dados, planejamento da preservação, administração e acesso. O modelo de informação define as informações, expressas por metadados, necessárias para a preservação de longo prazo e acesso aos objetos armazenados num sistema baseado no OAIS. O modelo de informação constitui uma conceitualização dos objetos de informação incorporados, armazenados e disseminados por um repositório digital orientado para a preservação (CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEM, 2002).

O OAIS define ainda o ambiente onde interagem os protagonistas envolvidos em todo o ciclo: o produtor - papel desempenhado pelas pessoas ou sistemas que fornecem a informação que deve ser preservada; a administração - papel desempenhado por quem estabelece as políticas gerais do repositório; o consumidor (usuário) -papel desempenhado por pessoas ou sistemas que interagem com os serviços do repositório com o propósito de identificar e adquirir a informação preservada que deseja; uma classe especial de consumidores, chamada de *comunidade-alvo,* é definida como o conjunto de consumidores que devem ser capazes de compreender a informação preservada. (CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEM, 2002).

Inicialmente, pode haver um estranhamento em relação aos termos adotados pelo modelo, mas no documento OAIS há uma intencionalidade óbvia em se adotar um discurso independente de áreas específicas. Isso consubstancia a ideia de um

15 Disponível em: http://public.ccsds.org/default.aspx. Acesso em: 20 jul. 2021.

16 Disponível em: http://www.nasa.gov/home/index.html. Acesso em: 20 jul. 2021.

17 Disponível em: http://www.iso.org/. Acesso em: 20 jul. 2021.

modelo genérico e de domínio amplo de aplicação que torne possível a participação de instituições não arquivísticas nos processos de preservação digital.

Não obstante a amplitude da norma que toca em vários aspectos relevantes, a questão de metadados definida no seu escopo é determinante para a área de preservação digital. "O OAIS vem exercendo uma profunda influência no desenvolvimento da arte e da ciência da preservação digital e na área de metadados de preservação é onde este impacto é especialmente evidente"(LAVOIE; GARTNER 2005, p. 9, tradução nossa).

Quando oferece uma descrição de alto-nível dos tipos de informação que fluem no espaço onde se desenrolam os processos do que chamamos de preservação digital, o OAIS torna evidente o vínculo entre metadados e preservação digital e, dessa forma, reconstrói a ideia de metadados de preservação em bases mais sólidas. Como desdobramento, o modelo de informação OAIS vem constituindo o fundamento comum para a orientação e o desenvolvimento da maioria das iniciativas de metadados de preservação surgidas nos últimos anos.

5.1 O Modelo de Informação OAIS

O modelo de informação definido no escopo do documento OAIS especifica o espectro de diferentes tipos de informação - ou metadados - exigidos para assegurar a preservação por um período indefinido de tempo, que pressupõe ainda o acesso aos conteúdos e a sua correta interpretação pelas comunidades interessadas. Os tipos de metadados, que são necessários para a preservação, são definidos como parte de uma Taxonomia de Classes de Objetos de Informação (CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEM, 2002).

O pressuposto básico do Modelo de Referência OAIS é que um recurso de informação tenha dois componentes: o objeto que precisa ser preservado e as informações que tornem o objeto compreensível para os usuários do repositório OAIS; mais formalmente, significa dizer que todo *Objeto de Informação* é composto por *Objetos de Dados* – que pode ser um objeto físico (por exemplo, uma amostra lunar) ou um objeto digital (sequências de *bits*), e por *Informação de Representação,* que permite a completa interpretação dos dados em informações com significado (CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEM, 2002).

Para um objeto digital, que é composto por uma ou mais sequências de *bits*, o propósito da Informação de Representação é converter seus *bits* em conteúdos mais expressivos aos sentidos, ou seja, em texto, em imagem, em tabela, etc. Isso é realizado através da descrição de formatos de arquivo ou de conceitos de estruturas de dados aplicado à sequência de *bits*. Pode incluir também informações adicionais necessárias para estabelecer significados particulares de um conteúdo (DAY, 2005).

Este dispositivo de reconstituição do significado da informação assume dois tipos: *informação estrutural* e *informação semântica*.

A informação estrutural inclui especificações, tais como formato dos dados, descrição do ambiente de *hardware* e de *software* em que os dados foram criados; já a informação semântica acrescenta significado à estrutura de dados identificada através da informação estrutural. Por exemplo, a informação estrutural identifica que a sequência de bits é um texto ASCII, enquanto a informação semântica indica que o texto se encontra escrito em língua inglesa (SARAMAGO, 2004).

A ideia de Objeto de Informação composta por Objeto de Dados e Informação de Representação é aplicada a todo o tipo de informação discutida no âmbito do OAIS. Isso implica na necessidade de definir estruturas lógicas que vinculem o conteúdo a ser preservado à diversidade de metadados que apoiarão a gestão da sua preservação. Decorre daí a ideia de *pacote de informação*.

No ambiente de um repositório aderente à norma OAIS, os fluxos de informação se realizam por meio de unidades discretas chamadas *Pacotes de Informação* - contêineres que encapsulam logicamente os conteúdos, objeto da preservação e os metadados associados a eles (CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEM, 2002). Esse é um conceito-chave subjacente a todos os processsos que se desenrolam no âmbito do modelo OAIS.

A norma define três tipos de pacotes de informação: *pacote de informação de submissão*[18], formado pelo conteúdo e metadados que são submetidos pela entidade externa, *Produtor,* ao repositório no momento do depósito; *pacote de informação de armazenamento*[19], formado pelo conteúdo e pelos metadados que são efetivamente armazenados e gerenciados pelo repositório por longo prazo; o *pacote de informação de disseminação*[20], que é o conteúdo e os metadados entregues pelo repositório em resposta a uma requisição de acesso demandada pelo usuário, ou melhor, pelo *Consumidor*.

Deve ficar claro que o pacote de informação de armazenamento é o pacote destinado à preservação de longo prazo; ele é um contêiner que agrega quatro tipos de objetos de informação que circunscrevem os vários tipos de informações necessárias para a preservação de longo prazo (CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEM, 2002), ou seja:

- *Informação de conteúdo* – é a informação que o repositório tem obrigação de preservar, inclui *a informação de representação,* que são informações ne-

18 Do inglês *Submission Information Package* (SIP).

19 Do inglês *Archival Information Package* (AIP).

20 Do inglês *Dissemination Information Package* (DIP).

cessárias à apresentação e à interpretação da cadeia de bits que constituem o objeto armazenado como informação com significado para uma determinada comunidade alvo;
- *Informação de descrição de preservação* - informação que apoia e documenta a preservação dos objetos arquivados no repositório;
- *Informação de empacotamento* - informação que agrega todos os componentes de um pacote de informação - conteúdo e seus metadados - numa única unidade lógica;
- *Informação descritiva* - informação que apoia o usuário na descoberta e na recuperação de objetos armazenados no repositório.

A *informação de descrição de preservação*, identificada pelo OAIS pela sigla PDI[21], é o tipo de informação que nos interessa nesse momento. O PDI está:

> especificamente focado na descrição do estado, tanto passado quanto presente, da Informação de Conteúdo, assegurando que ela está univocamente identificada e que não sofreu alterações não documentadas" (CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEM, 2002, p. 4-27, tradução nossa).

A Taxonomia de Classes de Objeto de Informação do OAIS detalha a *informação de descrição de preservação* em quatro grupos distintos de dados, definidos como se segue (CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEM, 2002):

- *Informação de referência* – tem origem na necessidade de identificar e de localizar um objeto ao longo do tempo para manter a sua integridade; a referência identifica ou, se necessário, descreve um ou mais mecanismos usados para assinalar identificadores aos objetos armazenados, de forma que eles possam ser identificados inequivocamente interna e externamente ao repositório. Por exemplo, um identificador local (um número de chamada) e um DOI[22] ou um ISBN[23]; pode incluir ainda informações que descrevem o objeto, por exemplo, um resumo;
- *Informações de contexto* – está relacionado ao fato de que muitos objetos não podem ser adequadamente interpretados sem a compreensão do seu con-

21 Sigla para *Preservation Description Information*.
22 *Digital Object Identifier* - Disponível em: http://www.doi.org/. Acesso em: 20 jul. 2021.
23 *International Standard Book Number* - Disponível em: http://www.bn.br/portal/?nu_pagina=26. Acesso em: 20 jul. 2021.

texto; informação que documenta o relacionamento do objeto armazenado e seu ambiente; isso inclui a motivação da criação do objeto e como ela se relaciona com outros conteúdos; circunscreve as dependências técnicas – *hardware*, *software*, *linkage,* etc. - inclui ainda modo de distribuição, por exemplo, via rede;

- *Informação de proveniência* – refere-se ao principio de que parte da integridade de um objeto depende da sua história; informação que documenta a história do objeto armazenado; pode incluir informações sobre sua fonte ou origem, sua cadeia de custódia; registra também as ações de preservação sofridas pelo objeto e seus efeitos, por exemplo: as migrações efetuadas;
- *Informação de fixidade* – refere-se a qualquer informação que documenta mecanismos particulares de autenticação usados para assegurar que o objeto armazenado não sofreu nenhuma alteração não documentada, e que sua integridade não foi comprometida, por exemplo, assinaturas digitais e *checksums*.

Esses grupos de informação – que formam as bases das principais estruturas de metadados de preservação - são baseados em categorias definidas pelas discussões apresentadas em 1996 no relatório *Task Force on Archiving of Digital Information* comissionados pela *Commission on Preservation and Access* (CPA) e pela *Research Library Group*[24] (RLG), que registra textualmente que:

> no ambiente digital, as características que determinam a integridade da informação e merecem uma atenção especial para propósitos de arquivamento incluem: conteúdo, fixidade, referência, proveniência e contexto. (COMMISSION ON PRESERVATION AND ACCESS, 1996, p. 12, tradução nossa).

Os tipos de informações explicitadas pela taxonomia presente no modelo de informação OAIS podem ser interpretados como a descrição mais geral de metadados necessários para instruir a preservação de longo prazo e o uso de materiais digitais. Essas informações estabelecem um ponto de partida para a maioria dos esforços subsequentes em desenvolver esquemas formais de metadados.

5.2 Aplicações do Modelo de Informação OAIS

Enquanto um modelo de referência, o OAIS não toca nos níveis de implementação (CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEM, 2002); cada comunida-

24 Disponível em: http://www.oclc.org/programs/about/default.htm. Acesso em: 20 jul. 2021.

de interessada deve aplicar o modelo – incluindo o modelo de informação - no seu contexto técnico e organizacional, adequando-o as suas especificidades e objetivos. Ainda no seu papel de uma descrição de alto nível, a norma não transmite pressupostos sobre os tipos de recursos digitais manuseados pelo repositório e nem acerca das especificações tecnológicas adotadas por ele para cumprir os seus objetivos de preservação e acesso de longo prazo (SARAMAGO, 2004).

Entretanto, a demanda por desenvolvimento de soluções operacionais está refletida na longa lista de instituições envolvidas na criação de conjuntos de elementos de metadados para apoiar a preservação digital. Michael Day (2003), nos informa que a maioria dessas implementações surge em três contextos distintos que, porém, possuem o interesse comum pela preservação digital: bibliotecas nacionais e de pesquisa, projetos de digitalização e arquivos. Algumas dessas iniciativas, originadas por instituições internacionais de maior renome, são projetos com desdobramentos significativos para a área. Vamos nos ater, neste momento, às iniciativas que têm em comum uma fundamentação - embora em graus variados - inspirada no modelo OAIS.

Uma das primeiras respostas práticas ao desafio foi dada pelo *National Library of Autralia* (NLA)[25], tendo como ambiente o repositório de publicações eletrônicas PANDORA[26], sigla para *Preserving and Acessing Networked Documentary Resources of Australia*; logo após, a minuta de outro conjunto de elementos foi publicada no Reino Unido, no âmbito do projeto CEDARS[27] (CURL *Exemplars in Digital Archives*); o projeto *Networked European Deposit Library* (NEDLIB)[28] desenvolveu um sistema de depósito para bibliotecas eletrônicas e tentou definir, nesse contexto, um conjunto mínimo de metadados que seria necessário para apoiar a gestão da preservação.

Lavoie e Gartner (2005) observam que esses primeiros esforços resumem uma tendência por natureza altamente especulativa, dado que procuravam antecipar os elementos de metadados de preservação necessários para sustentar as iniciativas programáticas de preservação digital que iriam emergir no futuro. Não havia consenso sobre questões básicas, tais como que tipos de informações seriam necessárias e como elas poderiam ser usadas para apoiar os processos de preservação digital. Por outro lado, os projetos importantes que se seguiram – por exemplo, os

25 Disponível em: http://www.nla.gov.au/. Acesso em: 20 jul. 2021.
26 Disponível em: http://pandora.nla.gov.au/. Acesso em: 20 jul. 2021.
27 Disponível em: http://www.rluk.ac.uk/projects. Acesso em: 20 jul. 2021.
28 Disponível em: http://nedlib.kb.nl/. Acesso em: 20 jul. 2021.

conjuntos de elementos produzidos pela *Online Computer Library Center* (OCLC)[29], pela *National Library of New Zealand*[30] e pela *University of Edinburg*[31] - estavam mais proximamente alinhados com o planejamento e a implementação de sistemas de repositórios digitais e se beneficiaram amplamente da fundamentação estabelecida pelos primeiros conjuntos de elementos.

Na trajetória que se delineava, tornava-se imperativo, para a área de preservação digital, harmonizar os três esquemas referenciados acima – NLA, CEDARS e NEDLIB - em uma infraestrutura única. Nessa direção, por volta do ano 2000, a OCLC e o *Research Library Group* (RLG) convocaram um grupo de trabalho internacional, que reunia expertise de vários domínios e organizações, para endereçar os novos desenvolvimentos na área. O grupo produziu dois relatórios que constituíram documentos determinantes para o avanço na direção de uma efetiva implementação fundamentada no OAIS, são eles: "*Preservation metadata for digital object: a review of the state of the art*" (ONLINE COMPUTER LIBRARY CENTER, 2001) e o "*Preservation metadata and the* OAIS *Information Model: a metadata framework to support the preservation of digital object*" (ONLINE COMPUTER LIBRARY CENTER, 2002).

O primeiro documento – um livro branco - sintetizava o estado da arte em metadados de preservação digital, oferecia uma definição para eles, descrevia os papéis dessa classe de metadados no processo de preservação, ao mesmo tempo em que revisava as iniciativas existentes, identificando convergências e divergências. A tarefa seguinte estava fundamentada sobre os alicerces consolidados por este livro branco e tinha como resultado o desenvolvimento de uma infraestrutura de metadados de preservação abrangente e de larga aplicação baseada nas categorias de informação especificadas no modelo de informação do OAIS.

> A infraestrutura produzida pelo grupo de trabalho efetivamente substituiu o conjunto de elementos desenvolvido pelas iniciativas anteriores e representaram um ponto de partida importante para a futura implementação prática de metadados de preservação. (DAY, 2003, p. 5, tradução nossa).

Ao mesmo tempo em que deixava óbvio que a colaboração e o consenso formam a pedra de toque para superar os desafios e as incertezas da preservação

29 Disponível em: http://www.dpconline.org/docs/reports/dpctw05-01.pdf. Acesso em: 20 jul. 2021.
30 Disponível em: http://www.natlib.govt.nz/catalogues/library-documents/preservation-metadata-revised. Acesso em: 20 jul. 2021.
31 Disponível em: http://www.lib.ed.ac.uk/sites/digpres/metadataschema.shtml. Acesso em: 20 jul. 2021.

digital. Entretanto, ainda era necessário um esforço considerável antes que fosse possível implementar operacionalmente esquemas de metadados de preservação para repositórios particulares, posto que algumas questões importantes sobre os metadados e seus usos ainda precisavam ser respondidas. Por exemplo: de todas as informações cobertas pela infraestrutura, qual é o subconjunto de informações essenciais para preservação de longo prazo? Como essas informações podem ser traduzidas em elementos implementáveis de metadados de preservação? Como os metadados de preservação podem ser criados e mantidos no âmbito operacional de um sistema de arquivamento digital? (LAVOIE; GARTNER, 2005).

Para responder questões como essas, a OCLC e a RLG patrocinaram, logo em seguida, um novo grupo de trabalho chamado PREMIS – sigla para *Preservation Metadata: Implementation Strategies*[32] - com o objetivo de detalhar os aspectos práticos de implementação dos metadados de preservação no contexto de sistemas de preservação digital. É sobre isso que discutiremos brevemente a seguir.

6 PREMIS: o modelo OAIS em ação

O objetivo subjacente à idéia de constituir o Grupo de Trabalho PREMIS era delinear uma ferramenta concreta, uma ponte, que pudesse superar o abismo entre a teoria e a prática na área de metadados de preservação digital; o que também pode ser traduzido por colocar em ação os conceitos preconizados pela infraestrutura de alto nível fixada pelo Modelo de Informação do OAIS. Nessa direção, o PREMIS se estabeleceu tendo como base o consenso extraído das experiências acumuladas de muitas e variadas instituições – museus, bibliotecas, arquivos, governo e iniciativa privada – e a *expertise* dos principais profissionais da área, provenientes da Austrália, Nova Zelândia, Estados Unidos, Grã-Bretanha, Holanda e Alemanha. O empreendimento foi inicialmente planejado para um ano, porém se desdobrou por mais outro. Os resultados, entretanto, compensaram o alongamento dos prazos: o Grupo de Trabalho desenvolveu um conjunto de elementos de metadados altamente refinados, que potencialmente servia de fundamento para possíveis implementações (MCCALUN, 2005).

O esforço considerava vários objetivos relacionados. Porém, o interesse do Grupo de Trabalho convergia de forma contundente para dois pontos que sintetizavam o que se esperava do OAIS, enquanto uma fundamentação para a prática da preservação digital, para o intercâmbio de informações de preservação e para a interoperabilidade entre repositórios. Esses pontos eram os seguintes:

32 Disponível em: http://www.loc.gov/standards/premis/. Acesso em: 20 jul. 2021.

- Tomando como ponto de partida a infraestrutura delineada anteriormente, definir um conjunto essencial de elementos de metadados de preservação que seja implementável e de larga aplicação; esse núcleo essencial de metadados deve ser apoiado por um dicionário de dados, que será desenvolvido para oferecer diretrizes e recomendações para o preenchimento e para a gestão dos elementos de metadados;
- Identificar e avaliar estratégias alternativas para codificar, armazenar, gerenciar e intercambiar metadados de preservação, especialmente os essenciais, no contexto de um sistema de repositório digital.

O trabalho do Grupo começou pelo levantamento dos projetos de repositórios digitais em operação e ainda os planejados, tendo como objetivo identificar as práticas correntes e as tendências para projetos digitais. Dentre os vários aspectos endereçados – missão, comunidade de usuários, serviços, fundos de financiamento, gestão de direitos e conteúdos – estavam, naturalmente, interrogações sobre como os metadados estavam sendo usados para apoiar os processos, as funções e as políticas do repositório. (LAVOIE; GARTNER, 2005). O levantamento obteve 48 respostas, originadas principalmente por bibliotecas, arquivos e museus provenientes de 13 países diferentes. Os resultados do *survey* foram sumarizados no relatório "*Implementing preservation repositories for digital materials: current practice and emerging trends in the cultural heritage*" (ONLINE COMPUTER LIBRARY CENTER, 2004).

Contudo, a principal materialização do trabalho do PREMIS foi o relatório de 237 páginas lançado em maio de 2005, intitulado "*Data dictionary for preservation metadata: final report of the* PREMIS *Work Group*" (ONLINE COMPUTER LIBRARY CENTER, 2005).

6.1 Dicionário de dados PREMIS

O coração e a alma deste relatório é o PREMIS *Data Dictionary*, traduzido aqui por Dicionário de Dados PREMIS. Trata-se de um guia abrangente que define um conjunto de metadados necessários para apoiar a preservação digital de longo prazo.

O Dicionário de Dados não tem como objetivo definir todos os elementos possíveis de metadados de preservação, verdadeiramente ele se concentra no núcleo básico de elementos que a maioria dos repositórios precisa compreender para apoiar a preservação de longo prazo; esse núcleo é chamado de *metadados essenciais*.

O relatório inclui complementarmente vários outros textos e ferramentas: os "tópicos especiais" que discutem aspectos relacionados ao Dicionário de Dados; um glossário; e um conjunto de exemplos que ilustram o uso do Dicionário de Dados para vários materiais em diferentes contextos de preservação digital. O Grupo

de Trabalho desenvolveu também um conjunto de esquemas XML[33] para apoiar o uso do Dicionário de Dados por instituições que gerenciam e intercambiam metadados de preservação que estejam em conformidade com a proposta do PREMIS.

Rigorosamente, o Dicionário de Dados não define elementos de metadados e sim unidades semânticas. Essa diferença é sutil, porém importante: uma unidade semântica é uma peça de informação ou de conhecimento, enquanto um elemento de metadados é uma forma definida de representar essa informação em um registro de metadados, em um esquema ou numa base de dados. Nessa direção, o PREMIS não especifica como os metadados devem ser representados em um sistema, ele simplesmente define o que o sistema precisa entender e o que ele deve ser capaz de exportar para outros sistemas (CAPLAN, 2009).

O Dicionário de Dados está organizado em torno de um modelo de dados (Figura 1) que relaciona cinco entidades que têm papeis associadas com a preservação digital, são elas: *Entidade Intelectual, Objeto, Evento, Agente* e *Direitos*. O PREMIS as define da seguinte forma:

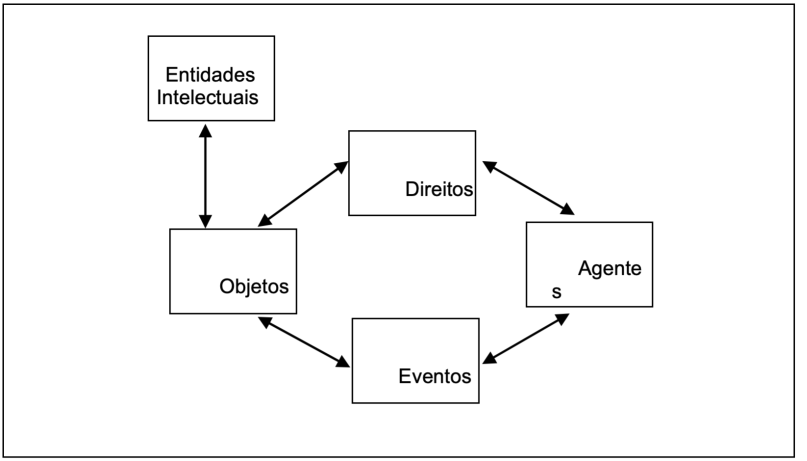
- *Entidade intelectual* – um conjunto coerente de conteúdos que é reconhecido como uma unidade, por exemplo, livros, artigos, bases de dados;
- *Objeto* – uma unidade discreta de informação em forma digital, constituindo o que realmente é armazenado e gerenciado pelo repositório, por exemplo, um arquivo PDF. As unidades semânticas para Objetos podem ser especificadas em três níveis: cadeia de bits (*bitstream*), arquivos (*files*) e o conjunto de arquivos que completam a apresentação de uma Entidade Intelectual, ou seja, a representação (*representation*);
- *Evento* – são ações que envolvem ou afetam os objetos no repositório, por exemplo, uma ação de migração;
- *Agente* – é uma pessoa, organização ou programa de computador que desempenha papéis associado com um Evento ou declarações de Direitos;
- *Direitos* – são direitos e permissões vinculadas ao Objeto relevantes para a preservação, por exemplo, permissão para se fazer uma cópia em PDF.

O Dicionário de Dados oferece uma descrição detalhada dos metadados associados com cada uma das entidades, entretanto os metadados para Entidades Intelectuais são considerados fora do escopo do PREMIS, dado que a estas informações já são supridas pelos esquemas focados em metadados descritivos (CAPLAN, 2009; LAVOIE; GARTNER, 2005). Por exemplo, MARC para materiais bibliográficos.

33 Disponível em: http://www.loc.gov/standards/premis/schemas.html. Acesso em: 20 jul. 2021.

Intencionalmente, o Grupo de Trabalho PREMIS não tratou de alguns aspectos bem conhecidos da preservação digital, tal como o detalhamento dos metadados técnicos para diferentes mídias; somente os metadados técnicos que são geralmente aplicados transversalmente a formatos de arquivos foram trabalhados pelo Grupo. Outra importante consideração adotada pelo PREMIS é que os metadados especificados devem ser, tanto quanto possível, assinalados e usados automaticamente. Isso leva preferencialmente para a escolha de valores extraídos de listas contendo formas padronizadas, ao invés de descrição textual (MCCALLUM, 2005).

Figura 1 - Modelo de Dados do PREMIS



Fonte: Elaborado pelo autor.

Lavoie e Gartner (2005, p. 14) observam que "há ainda muito trabalho a ser feito, especialmente em termos de testar o Dicionário de Dados em diferentes domínios e contextos de preservação digital"; eles concluem refletindo que no futuro, a ampla adoção do Dicionário de Dados pode ajudar no estabelecimento de práticas padronizadas voltadas para a gestão de metadados de preservação que enfatizem a interoperabilidade de repositórios digitais distribuídos em redes. A adoção de padrões pode ainda gerar uma economia potencial possibilitada pela prática de compartilhar e reusar determinadas formas de metadados de preservação entre repositórios digitais.

Nessa direção, o PREMIS *Maintenance Activity* desenvolveu um esquema XML que corresponde diretamente ao Dicionário de Dados, viabilizando que o PREMIS seja usado para intercâmbio de metadados representado em XML.

7 METS: empacotando os metadados de um objeto digital

Não há dúvidas de que um objeto digital vai acumulando uma quantidade crescente de metadados de todo o tipo ao longo de tempo; somadas aos metadados de preservação muitas outras formas são incorporadas, tais como metadados para a descoberta de recursos e administrativos. Este fato nos coloca diante de uma questão crítica: como todos esses metadados podem estar organizados e vinculados ao objeto correspondente? Algumas soluções foram propostas na forma de infraestruturas para empacotamento de metadados, dentre elas estão o MPEG-21[34] e o METS, sigla para *Metadata Encoding Transmission Standard*. No contexto que nos interessa, o mais importante é a norma METS, posto que ela foi projetada por iniciativa da *Digital Library Federation* (DLF) para implementar os pacotes de informação referenciados pelo Modelo de Referência OAIS (LAVOIE, 2004).

O METS é um esquema XML que oferece um mecanismo flexível para codificar todos os tipos de metadados associados a um objeto digital – descritivos, administrativos, estruturais - e para exprimir as ligações complexas entre esses metadados no ambiente de um repositório. Por conseguinte, o METS estabelece um padrão útil para a gestão de objetos digitais no âmbito de um repositório e o intercâmbio deles entre repositórios (ou entre repositórios e seus usuários); além do mais, oferece a possibilidade de associar um objeto digital com comportamentos ou serviços. Dessa forma, um documento METS pode ser usado para estruturar Pacotes de Informação de Submissão, Pacotes de Informação de Arquivamento e Pacotes de Informação de Disseminação, que é a forma como as informações são gerenciadas e fluem no contexto do Modelo de Referência OAIS (LIBRARY OF CONGRESS, 2009).

Um documento METS compreende cinco principais seções:

- Grupo de arquivos – é um inventário de todos os arquivos associados com o objeto digital e de suas versões eletrônicas;
- Metadados Administrativos – essa seção aninha as informações técnicas sobre: como os arquivos foram criados e armazenados, a gestão de direitos, o objeto original da qual o objeto deriva e a proveniência dos arquivos que compõem o objeto. Pode apontar para metadados externos ao documento METS;
- Metadados Descritivos – essa seção inclui informações sobre o conteúdo intelectual do item – incluindo informações bibliográficas - necessárias para a sua recuperação e avaliação por parte do usuário. Essa seção pode apon-

34 Disponível em: http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm. Acesso em: 20 jul. 2021.

tar para metadados externos ao documento METS, por exemplo, um registro MARC num catálogo *on-line* (OPAC);
- Mapa Estrutural – indica de forma hierárquica como os vários componentes do item se relacionam mutuamente, permitindo, dessa forma, que seus elementos constituintes possam ser navegados pelos usuários;
- Comportamento – essa seção pode ser usada para associar comportamentos executáveis com o conteúdo no objeto METS.

De muitas formas, o METS representa uma solução que se enquadra nas exigências de estabilidade da preservação digital. Em primeiro lugar, um documento METS está escrito em XML, que há muito tem sido consensualmente reconhecido por todos os domínios como uma forma robusta e legível para o arquivamento de metadados; depois, enquanto uma linguagem não proprietária, o XML pode assegurar que a informação, por ele codificada, não será dependente de nenhum pacote específico de *software* e, portanto, não sofrerá – ou sofrerá menos - as consequências da obsolescência tecnológica que ameaça as aplicações vinculadas a programas. Portanto, os metadados arquivados em dispositivos XML, tal como o padrão METS, deverão estar prontos para uso pelos mecanismos futuros de disseminação e de intercâmbio com outros repositórios (LAVOIE; GARTNER, 2005).

8 À guisa de conclusão

Os metadados têm um papel de fundamental importância na organização e no acesso às informações nos sistemas tradicionais, como nas coleções de livros de uma biblioteca ou nos ambientes informacionais baseados em redes de computadores, como é a própria *web*. Entretanto, o conceito de metadado pode ser expandido para apoiar a gestão de objetos digitais, cujo escopo inclui os processos de preservação digital de longo prazo.

Progressivamente, essa idéia foi se consolidando. Hoje, há um consenso absoluto de que os conteúdos digitais que precisam ser acessados e compreendidos no futuro devem estar acompanhados de dados e informações, expressos na forma de metadados, que tornem viável a sua acessibilidade, integridade e autenticidade.

Nessa direção, iniciando-se na década de 1990, inúmeros projetos e iniciativas vêm enfrentando o desafio de dimensionar o papel dos metadados no apoio às atividades de preservação digital e de identificar quais são as informações necessárias para tal. Esses esforços têm como característica comum o desenvolvimento baseado no consenso e na cooperação.

A universalidade do problema da fragilidade da informação digital, bem como a convergência de interesses das diversas instituições de patrimônio digital – bi-

bliotecas, museus e arquivos – falam a favor da colaboração e da construção do consenso para resolver os desafios e as incertezas de gerenciar materiais digitais por longo prazo. Numa trajetória evolutiva, diretrizes, padrões, práticas e experiências em implementação estão emergindo e se consolidando baseados em modelos conceituais concebidos num passado recente. O PREMIS, considerada a iniciativa mais importante em metadados de preservação, é uma síntese de tudo isso. Baseado nas experiências acumuladas por muitas instituições, na transversalidade de vários domínios e consolidado pelo consenso, ele representa um passo importante na superação do hiato existente entre a teoria e a prática no domínio da preservação digital.

Por fim, é necessário enfatizar que a relevância e a complexidade do problema da preservação digital podem ser mensuradas pela dependência quase total de dados e de informações digitais de alguns segmentos importantes da sociedade, por exemplo, educação, governo, negócios, pesquisa científica e expressão artística; isso sem falar nas mensagens para o futuro, que são críticas para a sobrevivência da humanidade, por exemplo, a localização de depósitos de materiais tóxicos. Essa dependência dramática se reflete na urgência pela busca de soluções abrangentes que sejam tecnológicas, econômicas, éticas e legalmente viáveis.

Tudo isso somado transforma a área de pesquisa e da prática em preservação digital um espaço pleno de desafios instigantes para muitos domínios do conhecimento.

Referências

BESSER, HOWARD. Digital longevity. *In*: SITTS, Maxine (org.). **Handbook for digital projects:** a management tool for preservation and access. Andover, MA: Northeast Document Conservation Center, 2000. p. 155-166. Disponível em: http://besser.tsoa.nyu.edu/howard/Papers/sfs-longevity.html. Acesso em: 21 jul. 2021.

CAPLAN, Priscilla. **Understanding PREMIS**. Washington, DC: Library of Congress, 2009. Disponível em: https://www.loc.gov/standards/premis/understanding-premis.pdf. Acesso em: 20 jul. 2021.

COMMISSION ON PRESERVATION AND ACCESS. Research Libraries Group. **Preserving digital information:** report of the task force on archiving of digital information, [*S. l.*]: CPA : RLG, 1996. 71p. Disponível em: https://www.clir.org/wp-content/uploads/sites/6/pub63watersgarrett.pdf. Acesso em: 20 jul. 2021.

CONSELHO NACIONAL DE ARQUIVOS. **Carta para a preservação do patrimônio arquivísticos digital**: preservar para garantir o acesso. Rio de Janeiro: CONARQ, 2004. Disponível em: http://www.unesp.br/ccad/mostra_arq_multi.

php?arquivo=6962. Acesso em: 20 jul. 2021.
CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEM. **Reference model for an open archival information system (OAIS)**. Blue book (CCSDS 650.0-B-1). Washington, DC: CCSDS, 2002. Disponível em: http://public.ccsds.org/publications/archive/650x0b1.pdf. Acesso em: 20 jul. 2021.
DAY, Michael. **DCC digital curation manual**: installment on metadata. Bath: University of Bath, 2005. Disponível em: http://hdl.handle.net/1842/3321. Acesso em: 20 jul. 2021.
DAY, Michael. **Preservation metadata.** Bath: UKOLN, University of Bath, 2003. Disponível em: http://www.ukoln.ac.uk/metadata/publications/iylim-2003/. Acesso em: 20 jul. 2021.
DAY, Michael. Preservation metadata initiatives: practicality, sustainability, and interoperability. *In*: BISCHOFF, Frank; HOFMAN, Hans; ROSS, Seamus. (org.). **Metadata in preservation**: selected papers ERPANET Seminar at the Archives School Marburg. Marburg: Archivschule Marburg, 2004. p. 91-117. Disponível em: http://opus.bath.ac.uk/14365/1/day-marburg-paper.pdf. Acesso em: 20 jul. 2021.
LAVOIE, Brian; GARTNER, Richard. **Preservation metadata**. [*S. l.*]: Online Computer Library Center, 2005. Disponível em: http://www.dpconline.org/component/docman/doc_download /88-preservation-metadata. Acesso em: 30 set. 2009.
LAVOIE, Brian. Implementing metadata in digital preservation systems: the PREMIS activity. **D-Lib Magazine**, v. 10, n. 4, 2004. Disponível em: http://www.dlib.org/dlib/april04/lavoie/04lavoie.html. Acesso em: 21 jul. 2021.
LEE, Kyong-Ho *et al*. The state of the art and practice in digital preservation. **Journal of Research of the National Institute of Standards and Technology**, [*S. l.*], v. 107, n. 1, p. 93-106, 2002. DOI: 10.6028/jres.107.010. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4865277/. Acesso em: 21 jul. 2021.
LIBRARY OF CONGRESS. **METS**: an overview & tutorial. Whashington, DC: Library of Congress, 2009. Disponível em: http://www.loc.gov/standards/mets/METSOverview.v2.html. Acesso em: 21 jul. 2021.
LUKESH, Susan. E-mail and potential loss to future archives and scholarship or the dog that didn't bark. **First Monday**, Chicago, v. 4, n. 9, 1999. DOI: https://doi.org/10.5210/fm.v4i9.692. Disponível em: https://firstmonday.org/ojs/index.php/fm/article/view/692. Acesso em: 21 jul. 2021.
MARCONDES, Carlos Henrique. Metadados: descrição e recuperação de informação na web. *In*: MARCONDES, Carlos Henrique; KURAMOTO, Hélio; TOUTAIN, Lídia Brandão; SAYÃO, Luis Fernando (org.). **Bibliotecas digitais**: saberes e práticas. Salvador: Ed. UFBA; Brasília, DF: IBICT, 2005. p. 97-114.

Disponível em: http:/livroaberto.ibict.br/handle/1/1013. Acesso em: 21 jul. 2021.
MCCALLUM, Sally. Preservation metadata: what we have and what we need. *In*: WORLD LIBRARY AND INFORMATION CONGRESS, 71., 2005, Oslo. **Proceedings** [...]. Oslo: IFLA, 2005. p. 1-8. Disponível em: https://archive.ifla.org/IV/ifla71/papers/060e-McCallum.pdf. Acesso em: 21 jul. 2021.
NATIONAL INFORMATION STANDARD ORGANIZATION. **Understanding metadata.** Bethesda, MD: NISO Press, 2004. Disponível em: http://www.niso.org/publications/press/UnderstandingMetadata.pdf. Acesso em: 21 jul. 2021.
ONLINE COMPUTER LIBRARY CENTER. Research Library Group. **Data dictionary for preservation metadata:** final report of the PREMIS Working Group. Dublin: OCLC; Ohio: RLG, 2005.
ONLINE COMPUTER LIBRARY CENTER. Research Library Group. **Implementing preservation repositories for digital materials**: current practice and emerging trends in the cultural heritage - A Report by the PREMIS Working Group. September. Dublin: OCLC; Ohio: RLG, 2004.
ONLINE COMPUTER LIBRARY CENTER. Research Library Group. **Preservation metadata and the oais information model**: a metadata framework to support the preservation of digital object. Dublin: OCLC; Ohio: RLG, 2002.
ONLINE COMPUTER LIBRARY CENTER. Research Library Group. **Preservation metadata for digital objects**: a review of the state of the art. Dublin: OCLC; Ohio: RLG, 2001.
SARAMAGO, Maria de Lurdes. Metadados para a preservação digital e aplicação do Modelo OAIS. *In*: CONGRESSO NACIONAL DE BIBLIOTECARIOS, ARQUIVISTAS E DOCUMENTALISTAS, 8., 2004, Estoril. **Anais** [...]. Estoril: ACTAS, 2004. p. 1-6. Disponível em: https://www.bad.pt/publicacoes/index.php/congressosbad/article/view/640. Acesso em: 30 set. 2009.
SAYÃO, Luis Fernando. Conservação de documentos eletrônicos. *In*: GRANATO, Marcus; SANTOS, Claudia; ROCHA, Claudia. **Conservação de acervos**. Rio de Janeiro: MAST, 2007. p. 181-204.

Artigo Originalmente publicado em: SAYÃO, Luis Fernando Uma outra face dos metadados: informações para a gestão da preservação digital. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, Florianópolis, v. 15, n. 30, p. 1-31, 2010. DOI: 10.5007/1518-2924.2010v15n30p1. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2010v15n30p1. Acesso em: 21 jul. 2021.

Digitalização de acervos culturais: reuso, curadoria e preservação[1]

Luis Fernando Sayão[2]

1 Introdução

Existe uma quantidade crescente de recursos informacionais sendo criada em formatos digitais através de processos de digitalização de informação analógica já existente e da geração de conteúdos de gênese digital. Registros desse tipo são criados e aplicados em todo espectro social, mudando comportamento, negócios, formas de governar, de ensinar e inaugurando padrões inéditos de socialização e de metodologias de produção de conhecimento científico. As instituições de patrimônio cultural não são exceções: elas geram e consomem ativos digitais, tendo como ponto de partida grandes programas – muitas vezes de âmbito nacional e internacional – de digitalização de seus acervos físicos.

No domínio das instituições culturais, parece que "a era digital é também a era da digitalização". Este fenômeno não acontece somente para documentos e imagens, mas também para recursos sonoros e visuais, para objetos tridimensionais, para artes performáticas, figurinos folclóricos e para monumentos e passagens; e ainda para a herança cultural intangível tal como memória oral e tradições locais. Nenhum aspecto é esquecido nessa transição para o mundo digital (bachi et al., 2014), que conta com o poder amplificador e pervasivo da web como meço de disseminação.

A disponibilidade em larga escala de informações digitais, alçada à crescente oferta e o uso desses ativos informacionais por meio de serviços online de naturezas distintas – aplicativos para celulares, bases de dados, intercâmbio de informa-

1 Artigo originalmente publicado em: sayão, Luis Fernando. Digitalização de acervos culturais: reuso, curadoria e preservação. *In*: seminário serviços de informação em museus, 4., 2016, São Paulo. **Anais** [...]. São Paulo: Sesc Bom Retiro, 2016. Disponível em: http://biblioteca.pinacoteca.org.br:9090/bases/biblioteca/322697.pdf. Acesso em: 16 ago. 2021.

2 Doutor em Ciência da Informação (ibict-ufrj), Comissão Nacional de Engenharia Nuclear (cnen), lsayao@cnen.gov.br.

ções, etc. –, tem ocasionado o aumento da expectativa em torno de serviços digitais que potencialmente podem ser oferecidos por instituições de patrimônio. Como resposta a essas demandas, muitas bibliotecas, museus e arquivos já abraçaram o desafio inicial de oferecer infraestrutura para gestão de coleções digitais e oferta de serviços on-line. De forma crescente, essas instituições estão criando representações digitais dos seus acervos físicos e ainda adquirindo conteúdos nato- digitais, tais como arte em mídia digital, dados históricos e dados de pesquisa e os armazenando em repositórios digitais (pennock, 2006) como etapa inicial na oferta de materiais digitais online.

Esse movimento tem acompanhado proximamente o ritmo da evolução do que conhecemos como tecnologias digitais. Já em 2001, Addison observava que o rápido avanço dessas tecnologias – de gráficos 3D a multimídias e realidade virtual – renova as possibilidades da dinamização do patrimônio digital: "ferramentas digitais oferecem novas promessas em documentação, análise e disseminação da cultura" (addison, 2001, p. 1), enfatizando o amplo espectro de aplicações possíveis.

Porém, é preciso considerar ainda que as coleções culturais digitais podem ser desenvolvidas não só para provisionar serviços online para usuários externos à instituição. As representações digitais de acervos físicos podem constituir uma ferramenta imprescindível para a gestão dos acervos originais, para os processos de documentação, conservação, preservação, segurança, marketing e editoração, entre outros.

Num patamar mais inovador, o acervo digital, que está paralelo ao acervo físico original, pode ir além de uma representação funcional deste, ampliando o seu potencial informacional, comunicacional e de reinterpretação e apresentação. Os processos de digitalização permitem que os objetos culturais digitais possam ser agregados com outros objetos formando novos constructos, reinterpretados em outros contextos para outros propósitos, compartilhados, recriados, enriquecidos, anotados com informações que podem ser compartilhadas, incorporados em outras coleções e em outras memórias, e analisados sob outros olhares, fomentando a pesquisa interdisciplinar. Isto significa que as coleções de materiais digitalizados devem ser coleções de materiais digitais primários ou brutos que possam servir de base para as transformações que coletivamente se chama, no domínio da área de curadoria digital, de reuso ou reutilização (sayão; sales, 2012).

Entretanto, a criação e aquisição de ativos digitais colocam o desafio crítico para os museus, bibliotecas e arquivos: dispor de infraestruturas tecnológicas e gerenciais permanentes, de sustentabilidade financeira e de equipes especializadas que deem apoio ao ciclo de vida complexo e de longo prazo dos objetos digitais. Isto acontece porque os conteúdos digitais devem não somente ser coletados ou

criados e disseminados na web, mas também apropriadamente gerenciados, armazenados e preservados para maximizar o investimento inicial e assegurar que os objetos permaneçam disponíveis e compreensíveis para o seu público-alvo pelo tempo que for necessário. É preciso observar que as formas de gerenciamento, as condicionantes tecnológicas e as configurações administrativas e de sustentabilidade econômica e financeira no caso de coleções digitais diferem em muitos aspectos das infraestruturas necessárias para o desenvolvimento e manutenção de coleções físicas.

A curadoria digital oferece um arcabouço prático e conceitual que permite a elaboração de fluxos de trabalho voltados para uma gestão dinâmica de coleções de materiais digitais que pode ser aplicado aos acervos culturais digitais, ampliando o seu potencial de reuso e, como desdobramento, a concepção e o desenvolvimento de serviços online inovadores e de espaços de interação em torno dos acervos digitais.

É sobre o acervo digital paralelo aos acervos físicos e suas perspectivas de aplicação e de reuso possibilitado pelo arcabouço de ferramentas da curadoria digital que falaremos rapidamente neste ensaio.

2 A ideia de acervo digital paralelo

As instituições de patrimônio cultural têm gradualmente reconhecido a urgência de digitalizar suas coleções. A percepção dessa necessidade está globalmente refletida na execução de grandes programas de digitalização de acervos culturais que se desenrolam há algumas décadas. Muitos desses programas se desenvolvem em âmbito nacional ou mesmo internacional, financiados por instituições governamentais, organismos internacionais e empresas privadas, ou ainda com recursos próprios. Os projetos de digitalização em grande escala, talvez por seu enquadramento em parâmetros atuais mais perceptíveis, têm garantido fontes de financiamento (embora o mesmo não se possa dizer sobre a sustentabilidade futura das coleções digitais produzidas).

O desenvolvimento da web cada vez mais rica em conteúdos e cada vez mais sofisticada semanticamente, além da disponibilidade de um amplo repertório de tecnologias digitais e de padrões abertos que permitem níveis satisfatórios de interoperabilidade entre sistemas, torna cada vez mais possível que instituições de patrimônio respondam às demandas dos vários segmentos sociais por serviços on-line ancorados nos seus ativos culturais. Essas demandas se espalham por um amplo espectro que vai da pesquisa científica às ações de educação patrimonial. Nesse contexto de grandes novidades, repositórios digitais estão sendo criados com o objetivo de servir como memória digital do patrimônio cultural, enquanto uma forte tendência de convergência entre museus, arquivos e bibliotecas, em termos de funções infor-

macionais, se torna cada vez mais concreta e presente (CONSTANTOPOULOS, 2010).

Os processos intensos de digitalização têm como perspectiva os benefícios mais diretos para as instituições: dar visibilidade universal aos seus estoques informacionais e tornar mais evidente a sua presença na Rede, reforçando a sua identidade como construção conectada com o seu tempo; alcançar novas audiências – o que pode se traduzir no aumento de visitas presencias aos acervos físicos – e contribuir para a revelação e massificação do que antes estava protegido, implícito e elitizado. Para tal, é necessário que as informações estejam organizadas e indexadas em estruturas de base de dados e repositórios para que sejam encontradas e recuperadas: "um arquivo que não pode ser recuperado por um usuário, simplesmente não existe para ele" (BACHI et al., 2014, p. 2). Novamente as infraestruturas gerenciais e tecnológicas são fatores determinantes.

Porém, as instituições culturais não desenvolvem suas coleções digitais paralelas ao acervo físico somente para prover serviços online, como exposições virtuais e dispositivos de busca e acesso. Muitas instituições são movidas também pela necessidade de constituir acervos digitais que sejam também ferramentas de apoio à gestão de seus acervos físicos originais.

Nos museus, os registros fotográficos são comumente utilizados para documentação, apresentação, pesquisa, conservação e gerenciamento dos objetos e coleções (STARRE, 1996). Com a automação dos acervos, as imagens digitais – pela versatilidade, possibilidade de agregação com outros recursos digitais, tais como textos, planilhas e pela possibilidade de transmissão quase instantânea – passaram a ser parte importante dos registros de uma peça, substituindo gradualmente as fotos em papel. Tornou-se, dessa forma, mais fácil e barato realizar os registros dos acervos por meço de imagens digitais, sem contar que as imagens podem se visualizadas mais frequentemente sem comprometer a sua conservação através de diferentes dispositivos, independentemente de localização geográfica, e podem ainda ser eletronicamente transmitidas quase instantaneamente.

Nessa direção, os objetos digitais passam a ser um elemento importante na documentação dos acervos, na identificação e descrição dos objetos, no registro dos seus principais detalhes e do estado de conservação antes e depois de um processo de restauração; as imagens tornam-se importantes também no reconhecimento de uma peça num evento de furto, roubo ou desaparecimento; além do mais, elas contribuem também para a preservação dos objetos quando evitam o manuseio e a exposição daqueles particularmente frágeis, raros e únicos. Assim, a digitalização apoia um conjunto de funções tradicionais desempenhadas por museus, arquivos e bibliotecas, tornando mais fácil e produtivo o gerenciamento de suas coleções físicas. A tabela 1 a seguir esquematiza algumas dessas funções:

Quadro1 – Algumas funções do acervo digital

FUNÇÃO	DESCRIÇÃO
Acesso	O acesso via web às coleções tem sido o principal objetivo das instituições quando se engajam em projetos de digitalização; as coleções digitais são complementos importantes para as visitas presenciais e contribuem para o aumento destas, revelando detalhes, ângulos e destaques que muitas vezes passam despercebidos ao visitante presencial.
Documentação	As imagens fazem parte dos registros dos objetos físicos, incluindo a sua identificação, substituindo as fotografias convencionais.
Conservação	As imagens digitais apoiam o acompanhamento do desenvolvimento do aspecto físico da obra, a fim de constatar o surgimento de alguma avaria; assistem no planejamento de ações que retardem ou impeçam o andamento da deterioração.
Restauração	A digitalização apoia o registro do estado físico da obra anterior ao processo de restauração e do estado final resultante do processo; registra o desenvolvimento da aplicação dos processos de restauração, possibilitando a construção e publicação de dossiê específico.
Segurança	As imagens ajudam na identificação e no reconhecimento de peças em eventos de roubo ou furto.
Marketing e comunicação	Uso na preparação de brochuras, material promocional, relações públicas, *press releases*, pôsteres, *outdoors* etc.
Publicação	Como material fonte para ilustrações de publicação tais como catálogos, outdoors, livros, publicações acadêmicas e relatórios.
Mídia eletrônica	Como elementos imagéticos do website da instituição, de exposições virtuais e de produtos multimídias.
Memória	Os objetos digitais contribuem para a complementação de lapsos e descontinuidades da memória das instituições culturais.
Preservação dos originais físicos	As representações digitais – dependendo da qualidade através da qual foram geradas – podem substituir para a maioria das necessidades os objetos originais, tanto do ponto de vista gerencial quanto do ponto de vista de pesquisa. Dessa forma, evitam manipulações desnecessárias desses originais.

Fonte: Elaborado pelo autor.

3 Ampliando as potencialidades das coleções digitais

Os usos dos acervos digitais paralelos identificados acima limitam o uso da digitalização como parte das estratégias de gestão e de acesso das coleções aplicadas pelas instituições de patrimônio. "Tais praticas fazem uso de tecnologias digitais somente como ferramentas e não como meço de interação" afirmam (REIS;

serres; nunes, 2016, p. 61). Esses acervos – da forma como são modelados – são compreendidos como patrimônio digital, contudo são isentos de possibilidade de interação, reutilização, reinterpretação e de serem agregadores de comunidades virtuais (gruber; glahn, 2009).

As afirmações dos autores citados acima ressaltam o fato de que as coleções digitais paralelas aos acervos físicos – que de uma forma geral são simulacros funcionas dos seus originais, duplicando-os com a maior fidedignidade que a técnica e os orçamentos permitem – podem ir além das funções de acesso e gestão. Elas podem amplificar as potencialidades dos acervos físicos e dessa forma revelar novas formas de apresentação, contextualização e interpretação. Essas potencialidades são ampliadas pela natureza dinâmica de fragmentação, recomposição, edição e agregação (como peças de Lego) dos objetos digitais (kallinikos, 2010) em ambientes virtuais baseados em padrões de interoperabilidade.

Pode-se, como observam Reis, Serres e Nunes (2016, p. 59):

> compreender o meço digital como um facilitador de acesso e precursor de novas possibilidades de imersão nos lugares de memória, afastando-se de uma concepção simplista do digital como mero repositório de informação.

O armazenamento de informação não é memória (ramsey, 2016) e não transmite conhecimento inercialmente, para isso são necessárias ações intencionas que ativem suas potencialidades.

Uma base de dados de imagens tem o potencial de amplificar o domínio de interação e usabilidade das coleções e desencadear processos comunicacionais se for pensada como uma fonte de materiais brutos ou primários cujo potencial de reinterpretação e interação possa ser ativado em ambientes virtuais. Essa ativação se efetiva por informações de representação, tecnologias digitais e padrões que permitam graus de interoperabilidade e compartilhamento.

A digitalização e a aquisição e geração de materiais digitais e os processos contínuos de gestão ativa sobre os acervos digitais devem ser conduzidos como forma de destacar as potencialidades de agregação, representação e reinterpretação, que poderíamos chamar coletivamente de "reusabilidade", e ainda proporcionar mecanismos de experimentação e de interlocução. Além do mais, estes acervos e os artefatos gerados por suas reconfigurações precisam ser preservados com níveis adequados, para cada caso, de proveniência e autenticidade, e ter os diretos associados a eles considerados. A resposta a esses desafios pode vir da curadoria digital que reúne um conjunto de metodologias voltado para a gestão dinâmica de conteúdos

digitais de naturezas distintas, incluindo a preservação e o arquivamento confiável, como será visto com um grau a mais de detalhes na seção 5.

O conceito de reuso é aplicado intensamente no domínio da pesquisa científica, onde dados e outros materiais são submetidos a outros olhares, analisados em contextos e disciplinas diferentes para os quais orginalmente foram gerados, fomentando a pesquisa interdisciplinar e o compartilhamento de informação e conhecimento. Seus pressupostos podem ser aplicados às coleções digitais das instituições de cultura e patrimônio. É o que será discutido a seguir.

4 Reuso: reinterpretando as coleções digitais culturais

Lynch (2002), em uma conferência proferida no começo do século XXI sobre digitalização de informações de patrimônio cultural, observa que todo o esforço das instituições de patrimônio está em produzir grandes quantidades de conteúdo digital e oferecer tipos simples de ferramentas de acesso, ao invés de sistemas mais sofisticados para uso contínuo, ou oferecer dispositivos para reuso e interpretação dos conteúdos. O uso inovador de tecnologias digitais pode ir muito além da mera criação de representantes digitalizados: essas tecnologias têm o potencial não somente de engajar novas audiências para as coleções dos museus, mas também de produzir concepções inéditas de produtos e serviços culturais.

Lynch (2002) reforça a ideia de que é necessário empacotar os conteúdos brutos das coleções – matéria-prima gerada pelos processos de digitalização – de várias formas, tais como experiências de aprendizado, exposições curadas ou interpretações e análises, criando novos artefatos intelectuais e serviços. Enquanto os museus, arquivos, bibliotecas e as comunidades de ensino superior são os maiores criadores de coleções digitais, os criadores de apresentações e interpretações de materiais baseadas nessas coleções serão muito mais numerosos e diversificados. Em outras palavras, as derivações das coleções primárias terão – em termos espaciais e temporais – um alcance maior do que os originais e seus equivalentes digitais. É isto que fazemos cotidianamente na faina intelectual de geração de conhecimento: "Se observarmos os processos de pesquisa acadêmica, eles incluem uma contínua reinterpretação de fontes estabelecidas de conteúdos (incluindo a avaliações de novos materiais)" (LYNCH, 2002, p.4) e, desse modo, geram novos conhecimentos manifestados em livros, artigos, modelos, simulações, exposições e muito mais. As complicações da dicotomia entre materiais digitais brutos e interpretação parecem ter um alcance bastante amplo no mundo cultural e no de pesquisa e ensino.

Além do mais, é preciso pensar nos acervos digitais como um pretexto e um substrato para a socialização e compartilhamento de ideias e para a formação das memórias digitais distribuídas e virtualmente integradas. Esta possiblidade con-

figura um dos maiores desafios nas práticas de patrimônio cultural digital, que é "compreender as possibilidades trazidas pelo meço digital, em especial no que diz respeito aos espaços colaborativos para ativação patrimonial e acesso à memória coletiva" (reis; serres; nunes, 2016, p. 58).

A ideia de reuso de conteúdos culturais digitais começa a se institucionalizar e se tornar também um novo nicho de negócios para a indústria de conteúdos. O Projeto *Europeana Space*[3] cujo lema é "um espaço de possibilidades para o reuso criativo de conteúdo cultural" ilustra bem esse novo conceito de reinterpretação de informações culturais. O objetivo do Projeto – conforme informa seu website – é a criação de novas oportunidades de emprego e de crescimento econômico no setor das indústrias criativas europeias com base nos recursos culturais digitais. Nesse sentido, ele se constitui como uma rede de boas práticas onde as possibilidades de reuso criativo de coleções digitais são investigadas, testadas e encorajadas. O *Europeana Space* oferece ainda um ambiente aberto para o desenvolvimento de aplicações e serviços baseados nos conteúdos digitais culturais. O uso desta plataforma é incentivado por um forte, abrangente e sustentável programa de promoção, difusão e replicação das boas práticas desenvolvidas no âmbito do projeto. Como resultado final, o projeto espera gerar produtos e serviços inéditos prontos e testados para serem distribuídos no mercado, e, dessa forma, capitalizar o potencial de negócios da herança digital cultural, criando novos empregos e oportunidades de negócios (bachi et al., 2014).

Possivelmente a face mais elaborada conceitualmente do termo "reuso" diz respeito à gestão de dados digitais de pesquisa, cujas metodologias se tornam cada vez mais sofisticadas e universais. Os pressupostos da ciência aberta, o princípio fundamental da reprodutibilidade dos experimentos científicos e os altos investimentos na geração e coleta de dados tornam o reuso de dados de pesquisa um desafio importante do nosso tempo, cujos conhecimentos, porém, podem ser transpostos para outras áreas. É oportuno atentar ao fato de que muitos museus, especialmente os de história, história natural e museus de ciências, também produzem e coletam dados digitais de pesquisa de valor contínuo, tais como dados arqueológicos e históricos, e muitos deles possuem cursos de pós-graduação e atividades de pesquisa que geram mais dados. Essas coleções de dados brutos são matérias-primas para reuso e precisam ser preservadas, arquivadas e passar por processos de curadoria.

Abaixo seguem alguns exemplos no domínio da educação, pesquisa, curadoria e interação de reuso de materiais digitais culturais, tendo como base a agregação, aplicações computacionais e o enriquecimento de materiais digitais brutos.

3 Disponível em: http:// www.europeana-space.eu. Acesso em: 30 jul. 2021.

- Agregações – possibilidade de recombinar os objetos digitais em agregações – como blocos de construção ou peças de Lego – identificadas e com autoria reconhecida que possibilitam reinterpretações e apresentações inéditas; materiais (imagens, vídeos, hipermídia) de várias fontes rearranjadas em novas coleções que confere uma nova perspectiva à coleção como um todo (SCIME, 2009); padrões internacionais como *Open Archives Initiative Object Reuse and Exchange* (OAI-ORE), *Resource Description Framework* (RDF), *Linked Data* e ontologias como *Europeana Data Model* (EDM) e o *Conceptual Reference Model* do CIDOC (CIDOC-CRM) abrem a perspectiva importante na contextualização, agregação e interoperabilidade dos materiais digitais e na derivação de vários e novos produtos e serviços culturais; possibilita ainda a integração de museus, bibliotecas e arquivos. As agregações não são fixas e variam no tempo, dado que podem ser versionadas e receberem contribuições;
- Espaço colaborativo – uma contribuição importante vinda do mundo da pesquisa científica é a possibilidade de interagir com os dados de pesquisa, criando formas de interlocução (formalizadas ou não) que podem ser registradas e compartilhadas com toda a comunidade envolvida, conhecidas como anotações. Essas intervenções críticas podem fazer parte integrante das coleções e contribuem para enriquecer o contexto informacçonal dos objetos digitais, adicionando valor a esses materiais. Os sistemas culturais – assim como fazem os sistemas científicos – precisam constituir formas de interlocução ao redor de suas coleções, incorporando esses diálogos às informações contextuais dos materiais primários ou de suas representações, por meio de sistemas próprios ou das redes sociais e blogs. Os espaços colaborativos se tornam essenciais para ativação patrimonial e acesso à memória coletiva (REIS; SERRES; NUNES, 2016);
- Curadoria online – os materiais digitais brutos podem ser recombinados, enriquecidos com outras informações e reinterpretados por curadores convidados, formando exposições virtuais que conferem novas visões aos materiais digitais; usuários não especialistas podem dispor de softwares aplicativos que permitem que eles construam suas próprias exposições enriquecidas com materiais pessoais e que podem ser armazenadas e receber contribuições diversas, criando espaços colaborativos; as exposições virtuais podem ser manifestações de exposição física, complementando, roteirizando, interpretando e fazendo *links* com outros recursos;
- Educação – embora haja uma grande ênfase no ensino online, é fácil constatar que o uso (e reuso) dos materçiais digitais disponíveis nas bases de dados das instituições de patrimônio cultural ainda são bastante restritos no que

diz respeito à elaboração de materiais didáticos digitais. Muitos especialistas defendem que os acervos digitais devem ter como ponto focal mais importante a disseminação para a educação. Nessa direção, softwares aplicativos podem ser desenvolvidos para criação e arquivamento de aulas virtuais, cursos, palestras e tutoriais construídos a partir de materiais digitais de diversas fontes e mídias, como imagens, imagens em 3D, vídeos, simulações, videogames, etc. Professores/curadores podem ser convidados para desenvolver aulas e cursos. As aulas podem ser armazenadas em bases de dados distribuídas como recurso agregado segundo padrões consagrados como OAI-ORE, podendo ser recuperadas e recombinadas formando novos constructos;

- Pesquisa científica – muitos museus geram ou coletam dados digitais de pesquisa que são gerenciados tendo como perspectiva o reuso em novas pesquisas e para o ensino científico. Com esse propósito os materiais digitais são retrabalhados tanto para as áreas específicas para as quais eles foram gerados, quanto para pesquisas interdisciplinares e para análises em outros contextos. Entretanto, as coleções digitais e os processos de digitalização podem ser fontes importantes de conhecimentos implícitos que podem ser revelados por metodologças computacionais, como a mineração de dados, e que podem ser analisados em outros contextos, gerando novos conhecimentos. Por exemplo, as centenas de diários de bordo digitalizados, registrando viagens marítimas nos últimos três séculos, se tornam uma base de dados incomparavelmente rica sobre a fauna, a flora, correntes e ventos oceânicos – as condições atmosféricas a partir das quais os cientistas reconstroem a história dos sistemas dinâmicos da Terra e melhoram as projeções sobre o futuro do clima (RAMSEY, 2016).
- Aplicativos computacionais – oferta de programas aplicativos para plataformas móveis e computadores que apoiem o usuário na manipulação dos conteúdos das bases de dados; interfaces que permitam criação de áreas pessoais (minhas coleções) e possibilitem o compartilhamento via redes sociais; disponibilidade de APIs para desenvolvedores de games, realidade ampliada, realidade virtual, etc.;

O reuso confiável de materiais digitais só é possível se eles são curados de tal forma que sua autenticidade e integridade sejam mantidas ao longo do tempo (PENNOCK, 2006), isto porque um criador– seja um pesquisador, professor ou curador de uma exposição virtual – confia no material digital coletado ou gerado por outro para dar prosseguimento ao seu trabalho, delinear o seu projeto ou criar novos artefatos. Isto coloca em primeiro plano a questão de proveniência e de como ela é

endereçada pelos sistemas que cuidam da curadoria digital, da preservação digital e do arquivamento confiável.

A consideração de que a curadoria digital, tomada como ponto de partida para a ativação das possibilidades de reuso, depende de infraestruturas tecnológicas que tornem os recursos visíveis para pessoas e por sistemas, é fundamental. Porém, essas infraestruturas não devem se restringir a computadores e redes, mas tomadas num sentido mais abstrato, que inclua ontologias, taxonomias, modelos de interoperabilidade e, sobretudo, as informações de representação estruturais e semânticas, efetivado por esquemas apropriados de metadados, que assegurem agora e no futuro a decodificação das informações pelos seus públicos-alvo.

5 Preservando o efêmero

Uma das complicações no desenvolvimento de estratégias de preservação para os conteúdos digitais é que essas estratégias podem exigir enfoques inovadores que desafiam as práticas existentes, bem como as estruturas e hierarquias organizacionais necessárias a sua execução. A gestão de conteúdos digitais é completamente diferente, em muitos aspectos, do tratamento dado aos objetos físicos em termos práticos e conceituas (a própria ideia de preservação digital, que essencialmente objetiva preservar o acesso, esbarra no conceito de preservação física que é, na maioria dos casos, antagônica ao acesso).

O reconhecimento de que os conteúdos digitais não sobrevivem sem intervenções efetivas, e que eles precisam ser intencionalmente preservados de forma ininterrupta por todo o seu ciclo de vida – que se inicia no momento do planejamento da sua criação – é de dramática importância para a sua sobrevivência e para a posterior suposição de sua autenticidade. O sucesso das memórias digitais será decidido em comparação com rivais que possuem tradição secular. As vantagens oferecidas pelas tecnologias digitais, pelo armazenamento de massa, pela facilidade de cópia, de transmissão e de reformatação e pela facilidade de pesquisa e análise não serão suficientes a menos que a confiabilidade, a preservação por longo prazo, as facilidades para o reuso, a recombinação e a reinterpretação do conteúdo digital possam ser asseguradas. É neste momento que a preservação e a curadoria digital tornam-se arcabouços técnicos e gerenciais imprescindíveis (CONSTANTOPOULOS, 2010).

Os museus e demais instituições de patrimônio têm uma longa tradição de preservar e oferecer acesso de longo prazo aos seus ativos informacionais, incluindo artefatos em vários formatos. A cuidadosa atenção dos curadores desses registros tem assegurado que eles permaneçam disponíveis para pesquisadores e para o público como fonte de conhecimento, memória e identidade. No novo cenário, em que os objetos digitais se tornam parte da herança cultural, a curadoria digital con-

figura-se como um importante arcabouço para a contínua preservação de coleções digitalizadas ou nascidas digitais, ancorada numa tradição secular.

Nesse contexto de mudanças e de novas interpretações, o termo "curadoria" – que denota uma atividade tradicional no domínio das instituições culturais – só recentemente começou a ser aplicado a materiais digitais "curadoria digital, interpretada de forma ampla, está relacionada à manutenção e ao adicionamento de valor a um corpo confiável de informação digital para uso corrente e futuro" (pennock, 2006, p. 1), em outras palavras, reforma a autora, é o gerenciamento dinâmico e a avaliação de informação digital durante todo o seu ciclo de vida. Nessa direção, todas as atividades envolvidas em gestão de dados, do planejamento à sua criação, melhores práticas em digitalização e documentação e a garantia de disponibilidade e adequabilidade para a descoberta e reuso no futuro são partes da curadoria digital. A curadoria digital também inclui a gestão de grandes coleções de dados para o uso diário (abbott, 2010).

O modelo de ciclo de vida da curadoria digital do *Digital Curation Centre* (dcc)[4] sintetiza abstratamente os fluxos das atividades que se desenrolam num processo de curadoria digital, constituindo a principal referência da área. Na qualidade de um centro especializado em curadoria e preservação digital, o dcc é um ponto focal de pesquisa e desenvolvimento nesses tópicos, promovendo expertise e boas práticas, em âmbito mundial, para a gestão de produtos digitais.

Por fim, a inevitável relação entre curadoria, no seu sentido histórico, e a curadoria digital é colocada como questão de partida nas análises teóricas e práticas de Dallas (2007): como e em que medida a agenda da curadoria digital pode ser relevante para a prática da curadoria no domínio dos museus, da arte e do patrimônio cultural no momento em que as pesquisas que são fundamentadas em coleções e a comunicação com o público dependem de medicação de tecnologias digitais? E inversamente: em que medida uma compreensão sobre museus e sobre as práticas de curadoria de patrimônio cultural contribuem para melhorar a curadoria digital de maternas culturais? Não obstante partirem de um ponto de vista disciplinar – museus, arte e patrimônio digital –, essas indagações tocam sobre aspectos mais amplos que consideram escopo, métodos e natureza epistêmica da curadoria digital, como, por exemplo, sua aplicação nas áreas de pesquisa qualitativa e quantitativa, o que torna essas questões universais.

6 À guisa de conclusão

Os projetos de digitalização de acervos culturais desencadeiam compactos

4 Disponível em: http://www.dcc. ac.uk/. Acesso em: 30 jul. 2021.

importantes na sociedade na medida em que configuram algumas de suas mais importantes atividades. A disponibilidade desses acervos digitais geram benefícios diretos para o ensino, para a pesquisa, para o governo e para muitas atividades econômicas e ainda para os próprios detentores desses conteúdos em termos de visibilidade global, de novos modelos de negócios baseados numa economia de conhecimento, do surgimento de novas profissões e de maiores condições de sustentabilidade econômica. As tecnologias digitais e o poder persuasivo da web são determinantes para essas transformações, que têm como principal consequência uma rápida convergência entre museus, arquivos e bibliotecas, reforçando a metáfora dominante de que toda informação se encontra no mesmo lugar. Porém, as ideias inovadoras e disruptivas são os motores principais dessas mudanças.

Considerar e tratar os acervos digitais culturais como matéria-prima para o reuso em diferentes contextos, como vimos, amplifica o potencial informacional e comunicacional desses ativos, mas, sobretudo, reposiciona as instituições de patrimônio cultural numa dinâmica mais contemporânea e integrada aos fenômenos do nosso tempo. Mas, ao mesmo tempo, deixa questões importantes para os profissionais e pesquisadores da área que vão enriquecer e ampliar o escopo de suas pesquisas e práticas: como os direitos associados a esses novos artefatos gerados pelo reuso devem ser tratados? Como garantir a proveniência e autenticidade desses materiais digitais em ambientes em constante transição, enquanto valores, tecnológicas e padrões estão sempre em evolução? Quais são os possíveis modelos de negócio e de sustentabilidade para as instituições de patrimônio? Como o cidadão comum pode participar dessas mudanças?

Referências

ABBOTT, Daisy. **What is digital curation**? Edinburgh: Digital Curation Centre, 2010. Disponível em: https://www.era.lib.ed.ac.uk/bitstream/handle/1842/3362/Abbott?sequence=3. Acesso em: 20 out. 2016.

ADDISON, Alonzo C. Virtual heritage: technology in the service of culture. *In*: CONFERENCE ON VAST 01: virtual reality, archeology and cultural heritage, 1., 2001, Atenas. **Proceedings** [...]. Atenas: University of Atenas, 2001. p. 343-354. DOI: https://doi.org/10.1145/584993.585055. Disponível em: https://dl.acm.org/doi/10.1145/584993.585055. Acesso em: 30 jul. 2021.

BACHI, Valentina *et al*. The digitization age: mass culture is quality culture. *In*: EUROMED, INTERNATIONAL CONFERENCE, 5., 2014, [*S. l.*]. **Proceedings** [...]. [*S. l.*]: Euromed, 2014. p. [1-17]. Disponível em: http://resources.riches-project.eu/digitization-age-mass-culture-is-quality-culture/. Acesso em: 24 out. 2016.

CONSTANTOPOULOS, Panos. Digital curation and digital cultural memory. *In*: SETN'10 HELLENIC CONFERENCE ON ARTIFICIAL INTELLIGENCE: theories, models and applications, 6., 2010, Atenas. **Proceedings** [...]. Atenas: [*s. n.*], 2010. p. 1-1.

DALLAS, Costis. An agency-oriented approach to digital curation theory and practice. *In*: INTERNATIONAL SYMPOSIUM ON INFORMATION AND COMMUNICATION TECHNOLOGIES IN CULTURAL HERITAGE, 2007, Toronto. **Proceedings** [...]. Toronto: Archives & Museum Informatics, 2007. p. [1-16]. Disponível em: https://www.archimuse.com/ichim07/papers/dallas/dallas.html. Acesso em: 30 jul. 2021.

GRUBER, Marion R.; GLAHN, Christian**. E-learning for arts and cultural heritage education in archives and museums**. [*S. l.*: *s. n.*], 2009. p. [1-12]. DOI: 10.13140/RG.2.1.2941.6401. Disponível em: https://www.researchgate.net/publication/283301043_E-learning_for_arts_and_cultural_heritage_education_in_archives_and_museums. Acesso em: 20 out. 2016.

KALLINIKOS, Jannis. A theory of digital objects. **First Monday**, [*S. l.*], v. 15, n. 6-7, p. [1-20], 2010. Disponível em: http://firstmonday.org/ojs/index.php/fm/article/view/3033/2564. Acesso em: 20 out. 2016.

LYNCH, Clifford. Digital collections, digital libraries and digitalization of cultural heritage information. **First Monday**, [*S. l.*], v. 7, n. 5-6, p. [1-15], 2002. Disponível em: https://firstmonday.org/ojs/index.php/fm/article/view/949. Acesso em: 30 jul. 2021.

PENNOCK, Maureen. Digital curation and management of digital library cultural heritage resources. **Local Studies Librarian**, [*S. l.*], v. 25, n. 2, p. 1-6, 2006. Disponível em: https://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lsl-curation_mep.pdf. Acesso em: 30 jul. 2021.

RAMSEY, Abby Smith. How to preserve cultural memory in the digital age**. Huffpost**, [*S. l.*], 14 jun. 2016. Disponível em: http://www.huffingtonpost.com/abby-smith-rumsey/culture-memory-digital_b_10357622.html. Acesso em: 20 out. 2016.

REIS, Marina; SERRES, Juliane; NUNES, João. Bens culturais digitais: reflexões conceituas a partir do contexto virtual. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, Florianópolis, v. 21, n. 45, p. 54-69, 2016. DOI: https://doi.org/10.5007/1518-2924.2016v21n45p54. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2016v21n45p54. Acesso em: 30 jul. 2021.

SAYÃO, Luis Fernando; SALES, Luana Farias. Curadoria digital: um novo patamar para preservação de dados digitais de pesquisa. **Informação &**

Sociedade: Estudos, João Pessoa, v. 22, n. 3, p. 179-191, 2012. Disponível em: https://periodicos.ufpb.br/ojs/index.php/ies/article/view/12224. Acesso em: 30 jul. 2021.

SCIME, Erin. The content strategist as digital curator. **Content Strategy**, [*S. l.*], n. 297, p. [1-17], 2009. Disponível em: https://alistapart.com/article/content-strategist-as-digital-curator/. Acesso em: 30 jul. 2021

STARRE, Jan van der. 3D Article on multimedia imaging related to museum documentation. *In*: INTERNATIONAL COUNCIL MUSEUMS. Committee for Documentation (CIDOC). Study Series, p. 28-29. 1996. Disponível em: http://icom.museum/uploads/tx_hpoindexbdd/3_ICOM-CIDOC.pdf. Acesso em: 19 out. 2016.

Gestão de dados como serviço: proposta de um modelo

Luís Fernando Sayão e Luana Farias Sales

1 Introdução

Os paradigmas consolidados sobre os quais a pesquisa científica por séculos vem sendo conduzida têm mudado dramaticamente nas duas últimas décadas. Novos métodos, ferramentas, instrumentos, escalas e equipamentos, além de novas fontes de dados – reveladas e potencializadas pelas tecnologias digitais - que se aliam à conectividade global da atividade de pesquisa, estão redesenhando os percursos tradicionais da comunicação científica e ressignificando os arquétipos da metodologia científica. A natureza mutável da pesquisa científica está se deslocando do ambiente tradicional e autocontido para um ambiente de pesquisa digital, conectado, relacional e distribuído em tempo real e altamente colaborativo.

Não importando o ponto de observação, o que se constata é que a pesquisa científica – por essa tendência global aliada às suas conexões imanentes com os sistemas técnicos - está produzindo um enorme e crescente fluxo de dados e de informações digitais. Inúmeros sensores instalados nos mais diversos dispositivos que vão de longínquos satélites, aceleradores de partículas, sequenciadores automáticos de dna, até em despretensiosos implantes médicos, permitem que dados sejam capturados em uma quantidade sem precedentes em todos os domínios científicos, das ciências exatas às humanidades, arte e cultura.

Parte do mesmo fenômeno, o aprimoramento de ferramentas de pesquisa, análise e visualização dota os pesquisadores de capacidade de interpretar essa quantidade crescente de dados trazendo à tona novas relações e padrões ocultos na profusão de dados, inaugurando disciplinas híbridas, cujo coração do conhecimento está na integração de inúmeras fontes de dados. Numa síntese possível, pode-se afirmar que a ciência está crescentemente se deslocando para ser uma atividade que produz e consome intensamente dados de um largo espectro de variedade para a construção do conhecimento científico, em um processo cada vez mais coletivo e colaborativo.

Contudo, esse deslocamento paradigmático coloca alguns desafios críticos e expõe tensões que estão localizadas no próprio cerne da geração do conhecimento

científico. Dentre os desafios mais importantes dessa agenda crítica está a gestão da imensidão de dados engendrados pela pesquisa contemporânea. Pesquisadores, gestores e profissionais de informação se deparam com um vasto espectro de problemas revelados por esse dilúvio informacional entrópico, cuja regra é o excesso e não mais a escassez com que lidavam os sistemas legados de informação.

Esse fenômeno do nosso tempo torna essencial planejar o arquivamento de dados de pesquisa, bem como os postulados e marcos epistemológicos, éticos e legais de compartilhamento e reuso (MUSHI; PIENAAR; VAN DEVENTER, 2020). A abundância de dados digitais de pesquisa desafia os profissionais da Biblioteconomia e da Ciência da Informação, e ainda, outros profissionais, como os arquivistas e informáticos a explorar essa avalanche de *streaming* de informações e dados provenientes de descobertas científicas e de atividades acadêmicas, preservando as evidências únicas para o uso futuro (JOHNSTON, 2017).

A integração de coleções digitais de dados e os dispositivos avançados de análises oferecem novas oportunidades para pesquisas inéditas, disruptivas e integrativas. De fato, muitos tipos de pesquisa, incluindo sínteses de pesquisa, meta-análises, análises longitudinais e investigações em escala global, exigem que dados provenientes de muitas fontes sejam descobertos, acessados, integrados e colocados prontos para os dispositivos de análises. Entretanto, a descoberta, o acesso e o reuso eficiente de dados digitais de pesquisa são impactados por inúmeros desafios causados pela forma como as comunidades científicas tratam atualmente seus dados (MAYERNIK *et al,* 2012).

Face a essas constatações, a gestão de dados de pesquisa se configura, atualmente, como um foco de interesse e um dos maiores desafios para as instituições de pesquisa. Em escala planetária, as ações de gestão e curadoria de dados se destacam com proeminência no cenário da pesquisa do século XXI, acompanhando a ubiquidade das tecnologias digitais para a coleta, análise e arquivamento em quase todos os domínios disciplinares (MAYERNIK *et al.,* 2012). Assim sendo, as instituições de pesquisa, em gradações distintas, estão reconceituando a gestão de dados e a identificando como parte integrante dos processos de pesquisa, reconsiderando ou ampliando as suas estratégias de tratamento dos dados, implementando plataformas de gestão e curadoria, adquirindo ferramentas de análises e visualização e desenvolvendo programas de capacitação para as suas equipes.

Não obstante termos ultrapassado algumas décadas na compreensão e equacionamento da gestão de dados, e haver um interesse crescente por parte dos vários *stakeholds* – pesquisadores, agências de fomento, instituições de pesquisa e profissionais de informação, casas editoras científicas e tantos outros –, o nível de sucesso em implementar e oferecer serviços de gestão de dados no âmbito das várias or-

ganizações de pesquisa intensiva em dados não tem sido consistente e ainda está, de forma geral, nos estágios preliminares. O planejamento, desenvolvimento e implantação de plataformas de gestão de dados de pesquisa, devido ao número de variáveis que precisam ser equacionadas, é um problema complexo e multifacetado. A solução para esse problema precisa ser articulada em torno de fluxos de trabalho de domínios disciplinares específicos, parâmetros informacionais, tecnológicos, políticos, éticos e legais de sustentabilidade e de expertise numa odisseia marcada por constantes mudanças, cuja principal característica é a heterogeneidade.

Este ambiente complexo pode ser um terreno adequado para a adoção dos Princípios FAIR (MONS *et al.,* 2017) como horizonte para a implementação de serviços de gestão que tornem os dados de pesquisa encontráveis, acessíveis e interoperáveis, para que possam ser reusados por longo prazo. Neste sentido, cria-se condições para a transição de uma pesquisa autocontida para uma pesquisa mais aberta, em rede, cooperativa e relacional, que, ao mesmo tempo, atende a requisitos disciplinares e beneficia comunidades de culturas e restrições específicas.

No entanto, o alinhamento e a implementação dos Princípios FAIR em uma instituição de pesquisa exigem investimentos financeiros, mudanças culturais, treinamento e a construção de uma infraestrutura técnica apropriada (GRAFF; WAAIJERS, 2011). Estes fatores podem ser aglutinados em torno do conceito de "plataforma de gestão de dado de pesquisa", cuja implementação tem o potencial de operacionalizar as diversas camadas de gestão e estabelecer uma crescente infraestrutura de serviços informacionais, científicos, computacionais e administrativos que viabilize uma escala de aderência aos princípios FAIR dos objetos de pesquisa, sejam eles dados propriamente ditos ou algoritmos, códigos, *workflow* ou outros dispositivos físicos ou conceituais que levam aos dados, bem como os metadados e as infraestruturas para gestão de dados. Este processo é identificado como "FAIRificação" (MONS *et al.,* 2017).

Tentando equacionar essa diversidade, o presente trabalho tem como objetivo apresentar uma arquitetura genérica para apoiar o projeto de plataformas de serviços de dados, definindo, realinhando, agregando e articulando os vários módulos conceituais – diretrizes, políticas, serviços, ferramentas, infraestruturas etc. - em torno de um modelo de camadas, que como blocos de construção podem ser ajustados de acordo com a profundidade, alcance e filosofia de cada instituição ou disciplina. O modelo pode ser usado para constituir uma possível escala para a mensuração do nível de maturidade dos projetos de serviços de gestão. A arquitetura proposta tem como horizonte tornar os dados aderentes aos Princípios FAIR e tornar mais concreta a ideia de uma Internet de Dados e Serviços FAIR (IFDS, na sigla em inglês para FAIR Internet Data and Services).

Para o delineamento dos elementos da arquitetura proposta, foi tomado como metodologia a análise da literatura da área, com ênfase especial em artigos, relatórios, manuais e projetos de infraestrutura de dados elaborados por pesquisadores e instituições de pesquisa, dando proeminência aos pontos de observação dos atores mais proximamente envolvidos na questão.

Para chegar à proposição do modelo, o presente ensaio está organizado tendo como ponto de partida um esboço do que se compreende por 'gestão de dados de pesquisa', distinguindo-a e relacionando-a com outras atividade conectadas ao tratamento e cuidado com os dados de pesquisa; em seguida, procura-se conceituar a gestão de dados na forma de ações que se materializam como um conjunto de serviços, que denominamos "gestão de dados como serviço". Neste percurso, verifica-se que o ambiente técnico-social para esses serviços, mais as suas ferramentas, metodologias e estrutura organizacional, ultrapassam os limites de repositório de dados e encontram um ambiente apropriado num sistema com amplitude maior, que abrigue as idiossincrasias disciplinares, que é conduzido, neste estudo, pelo conceito de "plataforma de gestão de dados de pesquisa". Por fim, propõe-se uma definição conceitual de "serviço de gestão de dados de pesquisa", na qual o modelo proposto será alicerçado.

2 Gestão de dados de pesquisa

Dados de pesquisa frequentemente se manifestam na forma de conjuntos complexos de dados, compostos por diferentes tipos de informação, densamente condicionados por contextos construídos pelas especificidades de seus domínios disciplinares, cujos significados dependem da profundidade das formas de representação de sua cadeia de proveniência. A manutenção desses conjuntos de dados requer conhecimento especializado sobre os ambientes científicos onde são coletados ou gerados e de conhecimento avançado em tecnologia computacional e informacional para organizar e arquivar os dados de forma que eles possam ser apropriadamente preservados e reusados (NIELSEN; HJORLAND, 2014). Na perspectiva de Mayernik e colaboradores (2012), a gestão de dados é um problema multifacetado que demanda tecnologias, estruturas organizacionais, conhecimento humano e habilidades para juntar, de maneira complementar, um largo espectro de variáveis, caracterizando-as, dessa forma, como uma equação de resolução complexa.

O principal objetivo da gestão de dados de pesquisa é revelar o potencial de transmissão de conhecimento dos dados gerados numa investigação científica, transformando o conhecimento que é local e tácito em global e explícito para (re) uso no seu percurso espacial e temporal. Isto é realizado por meio de sucessivos graus de agregação de valor que se sucede por todo o ciclo de vida dos dados –

do seu planejamento inicial ao arquivamento no fim do projeto - que é alcançado por intermédio de processos informacionais, computacionais e científicos. Esses processos que chamaremos de serviços de gestão de dados são desenvolvidos no âmbito de arcabouços técnicos, gerenciais e sociais, que no decorrer deste trabalho serão coletivamente denominados de **plataforma de gestão de dados de pesquisa**.

Uma definição para gestão de dados de pesquisa frequentemente citada e que tem a amplitude conceitual necessária é a colocada por Cox e Pinfield (2014) que, em síntese, preconizam que a gestão de dados de pesquisa é uma série de atividades técnicas e gerenciais associadas ao ciclo de vida dos dados.

> gestão de dados de pesquisa consiste em um número de diferentes atividades e processos associados com o ciclo de vida dos dados, envolvendo o projeto de criação de dados, armazenamento, segurança, preservação, recuperação, compartilhamento e reuso, tudo isso levando em consideração as capacidades técnicas, considerações éticas, questões legais e infraestruturas de governança (COX; PINFIELD, 2014, p. 300).

Essas atividades e processos são exigidos para cobrir um amplo espectro de formas de dados que vão de cálculos em larga escala - originados por dispositivos computacionais de alto desempenho, dados observacionais coletados por instrumentos astronômicos, passando por resultados de experimentos científicos realizados em laboratórios -, até o registro sonoro de entrevistas e a coleta manual de espécimes em um ecossistema. A gestão de dados é, portanto, um conjunto complexo de atividades que envolve uma matriz de desafios técnicos, bem como um grande número de questões culturais, gerenciais, legais e políticas (PINFIELD; COX; SMITH, 2014). Com uma longa faixa temporal de aplicação, a gestão efetiva dos dados traz a promessa de benefícios durante e depois do desenvolvimento de um projeto de pesquisa (JONES; PRIOR; WHITE, 2013).

"O retorno de uma boa gestão de dados [...] são publicações digitais de alta qualidade que facilitam e simplificam os processos em andamento de descoberta, avaliação e reuso em pesquisas subsequentes" (WILKINSON, 2016, p.1). Nessa perspectiva, a gestão de dados tem como desafio final a otimização do reuso desses dados por seus próprios criadores, por seus pares e ainda por pesquisadores de outras áreas, catalisando, dessa forma, a pesquisa transversal e interdisciplinar – que é onde, via de regra, acontece a inovação. Dados de pesquisa bem gerenciados, no ambiente de pesquisa contemporânea, é reconhecidamente um fator essencial para uma pesquisa de alta qualidade; a boa gestão os torna mais fáceis de usar e reusar, o que se traduz em maior coeficiente de colaboração entre cientistas, maximização

do retorno do investimento das agências financiadoras de pesquisa e do atingimento dos objetivos de transparência dos métodos e dos fluxos de trabalho, e o alcance de níveis aceitáveis de reprodutibilidade dos experimentos científicos, paradigma tão caro para a ciência (STRASSER, 2015).

Uma escala renovada para a avaliação dos processos de gestão de dados e dos seus efeitos sobre as coleções de dados pode ser dimensionada pelo grau de aderência desses ativos aos Princípios FAIR e na sua ênfase em expandir a habilidade das máquinas de automaticamente identificar, encontrar, interoperar e usar os recursos digitais envolvidos nas pesquisas. Isto coloca em pauta a crescente relevância dos "*stakeholds* computacionais" – aplicações e agentes computacionais que performam recuperação e análise de dados em nosso nome. "Estes *stakeholds* computacionais [...] demandam mais e mais atenção à medida que sua importância cresce" (MONS *et al*., 2017, p.1), requisitando que os recursos que desejam aderir ao máximo aos Princípios FAIR utilizem estruturas legíveis por máquina amplamente aceitas para a representação e o compartilhamento de dados e conhecimento (MONS *et al*., 2017). Nessa direção, apoiar a ação dessas aplicações e de agentes computacionais se torna uma consideração crítica para todos os participantes do ciclo de gestão de dados – do pesquisador/produtor de dados aos repositórios onde eles serão arquivados.

É preciso observar também que, para a efetiva "FAIRificação" do ecossistema de dados, a aderência não deve se aplicar somente aos dados, no sentido mais estrito, mas também aos algoritmos, ferramentas, códigos e *workflows* que levam aos dados, posto que todos os componentes dos processos de pesquisa devem estar disponíveis para assegurar a transparência, a reprodutibilidade e a reusabilidade, enfatizam Wilkinson e seus colaboradores (2016, p.1). Este alinhamento tem como contrapartida um aporte de investimentos infraestruturais e de políticas e ações voltadas para as mudanças comportamentais e organizacionais que se articulam em torno do conceito de "plataforma de gestão de dado de pesquisa".

É necessário lembrar ainda que o termo gestão de dados de pesquisa não é o único usado para caracterizar as atividades relacionadas ao tratamento e cuidados com os dados de pesquisa, outros termos se sobrepõem conceitualmente ou são usados para descrever processos e atividades específicas nesse universo multifacetado, como por exemplo, governança de dados, curadoria digital, administração de dados, arquivamento de dados, conservação e preservação de dados, entre outros. Embora as tecnologias e processos para preservar e disseminar informações digitais estejam consolidados em muitas disciplinas por décadas, é necessário ir mais além. Identificar a composição de práticas, conhecimentos e culturas é essencial para a construção de um domínio de conhecimento que possa ser chamado de gestão de dados de pesquisa (NATIONAL RESEARCH COUNCIL, 2015, p.18). Essa con-

dição coloca em pauta a discussão conceitual entre gestão e curadoria de dados de pesquisa, cujos contornos se tornam essenciais para o presente estudo.

3 Gestão, curadoria e preservação de dados: ainda uma controvérsia

A curadoria de dados de pesquisa ainda é um conceito em construção na Biblioteconomia, Ciência da Informação e Arquivologia e mesmo nas instâncias artísticas e culturais onde a intensificação das ações de digitalização impõe novos requisitos no planejamento, tratamento e reuso dos acervos digitais resultantes. Porém, essa imprecisão conceitual é mais evidente nos diversos domínios disciplinares, onde a idiossincrasia dos fluxos de geração, processamento e análise de dados é dramaticamente diversificada. O relatório *e-Science Curation Report,* editado pelo JISC em 2003, alertava que "este é um campo relativamente novo, e a terminologia não está ainda estável" (LORD; MACDONALD, 2003, p. 12). Decorridas quase duas décadas, as controvérsias terminológicas e conceituais persistem.

Não obstante as indefinições teóricas e epistemológicas da curadoria digital no domínio da pesquisa científica, as aplicações práticas são identificadas, em escalas distintas, pelos pesquisadores e profissionais de informação que, dessa forma, consolidam um corpo de conhecimento baseado no exercício profissional cotidiano. As práticas, portanto, colaboram para revigorar as bases conceituais desse novo domínio. Porém, esse contexto em transição provoca uma insegurança na localização da curadoria na cartografia informacional e na sua relação com a gestão de dados. Tentando superar as diversas interpretações e controvérsias e ir adiante, no modelo que aqui será proposto, a curadoria de dados de pesquisa é identificada como um coletivo de serviços aplicados aos dados de pesquisa, que amplia o seu valor informacional. O objetivo finalístico da curadoria de dados de pesquisa é permitir a descoberta e o reuso dos dados [agora e no futuro], posto que novos usos para as coleções de dados podem evoluir ao longo do tempo, revelando novas perspectivas disciplinares e interdisciplinares; assim sendo, "os dados precisam permanecer num estado ótimo que permita que eles estejam aptos para serem reusados em ambientes tecnológicos correntes" (CHOUDHURY *et al.*, 2018, p. 3). Nesse contexto, a dimensão temporal coloca como ponto crítico as metodologias da preservação digital e, por conseguinte, as tecnologias de arquivamento e armazenamento subjacentes.

Na esfera mais técnica, a curadoria digital se configura como um conjunto de ações que tem como objetivo agregar valor às coleções de dados, por meio de uma gestão dinâmica e contínua, que visa otimizar o uso e o reuso dos dados – o que inclui atividades de promoção -, apoiar a reprodutibilidade dos experimentos científicos, a acessibilidade, a encontrabilidade, a qualidade e a confiabilidade dos dados para aplicação imediata, porém com um olhar nas condições futuras. Por exemplo:

documentar e descrever ricamente as coleções de dados por meio de metadados acionáveis por máquina, com uso de vocabulário padronizado e internacionalmente aceito; assegurar que os dados estejam completos, autoexplicáveis e acurados; e que estejam preparados para o acesso de longo prazo mantendo integridade, confiabilidade e proveniência; e estejam arquivados em ambientes confiáveis. Dessa forma, os diversos serviços de curadoria podem se sobrepor ou se utilizar de outros serviços e processos que adicionam valor espacial e temporal aos dados como, por exemplo, o arquivamento em repositórios, a padronização para a interoperabilidade e a preservação de longo prazo.

Um reordenamento que permita identificar a curadoria de dados no contexto mais amplo do gerenciamento de dados é essencial para a construção de uma arquitetura de significados que se sobreponha aos serviços, funções e processos necessários à FAIRificação dos dados e de outros objetos de pesquisa que orbitam em torno deles. Nesse sentido, o diagrama colocado por Lord e Macdonald (2003), com o propósito de estabelecer contornos mais claros ao trabalho de assegurar infraestruturas apropriadas e serviços de apoio à pesquisa, pode indicar um ponto de partida. Os autores, em um contexto teórico-prático, propõem definições operacionais para "curadoria", "arquivamento" e "preservação", salientando o grau de interdependência entre elas. Neste ponto de observação, "a preservação é um aspecto do arquivamento, e o arquivamento é uma atividade necessária para a curadoria. Os três conceitos endereçam vertentes específicas da gestão de mudanças ao longo do tempo", resumem Lord e Macdonald (2003, p.12).

Indo mais adiante nas interrelações que se manifestam em torno da ideia de gestão de dados, o Projeto *Data Conservancy*, apresentado por Mayernik e colaboradores (2012), tendo como perspectiva a elucidação das fronteiras conceituais de proveniência, rastreabilidade e linhagem de dados, apresenta um modelo de camadas de serviços que representa os conceitos-chave de armazenamento, arquivamento, preservação e curadoria, onde cada camada depende da anterior. Dessa forma, o armazenamento é uma condição necessária, mas não suficiente para o arquivamento; o arquivamento é necessário, mas não suficiente para a preservação; assim por diante. "Este modelo tem sido útil para a comunicação com os usuários que frequentemente usam estes termos como sinônimos" (mayernik *et al.*, 2012, p.160). Nessa perspectiva, na camada mais baixa de serviços do modelo de gestão de dados está o **armazenamento** que descreve os *bits* registrados em discos, fitas ou na nuvem e os serviços de *backup* e restauração, inclui ainda os dispositivos de segurança física, lógica e de rede; o **arquivamento**, algumas vezes chamado de preservação ao nível de *bits*, foca na integridade dos dados por meio de ações ou conceitos, tais como replicação, fixidez e identificação, e de dispositivos tecnoló-

gicos, como os repositórios confiáveis; a **preservação** por sua vez envolve fornecer um *corpus* suficiente de informações de representação, contexto, metadados e de informações de proveniência de forma que algum *stakeholder*, seja humano ou computacional, que não seja o produtor original dos dados, possa interpretá-los, em termos informacionais e de transformações, em pesquisas correntes e futuras; a **curadoria** compreende os processos de adicionar valor aos dados por todo o seu ciclo de vida com o objetivo de ampliar o potencial de descoberta e reuso desses ativos informacionais, em particular, por comunidades diferentes daquelas que os produziram ampliando as suas possibilidades de pesquisas integrativas, transversais e interdisciplinares.

Tomando como base essas considerações, propomos a seguinte estrutura conceitual para dar continuidade à proposta de trabalho:

- **Curadoria** – atividade de **gestão de dados** que compreende os diversos processos de agregação de valor aos dados, por todo o seu ciclo vida, que se inicia antes de sua criação/coleta; tem como objetivo assegurar que eles estejam adequados aos propósitos atuais e potenciais e disponíveis para descoberta e reuso, agora e no futuro. Para *datasets* dinâmicos, isto significa um enriquecimento ou atualização contínua com o objetivo de mantê-los adequados e prontos para possíveis repropósitos.
- **Preservação** - atividade de **curadoria** em que coleções de dados de valor contínuo, apropriadamente avaliadas, são mantidas ao longo do tempo de forma que elas possam ser compreendidas por seres humanos e máquinas face às mudanças tecnológicas e às fragilidades das mídias digitais, e que sejam mantidas as suas propriedades arquivísticas como proveniência, confiabilidade e autenticidade, permitindo, dessa forma, que um pesquisador, ou outro *stakeholder*, possa confiar nesses dados para dar prosseguimento aos seus empreendimentos, sejam eles científicos ou de outra natureza.
- **Arquivamento** – atividade necessária à **preservação** que assegura que a integridade dos dados seja mantida ao longo do tempo por dispositivos tecnológicos confiáveis, e que eles não sofram corrupções ou intervenções não documentadas e possam ser identificados univocamente, recuperados e acessados.
- **Armazenamento** – atividade necessária ao **arquivamento**, que se preocupa com os *bits* registrados em mídias – discos, fitas, nuvem etc. - e a otimização do uso desses dispositivos de armazenamento no atendimento às demandas das pesquisas; endereça também os requisitos de segurança física, lógica e de rede desses registros.

Neste cenário multifacetado, dispositivos técnicos delimitados pelo conceito de repositório de dados parecem ser insuficientes para criar uma ambientação social, relacional e infraestrutural para abrigar a complexidade de ações, serviços e ferramentas necessários à gestão plena dos dados, é preciso ir mais adiante na busca de ambientes de maior envergadura técnica e conceitual.

4 Plataformas de gestão de dados

As atividades, serviços, processos e ferramentas que compõem a gestão de dados geralmente se agregam e se complementam num ambiente que chamaremos de plataforma de gestão de dados, compreendido como um arcabouço técnico, social e gerencial onde se efetivam os cuidados com os dados, segundo políticas e diretrizes institucionais definidas para tal. Historicamente, grande parte do esforço no planejamento dos dados e de desenvolvimento de sistemas de gestão de dados ocorreram de forma isolada, escondido por trás das portas dos laboratórios, e com um enfoque comunitário e disciplinar. Esta configuração inicial evoluiu para um cenário que apresenta arquitetura de sistemas que vão de projetos altamente customizados e de pequena escala, até grandes sistemas de perspectivas mais abrangentes, com alto grau de institucionalização e de internacionalização e de alcance global. A multiplicidade, diversidade e interoperabilidade das plataformas de gestão de dados põem em pauta o conceito técnico-social de ecossistema de dados de pesquisa, que costura as dinâmicas e interlocuções associadas a esses sistemas pelas pessoas e tecnologias.

De forma ideal, essas plataformas poderiam alternativamente ser criadas em nível nacional ou internacional, onde poderia se esperar uma grande economia de escala, uma centralização de expertises e os serviços não necessitariam ser replicados em inúmeros lugares. O *UK Data Archives*[1] é um exemplo desse modelo nacional para as ciências sociais no Reino Unido. Para certos tipos importantes de dados e de outros produtos digitais de pesquisa existem plataformas internacionais com propósitos específicos. Essas plataformas de gestão proporcionam uma curadoria profunda e contínua, um alto grau de integração e uma conexão próxima com as demandas das comunidades disciplinares-alvo, tornando-se, dessa forma, sistemas referencias para seus respectivos campos de estudos. O *GenBank*[2], na área de genômica, assim como o *Protein Data Bank*[3] e o *UniProt*[4], são exemplos no escopo das biociências;

1 Disponível em: https://www.data-archive.ac.uk/. Acesso em: 25 abr. 2021.
2 Disponível em: https://www.ncbi.nlm.nih.gov/genbank/. Acesso em: 25 abr. 2021.
3 Disponível em: https://www.rcsb.org/. Acesso em: 25 abr. 2021.
4 Disponível em: https://www.uniprot.org/. Acesso em: 25 abr. 2021.

o *Space Physics Data Facility* (SPDF)[5] e o *Set of Identifications, Measurements and Bibliography for Astronomical Data* (SIMBAD)[6] estão no escopo das ciências espaciais (WILKINSON *et al.*, 2016). Estes sistemas referenciais oferecem dispositivos que assistem aos usuários humanos e máquinas, no acesso aos seus conteúdos de forma dinâmica e precisa, além de proporcionarem uma ampla gama de serviços.

Entretanto, nem todas as disciplinas acadêmicas são cobertas pelos vários centros nacionais e internacionais de dados especializados, atualmente em operação; nem é provável que cada tópico potencial de pesquisa disponha algum dia de uma plataforma específica; além do mais, nem todos os tipos de dados podem ser capturados ou submetidos a essas plataformas, posto que elas geralmente interpõem vários níveis de exigências para a publicação de dados. Todavia, muitos *datasets* importantes emergem de pesquisas tradicionais realizadas nas bancadas dos laboratórios e não se ajustam aos modelos de dados das plataformas de propósitos temáticos existentes e às barreiras interpostas. Nada obstante, esses conjuntos de dados não são menos importantes em relação à integralidade e à reprodutibilidade da pesquisa e às possibilidades de reuso (WILKINSON *et al.,* 2016), sendo assim, eles precisam ser gerenciados.

Portanto, neste cenário multifacetado, é preciso considerar que existem muitos pequenos grupos de pesquisa ou mesmo pesquisadores individuais, localizados na distribuição estatística conhecida como "cauda longa da pesquisa" (SALES; SAYÃO, 2018), que trabalham em diversos campos produzindo dados com características muito específicas e que têm requisitos que não são facilmente generalizáveis; ou áreas disciplinares que são tão estreitas para justificar o custo de se estabelecer e manter grandes centros de dados. Além disso, há as universidades, centros de pesquisa e outras organizações produtoras de conhecimento científico que desejam integrar suas coleções de dados às suas memórias acadêmicas por meio de plataformas de gestão de dados desenvolvidas em torno de repositórios institucionais.

Aparentemente em resposta a essa demanda, vão surgindo inúmeros repositórios multidisciplinares e de múltiplos propósitos, numa escala que vai de repositórios institucionais, por exemplo, pertencentes a uma única universidade, à repositórios abertos de escopo global tais como *Dataverse*, *FigShare*, *Dryad*, *Mendeley Data*, *Zenodo*, *DataHub*, DANS e *EUDat*. Estes repositórios aceitam um amplo espectro de tipos de dados que variam em termos de formatos, volume, modelos e estruturas. Observa-se também que eles não tentam integrar ou harmonizar os dados depositados e interpõem poucas restrições aos metadados assinalados na publicação dos dados. O ecossistema de dados resultante, portanto, parece afas-

5 Disponível em: https://spdf.gsfc.nasa.gov/. Acesso em: 25 abr. 2021.

6 Disponível em: http://simbad.u-strasbg.fr/simbad/. Acesso em: 25 abr. 2021.

tar-se da tendência relacional e está se tornando mais diverso e menos integrado, exacerbando, como consequência, os problemas de descoberta e reusabilidade para seres humanos, e muito mais para *stakeholds* computacionais (WILKINSON *et al.*, 2016). "Não obstante, são precisamente os tipos de análise integrativa, profunda e ampla que constituem a maior parte da *eScience*", concluem os autores (p. 3).

Dito de outra maneira, o investimento na construção de repositórios genéricos vem colocando em pauta outro desafio que é a oferta de serviços úteis que possam apoiar a *eScience*. Assim, a seção a seguir visa delinear os contornos de serviços de gestão de dados que possam ser oferecidos por meio de plataforma de gestão disciplinares.

5 Delineando os contornos dos serviços de gestão de dados

Como ratificado por Sara Jones, Grahan Prior e Angus White (2013, p. 5), "para dar apoio efetivo à gestão e ao compartilhamento de dados, uma instituição necessita de uma estratégia coerente e de um conjunto de serviços". Mas o que poderia significar este conjunto de serviços de gestão de dados? Naturalmente ele tem um espectro contínuo que varia em termos disciplinares, cultural e epistemológico, institucional e político, e ainda depende das bases tecnológicas disponíveis para a gestão de dados. De fato, as instituições de pesquisa podem oferecer serviços de dados numa grande multiplicidade, que não varia somente nos tipos de serviço, mas também na profundidade e alcance que esses serviços são disponibilizados, nos níveis de especificidade e comprometimento e para quem e com que objetivos esses serviços são oferecidos (CHOUDHURY *et al.,* 2018).

Para exemplificar, FEARON JR. *et al.* (2013) apontam que serviços de gestão de dados englobam o fornecimento de informações, consultoria, treinamento e ainda o envolvimento ativo no planejamento da gestão de dados, orientação durante a pesquisa (por exemplo, aconselhamento sobre o armazenamento de dados e segurança de arquivos), documentação e metadados, compartilhamento de dados de pesquisa e curadoria (seleção, preservação, arquivamento, citação) de projetos concluídos e dados publicados. Já na perspectiva de Choudhury e colaboradores (2018), os serviços de gestão de dados incluem a oferta de infraestrutura necessária para realizar a curadoria de dados por meio de licenças para preservação, análises e ferramentas de acesso; a disponibilidade de espaço em sistemas de armazenamento financiados pela organização para dados curados; treinamento e consultoria que permitam o pesquisador explorar os serviços de dados oferecidos pelas várias unidades da instituição. Complementando, Tang e Hu (2019) apontam que no diagrama de componentes de gestão de dados de pesquisa, as atividades abrangentes incluem "política e estratégias de gestão de dados" e "plano de negócios e sustenta-

bilidade". Subjacente ao estabelecimento de serviço de gestão de dados de pesquisa, vários níveis de orientação, treinamento e suporte são necessários. Para esses autores, o ponto focal do processo de gestão de dados deve dar proeminência aos componentes de serviço de gestão de planejamento, gerenciamento de dados ativos, seleção e compartilhamento, bem como repositórios e catálogos de dados. Neste sentido, os repositórios ou os catálogos de dados são apenas mais um serviço dentre inúmeros outros que uma plataforma de serviço de gestão de dados pode oferecer.

É importante observar que são muitas as diferenças entre a gestão de recursos mais tradicionais e o nível de exigências técnicas e de infraestruturas e expertises necessárias à gestão de dados de pesquisa. Um livro, por exemplo, tem uma catalogação universal e padronizada, as diferenças de tratamento entre disciplinas são poucas e seus processos estão focados na pós-publicação; o mesmo não se pode dizer de dados de pesquisa e de outros objetos digitais de pesquisa, como base de dados e códigos, cuja gestão tem que se preocupar com o longo e idiossincrático ciclo de vida que se inicia ainda na fase de planejamento - muito antes da publicação e arquivamento, indo até a pós-publicação, mas num processo ainda mais complexo do que era executado na gestão das publicações bibliográficas. Some-se a isso toda a peculiaridade própria que exige a articulação da gestão com o ciclo de vida do projeto de pesquisa. Neste contexto, o que se observa é que "O leque de competências e conhecimentos necessários para entregar serviços de gestão de dados é ditado em grande parte pelas fases individuais do ciclo de vida do projeto", confirmam Jones, Prior e White (2013, p. 3). Assim, a escala de serviços que as instituições de pesquisa oferecem pode variar não apenas nos tipos de serviços disponibilizados, mas também no nível de profundidade em que eles atuam, e no universo de usuários para quem os serviços são oferecidos (CHOUDHURY *et al.*, 2018). Pesquisadores, professores e estudante de pós-graduação são os clientes-alvo mais prováveis dos sistemas de gestão de dados, porém outros *stakeholders* devem ser considerados, como os gestores de C & T, financiadores e comunidades de práticas especificas – como engenheiros e agrônomos -, que reusam os dados, especialmente os dados com alto grau de processamento nos seus projetos e empreendimento, como na construção das fundações de uma usina nuclear ou na seleção de cultivares. Os serviços podem estar distribuídos por várias unidades da instituição ou concentrados e coordenados por uma unidade, possivelmente a biblioteca de pesquisa.

A visão fragmentada e heterogênea sobre os serviços de gestão de dados – que por fim reflete as múltiplas faces da atividade de pesquisa - cria um obstáculo no delineamento dos seus contornos e na enumeração do diagrama dos seus componentes. Possivelmente, uma racionalidade partindo dos pesquisadores na qualidade de usuários desses serviços pode ajudar na compreensão e na construção de um

possível conceito de serviços de gestão de dados. É o que este ensaio tenta fazer na seção seguinte, propor uma definição para serviço de gestão de dados de pesquisa onde se considere as necessidades específicas de seus usuários.

6 Afinal, o que são serviços de gestão de dados de pesquisa?

Num ambiente científico em que há uma intensa geração e consumo de dados pela odisseia cotidiana dos cientistas na busca por novos conhecimentos, há uma necessidade crítica de dispositivos para controle e organização desses ativos que, idealmente, se concretizam por um conjunto de serviços alinhados a uma estratégia coerente de gestão de dados. Nesse contexto em constante transição, a gestão de dados de pesquisa, em escala mundial, se torna cada vez mais um serviço essencial que deve ser intermediado pelas unidades de informação das organizações de pesquisa (TANG; HU, 2019).

A amplitude das ações proporcionadas pelos sistemas técnicos e gerenciais das plataformas de gestão de dados propicia aos diversos stakeholds, que orbitam em torno dos fluxos de geração de conhecimento científico, uma série de benefícios que dependem das suas matrizes de interesse, por exemplo: para um coordenador ou gestor acadêmico, a plataforma se configura em um dispositivo de avaliação de produtividade de um programa, área ou grupo de pesquisa; para um editor científico e seu corpo de revisores, a plataforma serve como um instrumento de avaliação e validação de registros submetidos; para os formuladores de políticas científicas e financiadores de pesquisa, as plataformas tornam-se uma cartografia que apoia a ordenação e níveis de incentivos necessários a uma área de estudo; e poderíamos dizer que para o cidadão comum e para a grande imprensa, como um elemento a mais de transparência dos investimentos públicos em ciência. Todavia, o pesquisador tem a percepção sobre os dados como um instrumento de pesquisa e de geração de novos conhecimentos, que tem um potencial transformacional, conforme enfatizam Sara Jones, Grahan Prior e Angus White (2013).

Assim sendo, a gestão de dados não é um fim em si mesmo e se concretiza aos olhos do pesquisador na forma de um amplo espectro de serviços e ferramentas que apoiam todo o ciclo de vida dos dados, no âmbito de um projeto de pesquisa, cujos benefícios são diretamente perceptíveis por eles, como por exemplo: maior visibilidade para a pesquisa; mais citações e prestígio, reconhecimento da autoria dos dados; maior nível de colaboração em escala global; organização dos dados para o próprio uso do pesquisador e de seus colegas próximos; reconhecimento em termos de promoção e financiamento; proteção lógica e física dos dados; e preservação de longo prazo. Isto indica fortemente que no âmbito da gestão de dados, a conexão orgânica com as comunidades científicas se concretiza via serviços resso-

nantes com a dinâmica e a cultura disciplinar, porém, sem perder a perspectiva de inserção global e interdisciplinar.

O estudo profundo sobre a avaliação da maturidade – um enfoque comum para a determinação do nível de sofisticação de serviços e produtos - dos serviços de dados conduzidos por Inna Kouper e colaboradores (2017) indicam que os serviços de gestão de dados mais avançados são provavelmente aqueles cujas ações privilegiam as necessidades das comunidades de instituições individuais, mas também estão cientes das comunidades de pesquisa mais amplas às quais os indivíduos se inserem. Os programas de serviços de gestão de dados de pesquisa mais consolidados não são necessariamente aqueles que oferecem as maiores carteiras de serviços ou que empregam as equipes maiores, "mas aqueles cujas atividades são mais profundamente conectadas à missão da biblioteca e da instituição como um todo [...], em outras palavras, onde os serviços são cuidadosamente escolhidos cuidadosamente, organizados, monitorados e otimizados", concluem os autores (KOUPER et al, 2017, p. 164).

Nessa diversidade de enfoques, os princípios FAIR oferecem uma base conceitual e um horizonte mais substantivo para o delineamento e consecução de serviços de gestão de dados, posto que eles estão focados em assegurar que os objetos de pesquisa sejam encontráveis, acessíveis, interoperáveis e reusáveis, sintetizando dessa forma os possíveis objetivos de um serviço de gestão de dados, ou, num nível de abstração mais elevado, o alcance de uma Internet de Dados e Serviços FAIR. Os princípios FAIR "descrevem características e aspirações aplicadas aos sistemas e serviços voltados para apoiar a criação de resultados valiosos de pesquisa que podem, então, ser rigorosamente avaliados e amplamente reutilizados (...)" (MONS et al., 2017, p. 50). Os princípios deliberadamente não especificam requisitos técnicos e não são padrões, mas constituem um conjunto de diretrizes orientadoras que estabelecem um continuum crescente de possibilidades de reuso dos objetos de pesquisa por meio de uma variedade de implementações diferentes. Nesta direção, os princípios FAIR descrevem as qualidades ou comportamentos dos recursos de dados para conseguir – possivelmente de forma incremental – um ótimo nível de descoberta e de reuso acadêmico, abrindo possibilidade para muitos e diferentes enfoques na concepção e na renderização de dados e de serviços.

Levando em conta os argumentos acima e tomando os princípios FAIR como núcleo agregador conceitual, propomos a seguinte definição para serviços de gestão de dados de pesquisa: É o conjunto de serviços informacionais, computacionais, científicos e administrativos oferecidos no âmbito da gestão de dados de pesquisa e ancorados nas necessidades específicas das comunidades acadêmicas e científicas, que têm como propósito tornar os dados localizáveis, acessíveis, interoperáveis e

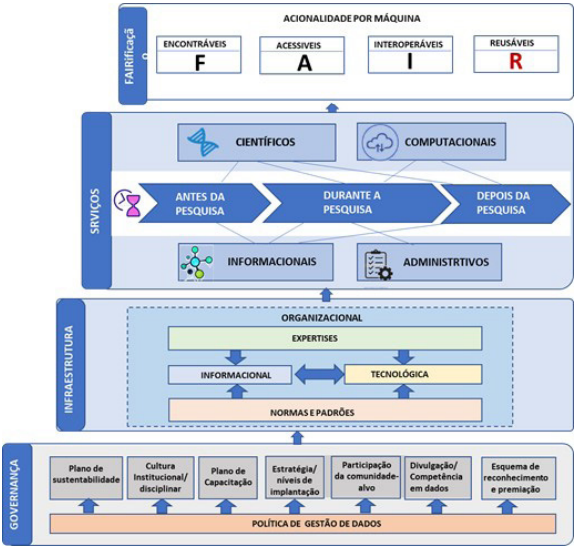
reusáveis, de forma que eles se traduzam em benefícios para a ciência e para todos os seus stakeholders.

A partir desta definição, a seguir propomos um modelo de serviço de gestão de dados que procura representar os diferentes aspectos da gestão de dados de pesquisa sem perder de vista a natureza interrelacionada da dinâmica das atividades que se desenrolam num ambiente científico intensivo de dados, cujo objetivo é torná-los FAIR.

7 Descrição de um modelo para a implementação de serviços de gestão de dados de pesquisa

Como uma abstração conveniente da realidade que se quer compreender, um modelo é uma criação cultural, um "mentefato" destinado a representar uma realidade ou alguns dos seus aspectos, a fim de torná-los descritíveis qualitativa e quantitativamente e, algumas vezes, observáveis (SAYÃO, 2001). A partir desse ponto, decidiu-se dividir o modelo em quatro camadas representacionais: 1) a governança, onde são discutidos os princípios norteadores do projeto de serviços de gestão de dados; 2) as infraestruturas técnicas, onde se incluem também as categorias de expertises necessárias; 3) os serviços informacionais, computacionais, científicos e administrativos; 4) os resultados da efetivação desses serviços manifestados pela FAIRificação dos dados. A figura 1 apresenta uma visão geral dos componentes agrupados em camadas e de suas interrelações, que a seguir são discutidos.

Figura 1 - Arquitetura para um Serviço de Gestão de Dados FAIR



Fonte: Os próprios autores

A arquitetura representada na figura acima é proposta com base na combinação dos modelos semânticos de agregação, de associação e classificação – em que apresenta as categorias gerais do modelo semântico de agregação em uma perspectiva *bottom-up*, a saber: governança, infraestrutura, serviços e fairificação. Essas categorias se associam a partir de uma relação de viabilidade, isto é, um componente viabiliza o outro. Dentro da categoria *Governança*, seus componentes se associam a partir de uma relação de direcionamento, mostrando como um elemento principal pode direcionar os demais componentes, tendo como parâmetro estruturante uma política de gestão. Já a categoria *Infraestrutura* revela a agregação de componentes necessários para a construção da infraestrutura da plataforma, que precisa estar permeada por uma infraestrutura organizacional, a saber: expertises, tecnologias, recursos informacionais e normas e padrões. Nesta categoria, mais uma vez em uma perspectiva *bottom-up,* observa-se as normas e padrões fundamentando as instancias tecnológicas e informacionais, que se relacionam com as expertises em uma perspectiva *top down,* mostrando quem as operacional. Na categoria *Serviços*, o modelo usado na construção da arquitetura é de agregação, mas os seus componentes se relacionam em uma linha temporal, isto é, os serviços científicos, computacionais, informacionais e administrativos se encontram na linha temporal que contorna todo o processo de pesquisa. Por último, a categoria *Fairificação*, também em um modelo semântico de agregação, mostra as categorias em que os princípios FAIR se organizam como componentes necessários para promover a fairificação e, em um nível mais elevado, um ecossistema de dados e serviços FAIR. As seções a seguir explicarão detalhadamente cada componente do modelo.

7.1 Governança de Dados de Pesquisa: planejamento, política, institucionalização e sustentabilidade

A configuração organizacional e institucional na qual a gestão de dados é realizada pode variar em relação a vários aspectos, como a intensidade de apoio à gestão e o nível de investimentos aplicados. Algumas instituições, como centros referenciais de dados científicos e agências estatísticas governamentais, podem estar inteiramente dedicadas à gestão de dados, tendo-a como finalidade principal. Em outras configurações, a gestão de dados é parte de uma atividade mais ampla que se conecta a outras atividades de pesquisa, como no caso das universidades (NATIONAL RESEARCH COUNCIL, 2015), cuja atividade de gestão de dados é decorrente de suas funções de ensino, pesquisa e extensão. Porém, mesmo no contexto acadêmico, são muitas as formas de planejar e executar as tarefas de gestão de dados, que variam de acordo com referências objetivas, como graus de investimento, sistemas técnicos disponíveis, volume e tipo de dados e de como a gestão de dados

está integrada aos seus fluxos de trabalhos e processos; e com percepções mais subjetivas, como cultura disciplinar e prestígio acadêmico. No presente modelo, esses parâmetros são equacionados por um nível mais administrativo compreendido pela categoria "governança de dados". Num plano mais conceitual, a governança de dados delineia os princípios, políticas e estratégias que são comumente adotados num ambiente que necessita de um programa de gestão de dados coerente; delineia também as ações, funções e papéis que são necessários para implementar essas políticas e estratégias. No âmbito de uma instituição de pesquisa, os princípios operacionalizados pela gestão governam todo o ciclo de vida dos dados – da conceitualização ao arquivamento e possível descarte. Assim sendo, o processo de governança de dados trata os dados não somente em seu aspecto espacial, mas também ao longo da sua dimensão temporal (SOLOMONIDES, 2019), este requisito implica uma ampliação do grau de complexidade e envergadura dos comprometimentos da governança.

Este arcabouço estruturante é necessário posto que dados de pesquisa digitais só podem ser gerenciados e preservados adequadamente ao longo do tempo por meio de um compromisso institucional sustentável (MAYERNIK *et al.,* 2012, p. 1). Em certa medida, a consolidação dos serviços de gestão de dados reflete o nível de aceitação organizacional incorporada a eles e o grau de planejamento das várias ações necessárias: orçamento sustentável em vigor, política de dados apropriada, conexão orgânica com as comunidades-alvo, conformidade com os códigos éticos e legais, alinhamento com os objetivos estratégicos institucionais e uma estratégia de desenvolvimento que considere os percursos possíveis para cada instituição. É necessário considerar também a inevitabilidade do fato de que as infraestruturas tecnológicas para acessar, interpretar e preservar a informação digital estão continuamente evoluindo; antecipar esses problemas e desenvolver estratégias para mitigá-los é uma atividade relevante para os compromissos de governança (NATIONAL RESEARCH COUNCIL, 2015). Sobre esses pilares podem ser desenvolvidos serviços avançados de dados que possam apoiar apropriadamente todo o ciclo de vida desses ativos informacionais, de acordo com os interesses dos vários *stakeholds* envolvidos. Considerando essas questões, propomos os seguintes enfoques como componentes do modelo:

- **Política de Gestão de Dados da Instituição** - Estabelece os fundamentos, diretrizes e compromissos da instituição concernentes à gestão, uso, propriedade, conformidade aos códigos éticos e legais, aderência às políticas das agências de fomento, políticas nacionais de ciência, tecnologia e inovação, às orientações e práticas internacionais e, por fim, mas de importância crítica, à

cultura, às práticas e às idiossincrasias das comunidades e domínios disciplinares. Uma política de gestão de dados de pesquisa abrangente deve também identificar as responsabilidades de cada um dos atores – biblioteca, laboratórios, tecnologia da informação, administração etc. - posto que a gestão de dados envolve diferentes setores da instituição (MUSHI; PIENAAR; DEVENTER, 2020) e o projeto é considerado como parte das atividades de pesquisa da instituição. É necessário enfatizar que o processo de desenvolvimento de uma política institucional de gestão de dados requer uma consulta extensiva a todos os agentes envolvidos e a aprovação das comunidades e organizações científicas relevantes (WILSON *et al.*, 2011). As orientações da política devem permear todo o ciclo da gestão. "Políticas podem ser um importante fator motivador para dados FAIR e outros objetos de pesquisa (*software*, *workflow*, modelos, protocolos etc.). Portanto é essencial que esforços "*botton-up*" baseados na comunidade sejam combinados com políticas com enfoque "*top--down*", completam Hong e seus colaboradores (2020).

- **Cultura Institucional/disciplinar** - A implantação de uma plataforma de serviços de gestão de dados de pesquisa deve ser precedida de uma análise de requisitos que considere o contexto e a cultura institucional, comunitária e disciplinar, bem como suas características únicas. Espera-se que este processo ajude a definir uma carteira de serviços de gestão de dados mais efetiva para apoiar as práticas de pesquisa da instituição e de suas comunidades (COATES, 2014; REED, 2015; MUSHI; PIENAAR; DEVENTER, 2020). É importante também reconhecer que algumas disciplinas necessitam de diferentes tipos de soluções técnicas para obter os mesmos benefícios dos dados FAIR (HONG *et al.*, 2020).
- **Plano de sustentabilidade –** Um dos grandes desafios de um programa de implementação e manutenção de uma infraestrutura de gestão de dados é assegurar que cada fase do projeto seja sustentável como um serviço contínuo ao longo do tempo (WILSON *et al*, 2011). Uma vez que a gestão de dados de pesquisa é reconhecida como algo necessário para as atividades de pesquisa, os seus custos devem ser estimados e suas fontes de financiamento – especialmente as perenes – identificadas. A escolha sobre onde investir os recursos, tipicamente limitados, e as projeções sobre as necessidades futuras se tornam essenciais (CHOUDHURY *et al.,* 2018). Desta forma, um projeto de implementação de serviços de gestão de dados de pesquisa precisa estar associado a um plano de sustentabilidade que delineie um comprometimento possível com o agora e com o futuro. A criação e o comprometimento com uma estratégia de longo prazo para os serviços podem revelar com mais cla-

reza os recursos necessários à continuidade dos serviços e das infraestruturas necessárias para tal. Isso, portanto, pode incluir, um plano de sucessão (MUSHI; PIENAAR; DEVENTER, 2020).

- **Divulgação/Competência em dados –** Para a implementação de um ambiente de pesquisa FAIR é necessário que as comunidades envolvidas desenvolvam uma compreensão compartilhada do que está circunscrito pelo conceito FAIR e pelos seus princípios. De uma forma geral, os pesquisadores e outros *stakeholders* têm baixo nível de percepção sobre a importância das práticas de gestão de dados e das exigências de gestão e compartilhamento das agências de fomento e dos compromissos de depósito dos dados firmados com os editores científicos, além das questões éticas e legais envolvidas na publicação dos dados. Por exemplo, em relação ao conceito FAIR, Hong *et al.* (2020) observam que o pesquisador não sabe o que é dado FAIR e muitas vezes acha que é o mesmo que dado aberto. Isto indica que é necessário planejamento e ações de divulgação e de conscientização que tragam à tona essas questões. Um programa de divulgação nesta direção deve contemplar a elaboração de material didático (cartilhas e guias), cursos, eventos, oficinas, entre outros.
- **Conhecimento/participação da comunidade-alvo –** Como criadores e usuários de dados de pesquisa, o engajamento dos pesquisadores é crucial no desenvolvimento de serviços de gestão de dados. O provisionamento de qualquer serviço precisa ser baseado numa compreensão próxima dos padrões e fluxos das pesquisas que se desenvolvem na instituição, das suas motivações, características e prioridades. Assim sendo, a definição precisa dos requisitos dos serviços necessita ser estabelecida com o comprometimento e a contribuição da comunidade de pesquisadores, sem essas considerações, as características dos serviços podem não estar em harmonia com os objetivos dos pesquisadores. A comunidade deve ser acompanhada nas mudanças de interesse sobre os dados, e a participação dela no desenvolvimento e escolha de padrões compartilháveis para as práticas e para as infraestruturas FAIR deve ser reconhecida e institucionalizada. A proximidade, interação e alinhamento das comunidades com as organizações nacionais e internacionais que lidam diretamente com a gestão de dados FAIR, como GO FAIR, RDA, CODATA, DCC e outras, devem ser incentivados.
- **Plano de capacitação –** Para oferecer serviços completos em gestão de dados, as bibliotecas precisam ter pessoal tecnologicamente qualificado ou aumentar muito as oportunidades de treinamento tecnológico para o pessoal existente. (TENOPIR; BIRCH; ALLARD, 2012). Sustentabilidade humana é crítica para assegurar a continuidade e a consistência da oferta de serviços

ao longo do tempo. Entretanto, poucos programas formais em estudos informacionais incluem em seus currículos gestão de dados, desta forma, os gestores de dados de pesquisa são normalmente treinados em serviço nas disciplinas específicas onde trabalham (BORGMAN, 2007, p.155).

- **Estratégia/níveis de implantação –** O desenvolvimento e a implantação de infraestrutura de gestão de dados, além de muitos recursos, requerem tempo para alcançar sua plena maturidade e espelharem as demandas das comunidades científicas, isto implica a necessidade do estabelecimento de níveis de implantação de infraestruturas e serviços. As bibliotecas de pesquisa, por exemplo, têm, em muitos casos e de forma proativa, procurado suprir as necessidades de gestão de dados para as suas comunidades de usuário. Frequentemente isso acontece sem aporte financeiro adicional destinado ao desenvolvimento e disponibilização de serviços de dados. Assim sendo, as bibliotecas têm que começar numa escala mais simples, construindo uma base sobre a qual possa desenvolver serviços mais sofisticados (ERWAY *et al*, 2016), começando com serviços básicos que exijam apenas recursos da própria biblioteca, até alcançarem serviços mais complexos que exijam alto nível de compromisso institucional e mais recursos financeiros, tecnológicos e humanos (KOUPER *et al.,* 2017).
- **Reconhecimento e premiação –** A gestão de dados de pesquisa consome tempo, recursos e exige grande dedicação do pesquisador, entretanto, esse esforço raramente é reconhecido pelos sistemas de recompensa acadêmicos, exceto quando linkado com publicações em periódicos científicos. Portanto, para incentivar essa nova tarefa dos pesquisadores e destacar sua importância, é essencial que ela seja apropriadamente reconhecida e que seja considerada nos critérios de avaliação, promoção e de financiamento.

7.2 Infraestruturas de Dados de Pesquisa

Infraestrutura é uma noção de grande amplitude e multidimensional. Ela pode ter uma conotação técnica, legal, organizacional e, em muitos casos, é imprescindível considerar também os aspectos sociais, culturais e políticos. De fato, é assim no domínio da ciência: o projeto de infraestrutura de pesquisa é simultaneamente uma questão tecnológica, uma questão de identificação das necessidades da pesquisa em áreas disciplinares específicas e uma questão política. Essa ótica mais geral se aplica às infraestruturas institucionais de gestão de dados de pesquisa que precisam oferecer tecnologias e ferramental, processos, políticas, recursos e treinamento para os vários e diversificados estágios da gestão de dados.

De fato, da mesma forma que as instituições devem providenciar infraestruturas básicas para a pesquisa – tais como laboratórios, instrumentação, computação de alto desempenho, redes, reagentes e muito mais – elas devem também tomar medidas para uma gestão adequada dos dados. Isto pressupõe um amplo espectro de atividades gerenciais, tecnológicas e informacionais que inclui profissionais de informação treinados para apoiar pesquisadores no planejamento e gestão de seus dados, no acesso a dispositivos de armazenamento seguro e *backups* durante o desenvolvimento do projeto e disponibilidade de plataformas de acesso e de preservação de longo prazo, necessárias após o fim da pesquisa (STRASSER, 2015). É imprescindível também um corpo de normas, padrões e boas práticas que permitam, principalmente, uma interlocução em níveis variados dos sistemas e serviços, tanto local quanto global, que pode ser traduzida por interoperabilidade. Nesta categoria, à guisa de exemplo, estão os **padrões de modelo de dados** - geralmente estabelecidos por um domínio disciplinar ou repositório – que determinam a estrutura dos vários componentes de uma coleção de dados, que, por fim, têm efeito sobre as interfaces de interação com os usuários humanos e computacionais e sobre os níveis de interoperabilidade do *dataset* (CHOUDHURY *et al.,* 2018).

Quando comparamos a publicação acadêmica tradicional com a publicação de dados verificamos que as infraestruturas subjacentes à publicação acadêmica criam uma ponte epistemológica entre disciplinas, tendo como ponto agregador as bibliotecas de pesquisa que selecionam, coletam, organizam e tornam acessíveis publicações de todo o tipo e de todas as áreas. Por sua natureza, as instituições sociais trabalham para estabilizar práticas particulares e formas de conhecimentos. Em certo sentido, as instituições são infraestruturas sociais em si mesmas. Nessa direção, as infraestruturas técnicas estão entrelaçadas com as infraestruturas sociais das instituições, muitas vezes mediadas por padrões, protocolos, documentos e artefatos que ligam os aspectos sociais e técnicos das infraestruturas (LEONARDI, 2010). Entretanto, não existe ainda infraestrutura dessa magnitude para os dados. Algumas poucas áreas têm mecanismos consolidados para publicar dados; outras estão nos estágios de desenvolvimento de padrões e práticas para agregar seus dados e torná-los amplamente acessíveis. Um problema-chave nas instituições de pesquisa, como observam Mayernik e seus colaboradores (2012, p.158), "é a falta de uma infraestrutura confiável que possa ser implantada num nível institucional", essa "falta de infraestrutura para dados amplifica a descontinuidade na publicação acadêmica", acrescenta Borgman (2007, p. 155).

Os arcabouços infraestruturais voltados para a gestão de dados são diversos e fragmentados em termos de fluxos, complexidade, aplicação e topologia, e organizados de forma diferente pelas várias disciplinas e em diferentes países (GRAAF;

WAAIJERS, 2011). Contudo, crescentemente as infraestruturas moldam os padrões e as práticas da gestão de dados. Diante desse fato, o conhecimento sobre a origem, domínio disciplinar, grau de processamento, sistemas de coleta, *workflows* etc. parecem ser de importância crítica na concepção de infraestruturas voltadas para a gestão de dados (SAYÃO; SALES, 2020).

Na presente proposta de modelo, consideramos cinco instâncias de infraestruturas necessárias à implantação de sistemas de gestão de dados: de padronização, tecnológicas, informacionais, profissionais e organizacionais.

- **Normas e padrões** - Normas e padrões são formas consensuais de codificar o conhecimento que circula transversalmente por comunidades para assegurar uniformidade e similitude aos seus produtos e processos através do tempo e do espaço. Eles refletem o conhecimento mais atual sobre as práticas profissionais e aumentam a interoperabilidade, a consistência, a preservação, a reusabilidade, a segurança e a proteção das coleções digitais. Portanto, assegurar que em um ecossistema científico, em que as infraestruturas estão globalmente dispersas, seus produtos se alinhem aos Princípios FAIR, tenham um grau satisfatório de qualidade e excelência e sejam apropriados às necessidades dos pesquisadores, exige um corpo de padrões e princípios amplamente adotados e compartilhados. Considerando este fato, propõe-se que um corpo consensual de normas e padrões consubstanciem infraestruturas que devem estar subjacentes aos processos de gestão de dados. Isto porque espera-se que as coleções de dados estejam aptas para serem utilizadas para uma grande variedade de propósitos – e não somente para as finalidades para as quais elas foram inicialmente coletadas. Para tal, elas precisam ser agregadas a outras coleções em outros sistemas, compartilhadas, acessadas, analisadas e arquivadas usando um amplo espectro de tecnologias. Essa condição torna um corpo de normas e padrões comuns infraestrutura essencial para a gestão e curadoria de dados de pesquisa. À medida que os princípios e práticas da gestão de dados de pesquisa se desenvolvem, eles começam a adquirir reconhecimento como um campo de conhecimento distinto e chamando a atenção de organizações interessadas no seu aprimoramento como, por exemplo, DCC, Codata, GO-FAIR, DataOne, DataCite, entre muitas outras. Nesta direção, padrões e normas comumente adotados para a gestão de dados estão tomando corpo em muitas disciplinas e setores diferentes e estão sendo redefinidos em outras disciplinas. Como resultado, práticas aprimoradas para garantir a qualidade e a durabilidade dos dados digitais estão sendo continuamente estabelecidas. (NATIONAL RESEARCH COUNCIL, 2015).

- **Infraestrutura Tecnológica** – Compreende um vasto conjunto de atividades, equipamentos, processos e expertises que possam viabilizar os requisitos tecnológicos operacionais necessários às ciberinfraestruturas de gestão de dados, tais como: organização lógica, física e virtual dos dados; dispositivos para processamento de alto desempenho, computação em grade e armazenamento das coleções de dados locais ou em nuvem; redes locais, comunicações, conexões externas, internet, serviços *web*; aquisição/desenvolvimento de códigos científicos, s*oftware* de *workflow*; equipamentos para análise de dados e visualização, conexões, estratégias de segurança física, lógica e de rede.
- **Infraestrutura Informacional –** Compreende todo o arcabouço conceitual e teórico materializado nas práticas da Ciência da Informação, Biblioteconomia e Arquivologia, que são plenamente aplicadas à gestão de dados, como seleção, catalogação, indexação, classificação e descarte e os instrumentos e dispositivos tecnológicos que viabilizam essas práticas como: esquemas de representação e identificação persistentes; metadados descritivos, técnicos, administrativos, de preservação e disciplinares; tesauros, vocabulários controlados, taxonomia, ontologias, esquemas de classificação; bases de dados, repositórios e bibliotecas digitais e plataformas confiáveis para o arquivamento de longo prazo.
- **Infraestrutura de pessoal –** As inúmeras instituições de pesquisa desenvolvem os mais diversos enfoques de gestão de dados. Isto pressupõe equipes de apoio compostas por diferentes profissionais (PINFIELD; COX; SMITH, 2014). Papéis como administrador de dados e cientistas de dados estão emergindo no mundo da ciência contemporânea e se incorporando às equipes mais tradicionais compostas por pesquisadores, técnicos de laboratório, assistentes de pesquisa e analistas; por outro lado, no âmbito das bibliotecas especializadas e dos repositórios, novos atores como bibliotecários e arquivistas de dados e curadores fazem a conexão entre a biblioteca e os laboratórios e apoiam a gestão das idiossincrasias disciplinares dos ciclos de vida dos dados (BALL, 2012). Entretanto, um requisito essencial - especialmente quando se trata dos serviços associados à curadoria - é a necessidade de conhecimento das disciplinas e domínios nos quais os dados são coletados, processados e utilizados. Sem alguma familiaridade com o problema a ser abordado, a cultura disciplinar, os objetivos a serem perseguidos, bem como os métodos utilizados, nomenclatura e práticas dos campos em que os ativos digitais são usados, os curadores não serão capazes de tomar as decisões mais corretas para gerenciar esses ativos para uso atual e futuro (NATIONAL RESEARCH

council, 2015). As equipes de gestão precisam dos papéis relacionados a seguir, que podem ser desempenhados, cada qual, por profissionais distintos ou acumuladamente - de forma mais próxima à realidade das instituições de pesquisa - por equipes menores.

- **Pesquisadores –** personagem mais envolvido com a pesquisa e com os dados; como autor/criador/coletor dos dados/avaliador devem assegurar que os metadados disciplinares, registro dos dados (proveniência), documentação, contexto e qualidade estejam em conformidade com os padrões da comunidade/instituição.
- **Bibliotecário de dados –** profissional de **Biblioteconomia** com formação em gestão de dados; cataloga, indexa, organiza, apoia a publicação dos *datasets*; assessora o planejamento e a operacionalização dos repositórios e dos serviços de gestão de dados; apoia a curadoria por meio da construção de instrumentos de representação e padronização; idealmente conhece os fluxos de pesquisa de sua instituição; promove cursos, divulgação e material didático e assessora os pesquisadores na elaboração do Plano de Gestão de Dados (pgd).
- **Arquivista de dados –** profissional de **Arquivologia,** responsável pelo arquivamento e preservação de longo prazo dos dados e garantia de integridade, autenticidade e confiabilidade. Apoia o planejamento de sistemas de arquivamento confiáveis.
- **Cientista de dados –** profissional da área de **Ciência da Computação** e/ou da área disciplinar que contribui no desenvolvimento de tecnologias de análise, manipulação, visualização, modelagem, algoritmização e aplicação de metodologias avançadas, como inteligência artificial e aprendizagem de máquina.
- **Gerente de dados –** profissional da área de **Tecnologia da Informação** responsável pela manutenção e implementação de bases de dados, repositórios, sistemas de armazenamento; apoia a segurança, *backups*, checagem de integridade.
- **Curador de dados – pesquisador ou profissionais de informação** com conhecimento disciplinar que adiciona valor aos dados por meio de documentação, metadados, identificadores, contextualização, integração, reformatação, *mashup* etc.; promove o compartilhamento e o reuso; apoia a avaliação para a preservação e a criação de serviços.
- **Gestor – administrador de C&T** que compreende a importância dos dados no âmbito institucional, nacional e internacional; apoia a definição de

políticas, negocia recursos junto às agências de fomento, implanta e-infraestruturas e adquire ferramentas, equipamentos, software e coleção de dados.

- **Infraestrutura organizacional** – O arcabouço infraestrutural pressupõe, assim como a governança, uma ancoragem baseada em alguma estrutura organizacional voltada para a pesquisa, como uma universidade, instituto de pesquisa, ou mesmo uma empresa cujos empreendimentos dependem da gestão de dados. Estas organizações precisam oferecer tecnologias e ferramental, processos, políticas, recursos e treinamento para os vários e diversificados estágios da gestão de dados.

Essas vertentes infraestruturais - que possibilitam imbricamento de saberes e práticas que estão subjacentes a equipamentos, instalações, metodologias e principalmente a pessoas - proporcionam uma vasta carteira de serviços, ferramentas e processos que continuamente levam os objetos de pesquisa para um alinhamento com os Princípios FAIR.

O que se percebe até aqui é que para se chegar na categoria que representa os serviços de gestão de dados de pesquisa, faz-se necessário o estabelecimento de uma governança de dados eficiente e a construção de uma infraestrutura adequada à formulação de serviços. Assim, o item a seguir descreve a categoria que é implicitamente o coração do modelo.

7.3. Serviços de Gestão de Dados de Pesquisa

Na construção do presente modelo, consideramos uma matriz de serviços baseados em dois eixos principais: um eixo temporal, que considera o desenrolar dos serviços de dados ao longo do tempo interligando o ciclo de vida dos dados ao ciclo de vida da pesquisa; o segundo eixo considera o ponto de ancoragem dos serviços, significando que eles podem estar fundamentados em processos informacionais, computacionais, científicos ou administrativos. Fica claro que esses contornos não são sempre bem definidos e as sobreposições estão presentes em ambos os eixos, o que demonstra a necessidade de interconexão de várias expertises para a consecução das atividades de gestão de dados de pesquisa. Do ponto de vista temporal, podemos considerar que a atuação da gestão na forma de serviços se efetiva em três momentos (JONES; PRIOR; WHITE, 2013):

- **Antes da pesquisa começar** – fase de planejamento e conceitualização dos dados onde é enfatizado a assistência à preparação do Plano de Gestão de

Dados, incluindo o suporte ao uso de ferramentas *on-line*, orientação sobre custos das atividades de gestão e conhecimento sobre os recursos, serviços e ferramentas disponíveis pela instituição, para a gestão de dados.

- **Durante a pesquisa** – compreende um vasto conjunto de atividades informacionais, computacionais e científicas que inclui orientação sobre documentação, formatos e padrões que potencializam a encontrabilidade, a interoperabilidade e o reuso; orientação sobre armazenamento, gerenciamento e análise de dados; aconselhamento e/ou fornecimento de sistemas de armazenamento e segurança que atendam às necessidades de uma ampla gama de tipos de dados, plataformas e necessidades de acesso.
- **Depois que a pesquisa finalizar** – esta última fase abriga as questões de arquivamento e de acesso e preservação de longo prazo; seleção das coleções de dados de valor contínuo; suporte para tornar os dados de pesquisa disponíveis para audiências específicas; e orientação aos pesquisadores sobre como arquivar seus dados no final do projeto.

Para atender o amplo espectro das necessidades de gestão de dados, materializadas em serviços específicos, é necessário a colaboração de diversas áreas e a integração de expertises, infraestruturas, práticas e metodologias. Mesmo identificando que há sobreposições importantes – e desejáveis -, consideramos quatro tipos de serviços: serviços científicos, serviços computacionais, serviços informacionais e serviços administrativos.

7.3.1 Serviços científicos

Compreendem os serviços que se desenrolam em ambientes predominantemente científicos, como laboratórios e centros de pesquisa, e que são executados por cientistas, acadêmicos ou especialistas em gestão de dados com profundos conhecimentos disciplinares. São serviços relacionados à preparação de dados para usos mais amplos e podem incluir atividades como avaliação, limpeza, normalização, transformação, organização dos arquivos, nomeação e, quando necessário, anonimização e outras estratégias para a preservação da privacidade, indexação disciplinar; documentação de códigos, *workflow* e processamento, agregação de dados. Mesmo considerando que esses serviços são protagonizados pelos próprios pesquisadores, eles precisam de considerável suporte computacional e informacional e, algumas vezes, administrativo.

- **Descrição disciplinar** – para determinar se um *dataset* é realmente útil para um projeto de pesquisa, a informação descritiva (metadados e do-

cumentação) que o acompanha deve ser rica e extensiva permitindo que ele seja encontrado e interpretado, nessa direção várias camadas de descrição podem ser aplicadas (CHOUDHURY *et al.*, 2018). Há um consenso claro de que os pesquisadores são os profissionais mais bem posicionados para descreverem os dados que eles coletam ou produzem, posto que eles compreendem com profundidade os processos pelos quais os dados foram derivados, processados, limpos, o contexto em que eles foram agregados e as limitações e fragilidades que possuem e que podem não ser aparentes para outros pesquisadores que desejam reutilizá-los no futuro (WILSON *et al.*, 2011). Assim, o serviço deve oferecer ferramentas e instrumentos terminológicos disciplinares – metadados, ontologias, taxonomias - que permitam os pesquisadores descrever seus dados a partir do momento da sua criação (MARTINEZ-URIBE, 2019); inclui também metodologias para o empacotamento de dados, metadados e documentação**.** Algumas disciplinas se utilizam do conceito de **dicionário de dados** que contém informação, tais como, significado dos dados, relacionamentos com outras coleções, origem, usos e formatos (STRASSER, 2015). É importante salientar que uma parcela do conjunto de metadados é assinalada automaticamente pelos instrumentos científicos, exemplo, data, hora, geolocalização, temperatura etc., constituindo o conjunto de **metadados intrínsecos** em contraste com os metadados assinalados pelos pesquisadores, também chamados de **metadados contextuais** (MONS *et al*., 2017); os **metadados descritivos** (autoria, título, data de publicação etc.) são geralmente assinalados pelos profissionais de informação.

- **Avaliação (Appraisal) das coleções de dados** – um dos maiores desafios em relação aos dados de pesquisa é decidir quais as coleções que precisam ser mantidas para o futuro e por quanto tempo (MARTINEZ-URIBE, 2019). Pesquisadores ou especialistas em dados devem avaliar e selecionar as coleções que serão arquivadas por longo prazo de acordo com orientações bem documentadas, políticas ou exigências legais. Como resultado da avaliação, alguns dados podem ser transferidos para outro custodiante ou para **destruição segura**; os dados considerados de valor contínuo são idealmente submetidos a um arquivo de dados, centro de dados, repositório ou a algum outro serviço equivalente (BALL, 2012). A avaliação também se faz necessária quando existem diversos conjuntos de dados a serem tratados e uma ordem de prioridade precisa ser estabelecida.
- **Limpeza dos dados** – consiste no processo de eliminação ou edição de parte dos dados que estão corrompidos ou sem a acurácia desejada, com o objetivo

de alcançar o nível conveniente de integridade e qualidade para a coleção de dados (SAYÃO; SALES, 2015).

- **Organização dos dados –** consiste na organização dos dados em coleções, pastas, diretórios etc., nomeados apropriadamente, convencionados à priori e registrados por meio de documentos, como por exemplo, o arquivo "leia-me". Para uma organização consistente, o uso de taxonomias com princípios classificatórios bem definidos se faz necessário.
- **Transformação** – consiste na reformatação ou criação de subconjuntos de dados ou de outra derivação da coleção de dados para reuso por pesquisadores (BALL, 2012).
- **Documentação do processamento** – os dados de pesquisa raramente são usados logo que são coletados ou gerados por um instrumento, geralmente eles passam por vários estágios de processamento que os tornam mais adequados às finalidades que se propõem. A publicação de uma descrição das etapas de processamento oferece um contexto para a interpretação e reuso dos dados e confere evidências sobre a **proveniência** dos dados (GOODMAN et al., 2014). Essa documentação pode incluir a descrição dos códigos usados na geração e processamento dos dados e do software de workflow que controlam e registram as várias etapas computacionais e de manipulação dos dados. Em algumas ocasiões, a documentação pode ser o próprio artigo onde o autor relata a metodologia de coleta e os resultados da pesquisa, ou ainda um documento textual adicional ao conjunto de dados ou um artigo de dados publicado pelo pesquisador em um periódico convencional ou periódico de dados (TORINO; ROA-MARTÍNEZ; VIDOTTI, 2020)
- **Anotação** – O reuso de dados e de outros objetos de pesquisa exigem níveis elevados de colaboração. Neste sentido, os dados de pesquisa estão crescentemente localizados em plataformas que permitem novas formas de comunicação e colaboração entre acadêmicos, entre elas está a possibilidade dos pesquisadores adicionarem informações aos dados, enriquecendo-os. Este tipo de colaboração é conhecido como anotação e pode ser aplicada a todos os tipos de dados com o propósito de descrever, corrigir, interpretar, estender ou classificá-los. "Anotação é uma parte essencial da prática acadêmica [...] permitindo que o conhecimento seja organizado, compartilhado, construído e reusado" (HARVEY, 2010, p. 208). O uso de vocabulários controlados ou ontologias se faz necessário para que as anotações possam ser comunicadas a outros pesquisadores e processáveis por máquina.
- **Documentação dos códigos** – O código (*software*) usado para criar ou processar os dados, em muitos casos, é um componente essencial para viabilizar

o seu uso e reuso, e a documentação sobre o código é de suma importância para a compreensão dos dados e de como os resultados foram obtidos. Isso significa dizer que o código e sua documentação devem ser pensados como parte do pacote de informações que descreve os dados e, idealmente, uma cópia e uma descrição do código devem estar incluídas no pacote depositado num repositório. É preciso enfatizar que os padrões que orientam a publicação de dados estão evoluindo de forma distinta nos vários domínios disciplinares, admitindo a submissão de um amplo espectro de objetos de pesquisa; nesse contexto, os *softwares* desempenham vários papeis importantes no desenvolvimento de experimentos científicos, especificamente em relação aos dados de pesquisa, porém, e em alguns casos, o *software* é o principal produto de dados, como por exemplo, na concepção de um novo algoritmo (GOODMAN *et al*, 2014).

- **Documentação dos *workflows*** – a combinação dos métodos de coleta dos dados, processamento e análises de um experimento é chamada de *workflow*, que minimamente indica como os dados intermediários, os produtos e os resultados finais são gerados. Em muitos casos, os pesquisadores utilizam pacotes de *software* de *workflow* para executar experimentos e registrar o que foi realizado. As informações usadas e capturadas pelo *workflow* fazem parte da **proveniência** dos dados, bem como o software de *workflow*, sua versão e as configurações utilizadas.
- **Análise de dados** – o serviço compreende a exploração, extração e validação de novos relacionamentos ou características de um corpo de dados (NATIONAL RESEARCH COUNCIL, 2015). Na medida em que os dados são processados – transformados, normalizados, integrados - e descritos, inúmeros tipos de análises podem ser realizadas sobre eles: análises quantitativas, qualitativas, visualizações, entre outras. A análise de dados é frequentemente realizada por meio de pacotes de *software* especializados, incluindo ferramentas estatísticas, *data analytics*, *de-identification*, processamento de linguagem natural, mineração de dados, aprendizado de máquina, algoritmos, técnicas de amostragem, desenvolvimento e teste de hipóteses. Os métodos emergentes, como a inteligência artificial, permitem análises mais refinadas, especialmente em dados não estruturados. O tipo de análise aplicada aos dados geralmente necessita de habilidades específicas, tendo como ponto de partida o conhecimento sobre o uso do pacote de *software*. Cada ferramenta de análise tem uma curva de aprendizagem, via de regra, as técnicas mais avançadas exigem conhecimentos especializados profundos (CHOUDHURY *et al*., 2018). Preparar os dados para a fase de análise normalmente exige conhecimentos

de programação, base de dados e expertise em manipulação e edição de arquivos de dados, transformação e obtenção de saídas em diversos formatos. Estas operações que são executadas para a preparação dos dados têm que ser armazenadas e/ou descritas com o objetivo de documentar os processos e contextos e propiciar níveis apropriados de reprodutibilidade. A integridade desses processos é, naturalmente, uma preocupação e uma matéria para o pesquisador, mas que necessita de um apoio da área de computação em termos de treinamento e assistência específica, especialmente dos cientistas de dados.

- **Apresentação e visualização de dados** – esses serviços são tipicamente compreendidos como o produto ou saída da análise de dados, são, porém, de grande importância no contexto dos serviços das plataformas de dados, posto que podem revelar uma compreensão mais acessível para um amplo espectro de interessados sobre os dados gerenciados. Envolve conhecimento sobre técnicas de visualização e apresentação, design e contextualização de informação e avaliação de produtos, algoritmos e pacotes de software (NATIONAL RESEARCH COUNCIL, 2015).
- **Empacotamento dos dados –** para que os dados sejam encontrados por seres humanos e computadores e sejam compreendidos, agora e no futuro, é preciso que os arquivos de dados (por exemplo, planilhas) estejam fortemente acoplados a metadados e à documentação de apoio, formando uma unidade conceitual identificada chamada de **pacote de dados**, que pode ser preparado para ser depositado em um repositório ou centro de dados. Os exemplos variam de um único documento informacional, resultado de uma pasta *zippada,* a objetos complexos padronizados, no âmbito de uma comunidade científica, e legíveis por máquina (TANG; HU, 2019).

Como mencionado, os serviços a serem ofertados podem se dividir em diversas categorias. Além dos serviços científicos aqui instanciados, os serviços computacionais constituem outra categoria de serviços necessários, que está descrita a seguir:

7.3.2 Serviços computacionais

A transição entre uma ciência fechada e autocontida para uma ciência mais aberta, distribuída em rede e cooperativa, pressupõe mudanças profundas na infraestrutura computacional necessária à condução das atividades de pesquisa, sintetizada pelo termo "ciberinfraestrutura de pesquisa". Este fato pode ser expresso pela demanda crescente de suporte computacional para a publicação de dados FAIR,

análises integrativas avançadas, inteligência analítica (*analytics*), máquinas virtuais, sistemas de *workflow* etc. Além do mais, subjacentes aos Princípios FAIR, há uma ênfase especial no conceito de "acionabilidade por máquina de dados e metadados", isto reque que os recursos que desejam cumprir ao máximo as diretrizes FAIR devem utilizar um arcabouço tecnológico amplamente aceito que viabilize a legibilidade por máquina de representação de dados e conhecimentos (MONS *et al.,* 2017).

Considerando esse contexto, os serviços compreendem a oferta de ferramentas de *software* e equipamentos de computação para apoiar o processamento, análise e visualização dos dados de pesquisa; apoiar os processos de interoperabilidade e acionamento por máquina de dados e metadados; prover orientação de como os dados podem melhor ser estruturados e armazenados e trabalhar, se necessário, junto aos pesquisadores na estruturação de bases de dados e marcação de texto (WILSON *et al.*, 2011); os serviços podem incluir ainda treinamento específico para a equipe de pesquisadores nos recursos oferecidos e, em situações mais avançadas, oferecer processamento de alto desempenho, armazenamento em nuvem de grandes volumes e computação em grade.

- **Sistema de armazenamento** – As exigências das agências de fomento têm aumentado a conscientização dos pesquisadores quanto à necessidade de armazenar os dados de forma segura, todavia, em muitos casos, os pesquisadores armazenam em seus próprios computadores e criam sistemas informais particulares de armazenamento. Isso acontece principalmente nos estágios iniciais de um projeto de pesquisa, como durante a coleta dos dados (CHOUDHURY *et al.*, 2018). Para minimizar esse problema, os sistemas de *storage* oferecem serviços críticos para a gestão de dados, colocando à disposição dispositivos de armazenamento para o amplo espectro de conjunto de dados gerados ou utilizados pelas instituições, numa escala que dê atenção aos usos correntes, mas que também antecipe os requisitos futuros das atividades de pesquisa das diversas equipes de pesquisadores (PINFIELD; COX; SMITH, 2014). Com esse horizonte, os sistemas de armazenamento performam várias ações de curto e longo prazos para garantir que os dados permaneçam seguros e íntegros, agora e no futuro, que incluem: apoio aos processos de *backups* e controles de versões; controle de acesso físico; manutenção do *hardware* de armazenamento; checagem de fixidade e atualização (*refreshing*) e migração de mídias entre outras facilidades (BALL, 2012). Os requisitos de desempenho dos sistemas de armazenamento podem variar em virtude dos níveis de utilização dos dados, por exemplo: *datasets* volumosos que serão ativamente analisados precisam de sistemas de arquivos de alta performance endereçados

por diferentes camadas de armazenamento. Essas camadas de armazenamento podem ir de sistemas de arquivo paralelo de alto desempenho, que podem ser acessados por ambientes analíticos avançados, a sistemas com alta latência para acesso a *datasets* raramente usados (CHOUDHURY *et al*., 2018). A criação e gestão de ambientes de armazenamento exigem assistência de profissionais de computação com expertise nesse domínio técnico, isto porque a complexidade do ambiente de armazenamento local ou em nuvem pode ser "intimidante" para o pesquisador e exigir conhecimentos e ferramentas avançadas.

- **Proteção de dados sensíveis** – assegurar que os dados, especialmente aqueles classificados como confidenciais ou sensíveis, estejam mantidos seguros, anonimizados, com autenticação apropriada e mecanismos de autorização válidos (PINFIELD; COX; SMITH,, 2014).
- **Normalização de formatos** – nem sempre os formatos de arquivo gerados pelos instrumentos científicos são os mais adequados para a publicação, disseminação e, principalmente, para a preservação de longo prazo, posto que são formatos proprietários e muito específicos (SAYAO; SALES, 2015). Assim, alguns dados podem necessitar ser migrados para formatos diferentes do original, para adequá-los às regras do sistema de arquivamento, seja para facilitar a gestão, seja para mitigar os riscos de obsolescência tecnológica, ou ambos. Nessa direção, este serviço apoia os pesquisadores sobre os formatos, práticas e ferramentas de conversão mais adequados para produzir e documentar dados específicos; este serviço pode também oferecer suporte para projetos de bases de dados (MARTINEZ-URIBE, 2019) e de outras formas de estruturação dos dados.
- **Serviços de** backup - realizar cópias de segurança é uma ação necessária para qualquer projeto de pesquisa envolvendo dados de pesquisa. A estratégia apropriada deve ser acionada pelo serviço, que pode considerar vários parâmetros: *backup* automático ou manual, testes e verificações, frequência, tempo de armazenamento, responsabilidades etc. Alguns dos processos contínuos e mais simples de *backup* podem ser conduzidos pelos próprios pesquisadores, mas à medida que o volume e a complexidade dos dados aumentam, eles precisam de apoio dos profissionais de computação.
- **Apoio a eliminação segura dos dados** - ao longo do processo de pesquisa, cópias de arquivos de dados que não são mais necessárias precisam ser destruídas de forma segura, principalmente as que contêm dados sensíveis. Estratégias confiáveis para apagar definitivamente arquivos de dados de pesquisa constituem um componente crítico para a gestão segura dos dados, que deve estar presente em vários estágios do ciclo de vida dos dados.

Além dos serviços científicos e computacionais, outra categoria de serviços bastante relevante é a de serviços informacionais, conforme a seguir:

7.3.3 Serviços Informacionais

Grande parte dos serviços informacionais são oferecidos pelas bibliotecas e executados com o apoio dos profissionais bibliotecários e arquivistas. Considerando que as bibliotecas acadêmicas historicamente têm um papel preponderante em oferecer acesso aos registros de pesquisa, nas diversas formas em que eles se apresentam, não é surpresa que a gestão de dados seja uma questão assumida globalmente pelas bibliotecas e seus profissionais (TENOPIR; BIRCH; ALLARD, 2012, p.25). Cada vez mais as bibliotecas – principalmente as que estão vinculadas às instituições de pesquisa – incorporam ao seu elenco de serviços tradicionais serviços avançados e inovadores de curadoria dos dados.

No âmbito mais amplo da gestão de dados, as responsabilidades das bibliotecas estão além dos limites de ações meramente administrativos sobre a vastidão de novos produtos de pesquisa engendrados pela ciência contemporânea. Elas podem desempenhar um papel relevante e dinâmico no desenvolvimento de esquemas de metadados, ontologias e de ferramentas que apoiem a curadoria, e em métodos de rastreamento da proveniência, no estabelecimento de políticas para o depósito e acesso a dados (BORGMAN, 2016, p.13) e na reconciliação com os códigos éticos e legais vigentes. Num plano mais elevado, as bibliotecas de pesquisa podem criar estruturas de apoio à reprodutibilidade dos experimentos científicos, posto que esta noção é essencialmente baseada em registros científicos. O princípio da reprodutibilidade exige uma extensão profunda da catalogação e da indexação para incluir uma rede completa de objetos associados; requerem também uma estrutura de relacionamento de metadados elaborada que está além das práticas correntes como FRBR[7]. Além do mais, as práticas de licenciamento necessitam também se expandir para acomodar os direitos associados aos novos produtos de pesquisa. Dessa forma, os serviços informacionais compreendem um amplo espectro de atividades que vai desde o apoio à elaboração de plano de gestão de dados, até o arquivamento de longo prazo para os dados de valor contínuo, atravessando todo o ciclo de vida dos dados, constituindo um ponto agregador e referencial de informações sobre dados. A seguir apresentamos algumas instâncias de serviços informacionais que podem ser oferecidos:

7 FRBR – Functional Requirements for Bibliographic Records . Disponível em: https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr/frbr_2008.pdf Acesso em 04 out.2021

- **Portal de dados científicos** – portal *web* de dados de pesquisa de carácter institucional que tem como objetivo conectar os pesquisadores com as informações básicas sobre todos os aspectos da gestão de dados, tornando-se um ponto agregador dos serviços de gestão de dados, especialmente dos serviços de balcão de referência; o portal deve incluir também referências aos recursos externos à instituição.
- **Balcão de referência de dados** – é uma extensão do serviço de referência tradicional oferecido pelas bibliotecas de pesquisa, consistindo em um conjunto de serviços de consultoria que têm como objetivo orientar os pesquisadores nos vários aspectos da gestão de dados, tais como: identificação de repositórios para publicação; recuperação de *datasets* para meta-análises e outros reusos; aderência dos dados às normativas éticas, legais e de propriedade intelectual; elaboração de documento de consentimento esclarecido; atendimento aos requisitos sobre depósito de dados dos periódicos científicos; formatação de *datapapers*, entre outros.
- **Apoio na elaboração do plano de gestão de dados** – o serviço tem como objetivo assistir o pesquisador na elaboração, manutenção/revisão e consistência do plano de gestão de dados de pesquisas tendo como perspectiva o atendimento às exigências institucionais e das agências de fomento. Considera os padrões aplicáveis e ferramentas de *software* e *templates* disponíveis. O serviço pode também apoiar o desenvolvimento de padrão alinhado às diretrizes da política de dados da instituição.
- **Apoio na descoberta e acesso à coleção de dados** – Descoberta de dados é o processo de encontrar e acessar dados já criados e depositados em repositórios, arquivos ou centros de dados. Neste sentido, este serviço tem como objetivo apoiar os pesquisadores na identificação, localização e acesso às coleções de dados que possam ser reusados em suas pesquisas, posto que, em muitos casos, os pesquisadores não usam somente as coleções de dados que eles criaram ou coletaram, mas também aquelas geradas e curadas por outros pesquisadores e instituições. É preciso notar que o uso de conjuntos de dados de serviços externos pode complicar o compartilhamento ou a publicação de dados posteriormente, se houver termos de licença que proíbam tais atividades.
- **Desenvolvimento de coleções de dados** – apoiar o planejamento e a construção de coleções de dados em termos informacionais, tais como: estrutura, padrões pertinentes, catalogação, metadados, identificadores, controles de versões e aderência aos padrões apropriados e às exigências éticas e legais, como por exemplo, anonimização e *copyright*. A aquisição de coleção de dados pode estar incluída neste serviço.

- **Desenvolvimento de metadados** – a aplicação de metadados confere informação e contexto aos dados, posto que isso nem sempre está aparente a partir deles somente, entretanto, padrões de metadados disciplinares só existem em alguns domínios, que implica que eles precisam ser desenvolvidos (CHOUDHURY *et al.*, 2018, p. 7). Nessa direção, o serviço objetiva apoiar a estruturação e o desenvolvimento de esquemas de metadados, vocabulários, taxonomias e ontologias voltados para as especificidades disciplinares dos dados gerados pelas pesquisas desenvolvidas na instituição.
- **Criação de referências padronizadas** – assim como as publicações acadêmicas, como artigos e livros, as coleções de dados adequadamente referenciadas são mais facilmente acessadas, compartilhadas e reusadas e têm sua autoria reconhecida com mais precisão. O serviço de citação apoia o pesquisador na criação de referências padronizadas para as suas coleções de dados de seus versionamentos de forma a aumentar a visibilidade e a citação de seus dados. A citação padronizada torna as coleções de dados e suas versões únicas e mais fáceis de ser identificadas e descobertas.
- **Identificação de dados e pesquisadores** – serviço de apoio à aquisição, criação, aplicação e manutenção de identificadores persistentes para as coleções de dados e de apoio à aplicação de identificadores a pesquisadores.
- **Catalogação/indexação das coleções de dados** – para apoiar a descoberta dos dados, metadados genéricos podem ser assinalados às coleções de dados expondo-as, dessa forma, aos grandes sistemas de descoberta, agregadores e máquinas de busca (CHOUDHURY *et al.*, 2018) Nesse sentido, o serviço está voltado para a adição de metadados descritivos às coleções de dados (autor, título, identificadores persistentes etc.). Tem como referência uma política de indexação/catalogação previamente estabelecida e como objetivo aumentar os níveis de identificação e encontrabilidade dos dados. Complementa a descrição disciplinar que está focada nos métodos de coleta, processamento, análise e proveniência, e que são assinalados por pesquisadores ou especialistas de assunto. Desempenha um papel importante no registro das informações administrativas, como os termos de uso, que inclui quem pode usar os dados e como podem ser usados, incluindo também questões éticas e legais, como privacidade, tempo de embargo e problemas relacionados à propriedade intelectual.
- **Arquivamento de Longo Prazo/Preservação** – a maioria dos dados gerenciados individualmente por pesquisadores será perdida devido à fragilidade e à degradação física das mídias nas quais os dados estão armazenados e pelo ciclo veloz da obsolescência tecnológica; porém, a perda de conhecimento ao

longo do tempo se dá, principalmente, à medida que o pesquisador esquece detalhes sobre a coleção de dados e os processos de análise; e ainda por razões pessoais, como transferência, morte e aposentadoria (MAYERNIK *et al.,* 2012). Para mitigar esse problema, o serviço oferece infraestrutura técnica e informacional confiáveis voltadas para o arquivamento de médio e longo prazos de conjunto de dados selecionados e de suas informações de representação (metadados e documentação) que garantam o acesso aos seus conteúdos com níveis aceitáveis de autenticidade, confiabilidade e garantia de proveniência. O serviço tem como objetivo assegurar que as coleções de dados mantenham suas qualidades arquivísticas ao longo do tempo e do espaço, mantendo níveis de confiabilidade que viabilizam o reuso por outros pesquisadores, agora e no futuro. Para tal, é essencial empregar e se manter alinhado às normas, padrões e modelos conceituais, tais como OAIS, RDC-arq, PREMIS, entre outros. A expertise em curadoria assegura a resiliência e a interoperabilidade ao longo do tempo dos dados digitais, equacionando como os requisitos de significado, integridade, autenticidade e proveniência dos dados gerados hoje serão (JOHNSTON, 2017) capturados e transmitidos para o futuro.

- **Publicação de dados** – A pesquisa científica nos dias de hoje não produz somente as publicações convencionais, como artigos e livros, mas produz também coleções de dados em vários formatos – planilhas, base de dados, modelos, algoritmos etc. Subsequentes- ou em paralelo ao atual esquema de publicação acadêmico, as coleções de dados de pesquisa podem ser publicadas de forma independente e se tornam fontes importantes de informação e de análises para novas pesquisas (GRAFF; WAAIJERS, 2011). O serviço de publicação de dados apoia a preparação de dados para a publicação, bem como assessora na escolha do repositório mais adequado em termos do tipo de dados, disciplinas, licenças, volume e eventual custo. Processos adicionais complexos estão envolvidos na preparação informacional das coleções de dados para a submissão em um repositório, incluindo a geração de metadados alinhados com esquemas relevantes e ontologias, a criação de identificadores persistentes para a gestão, não somente para a citação, mas também versões, subconjuntos e outros produtos derivativos, além dos meios mais comuns de publicação de dados, via depósito em repositórios ou centros de dados, e a publicação de uma descrição dos dados na forma de artigo de dados (*datapapers*) em periódicos especializados chamados periódicos de dados (*datajournals*) .
- **Contextualizaçãp/Linking** – dependendo do campo e do projeto, uma publicação pode se basear em vários conjuntos de dados, ou um conjunto de

dados pode ser base para várias publicações. Nesses casos, os artigos se tornam descrição e documentação dos dados (BORGMAN, 2007, p. 117). Se dados e documentos relacionados pudessem ser linkados formando uma ecologia informacional, novas formas criativas de dados e informação possibilitariam pesquisas e aprendizagem distribuídas, colaborativas, multidisciplinares (BORGMAN, 2007). Além do mais, quando alguém recuperar ou ler um artigo, pode ir diretamente aos dados, e, em alguns casos, usar as ferramentas de análise disponibilizada pelo periódico; inversamente, quando um pesquisador acessa uma coleção de dados de seu interesse, ele pode ir diretamente aos artigos que resultaram dessas coleções. "O movimento fácil entre publicações e dados [e vice-versa],, nesse modelo depende de que ambos – dados e artigos – estejam acessíveis abertamente". Dessa forma, uma infraestrutura informacional que pode estabelecer e manter conexões entre recursos associados, como dados e publicações, aumenta a cadeia de valor da pesquisa. Documentos acadêmicos e dados, nos quais eles são baseados, têm mais valor combinados, do que sozinhos.

- **Treinamento para pesquisadores** – A geração e o uso intensivo de dados da pesquisa criam novos desafios para os pesquisadores e demandam expertise em gestão de dados que, geralmente, não fazem parte da formação dos cientistas (TENOPIR; BIRCH; ALLARD, 2012). Entretanto, o treinamento de usuários é essencial em todos os estágios da gestão de dados. Num ambiente em constante mudança, para que o pesquisador (público-alvo) se mantenha atualizado à medida que a infraestrutura de gestão evolui, é necessário também um estreito comprometimento dos serviços de desenvolvimento e manutenção com a capacitação dos pesquisadores (WILSON *et al.*, 2011). Com o objetivo de contornar esse problema, o serviço tem como objetivo dotar os pesquisadores de conhecimentos básicos sobre metadados, identificadores, plano de gestão de dados, publicação de dados. Este serviço é complementado pelo Apoio Computacional, que assiste os pesquisadores na utilização de ferramentas da tecnologia de informação, como *software* estatísticos, de visualização e *workflow*. Algumas instituições tomam como princípio o entendimento de que a capacitação em gestão de dados deve ser integrada às infraestruturas de aprendizagem acadêmica, como as disciplinas de metodologia científica.

Soma-se ainda aos serviços científicos, computacionais e informacionais a categoria serviços de administração, também de fundamental importância e que completa o modelo de serviços de gestão de dados de pesquisa, como observado a seguir:

7.3.4 Serviços de administração

Nesta categoria são incluídos os serviços que não se enquadram nas categoriais científicos, computacionais e informacionais, mas que são importantes para dar apoio, sustentabilidade e visibilidade a esses serviços. Compreende serviços de orientação sobre custos, orçamento, aquisição de coleções de dados, conformidades ética e legal dos dados – especialmente dados sensíveis – às normativas e regulamentos institucionais, nacionais e internacionais; estatísticas de uso e reuso dos dados; esta categoria envolve também as questões de propriedade intelectual, licenças e tempo de embargo.

- **Aquisição de coleção de dados –** além de se utilizar de fontes externas, algumas instituições e suas bibliotecas estão adquirindo coleções individuais de dados motivadas por demandas de seus pesquisadores. Este processo muitas vezes requer uma extensa negociação em relação aos custos e à amplitude do acesso e aos termos de uso dos dados. As bibliotecas – especialmente as envolvidas com aquisição de recursos digitais – estão bem qualificadas para dar apoio a essas atividades (CHOUDHURY *et al.,* 2018).
- **Estatísticas de (re)uso –** é de grande importância para a instituição e para os pesquisadores compreender o nível de acesso e o mapa de reuso dos dados; por exemplo, que pesquisadores estão acessando os dados, em que projetos estão sendo reusados, em que áreas do conhecimento estão sendo aplicados, que tipos de análises estão sendo realizadas. Estas informações são catalisadores importantes para a identificação de oportunidades de pesquisa colaborativa, para o aperfeiçoamento dos serviços e ainda para a prestação de contas aos financiadores da plataforma.
- **Custo/orçamento** - as atividades de gestão e compartilhamento de dados envolvem um aporte considerável de recursos financeiros em diferentes rubricas que precisam ser orçadas em um nível institucional e no âmbito dos projetos de pesquisa. Nesse contexto, as estimativas de custo, orçamentação, elaboração de cronogramas de desembolso, entre outros, se tornam uma atividade essencial, especialmente na formulação do projeto e do plano de gestão de dados.
- **Propriedade intelectual –** apoio ao pesquisador em relação às condições legais de copyright, licenças, tempo de embargo e afins, tanto para o reuso de dados de outros pesquisadores, quanto para a disponibilidade de seus próprios dados.
- **Conformidade ética e legal –** orientação sobre a conformidade ética e legal dos dados em relação à legislação nacional e internacional e aos códigos da instituição.

- **Divulgação/disseminação** - elaboração e execução de um plano de ações de divulgação/disseminação das coleções de dados que amplie as possibilidades de colaboração disciplinar e interdisciplinar e eleve o nível de visibilidade, traduzido por citações das coleções de dados.

A ideia de tornar os dados aderentes aos Princípios FAIR - expressa pelo termo "FAIRificação" - não se realiza por si só, para tal é necessário um processo multidimensional de gestão de dados, consubstanciado por um elenco de serviços, que efetivamente possa ir agregando valor ao longo do tempo aos objetos de pesquisa. Esta é a última camada da arquitetura proposta, que será apresentada a seguir.

7.4 FAIRificação dos dados

O que se constata é que o nível de conformidade dos vários objetos digitais de pesquisa aos Princípios FAIR está vinculado ao alcance e profundidade da gestão à qual eles estão submetidos. Conforme visto, isto pressupõe a necessidade de um arcabouço de várias camadas – científica, tecnológica, informacional e adminsitrativo - que endereçam os múltiplos problemas que se interpõem entre o reuso, integridade, proveniência, reprodutibilidade, prestação de contas, bem como entre as novas necessidades e oportunidades de análise e reanalise, em larga escala, necessárias aos fluxos da *eScience* (WILKINSON *et al*., 2016). Por conseguinte, o grau de aderência aos Princípios FAIR – a FAIRificação - põe em evidência um conjunto de serviços, procedimentos e ferramentas que, mesmo não tendo esse parâmetro como objetivo final, coloca-o como uma parte de uma escala renovada de avaliação da maturidade de sistemas de gestão de dados de pesquisa.

Com o propósito de deixar mais claros os significados sintetizados no acrônimo FAIR, Mons e colaboradores (2017) elaboraram uma escala simples de FAIRificação, redesenhada neste estudo para incluir os serviços associados a este processo. Nesta representação, os elementos se tornam coloridos na medida em que se tornam FAIR: elementos coloridos em verde, são FAIR e abertos; elementos em vermelho, são FAIR e fechados. No percurso delineado pelos autores, no nível mais baixo da gradação estão os objetos de pesquisa indisponíveis para o reuso, que segundo eles, "representam 80% dos *datasets* nas práticas atuais" (MONS *et al.,* 2017, p. 52). Nessa categoria estão os dados não publicados ou publicados em ambientes instáveis e linkados precariamente, como material suplementar a artigos de periódicos; também **não dispõem de identificador único persistente resolvível por máquina que permite a renderização dos dados e de seus metadados, não legíveis por máquina**. O segundo passo, considerado o mais básico rumo à FAIRificação, consiste em atribuir ao *dataset* – na qualidade de uma entidade propriamente dita - um **identificador**

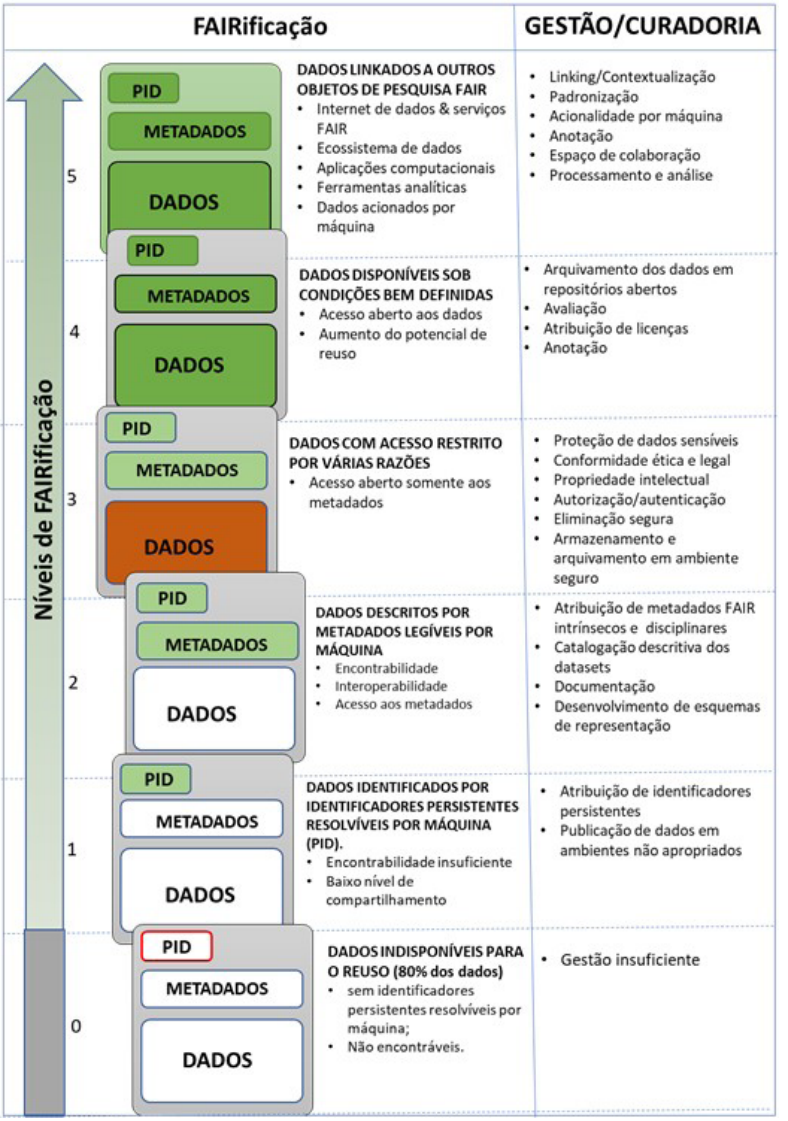
persistente (PID, na sigla em inglês). Contudo, sem um conjunto de metadados legíveis por máquina será difícil encontrar o recurso, a menos que se conheça, a priori, o seu PID. Isso nos indica que o identificador é necessário, porém insuficiente para atender o princípio da **encontrabilidade**, e que é preciso ir mais adiante.

Fica claro, portanto, que o passo seguinte é a **atribuição de metadados** que podem ser categorizados quanto à origem em: metadados intrínsecos, que são assinalados no momento da captura dos dados, geralmente por processos automatizados, por instrumentos ou pelo *workflow* que gera os dados; metadados assinalados pelos pesquisadores que criaram/coletaram os dados, profissionais de informação que catalogam os dados e outros pesquisadores que adicionam, por exemplo, informações contextuais na forma de anotações, que conferem proveniência e contextualização aos dados e incrementam seu grau de FAIRificação. Assim sendo, a adição de metadados ricos – e que também sejam FAIR – é um passo essencial para encontrar, acessar, interoperar e, como consequência, reusar os objetos de pesquisa. De fato, a identificação persistente e a agregação de metadados já atribuem um profundo efeito no potencial de reuso dos objetos de pesquisa, posto que podem ser identificados e recuperados, resumem Mons e colaboradores (2017). Contudo, mesmo que o dado seja tecnicamente FAIR, ele pode estar com o acesso restrito por razões claras e justas, tais como, contratos, proteção de espécies em extinção, questões legais e éticas; isto dito, compreendemos que o padrão máximo de FAIRificação deve acontecer quando os próprios elementos de dados estão disponíveis, sob condições bem definidas, para o reuso aberto por parte de qualquer outro interessado.

Indo ainda mais além na escala de FAIRificação, Barend Mons e seus colaboradores (2017) propõem que, quando os dados estiverem linkados a outros objetos de pesquisa FAIR, teremos alcançado a "Internet de Dados FAIR"; uma vez que um número crescente de aplicações e serviços podem linkar e processar dados FAIR, pode-se dizer que terá sido alcançada a "Internet de Dados & Serviços FAIR", significando um "ambiente global e compartilhado voltado para pesquisas orientadas por dados e inovação" (SALES *et al.*, 2020, p.3), onde todos os pesquisadores podem acessar, armazenar, analisar e reusar dados para a pesquisa, inovação e para propósitos educacionais. A partir dos contornos desse território, se estabelece uma ecologia de dados ativados por serviços associados que, para os diversos segmentos de usuários, se traduz num contínuo de benefícios acionados por aplicações computacionais.

A escala proposta, baseada no modelo original de Mons e colaboradores (2017), é sintetizada na figura 2, abaixo:

Figura 2 – Níveis de FAIRificação



Fonte: os próprios autores baseados em MONS *et al*. (2017).

Idealmente, os serviços que estão subjacentes aos processos de FAIRificação se ancoram conceitualmente em alguns pressupostos que são essenciais para a realização do objetivo finalístico de reuso, mas também para a concretização do ideal de uma Internet de Dados & Serviços FAIR, último e mais avançado passo na escala

acima proposta. Traduzindo essa ancoragem conceitual em três pontos focais, que realinham as ações dos novos sistemas de gestão de dados, temos:

- *FAIR é sobre acionalidade por máquina* - "O reconhecimento de que os computadores devem ser capazes de acessar dados publicados de forma autônoma, sem ajuda de operadores humanos, é central para os Princípios FAIR", afirmam de forma categórica Mons e colaboradores (2017, p. 51); assim sendo, "os princípios FAIR colocam uma ênfase privilegiada no aprimoramento das potencialidades das máquinas em encontrar e usar os dados, além de apoiar seu reuso por seres humanos", ratificam Wilkinson e colaboradores (2016, p. 1). Os "*stakeholders* computacionais", tais como, programas de aplicação e agentes computacionais, são exploradores que agem em nosso nome – seres humanos -, performando um papel crescentemente relevante na recuperação e análise de dados. Nesse contexto em constante transição, é preciso, portanto, considerar que os seres humanos não são os únicos interlocutores críticos no ecossistema de dados. Os Princípios FAIR são também, e principalmente, para as máquinas.

Essas configurações e condições impostas pela ciência contemporânea têm um reflexo profundo nos processos das modernas plataformas de gestão de dados; assim sendo, a adoção total ou parcial dos princípios FAIR, como parte da espinha dorsal desses sistemas técnicos-gerenciais, é um passo importante na direção da acionabilidade por máquina, na medida em que os habilitam a otimizar o uso dos recursos de dados por meio de escolhas de implementação técnicas adequadas. Por exemplo, o recurso digital pode ser usado como um agente ou um substrato em análises baseado em aprendizagem por máquina ou inteligência artificial. O segundo ponto é a conexão da acionabilidade por máquina aos metadados.

- FAIR *é sobre metadados* – Há uma ponte imprescindível entre acionalidade por máquina e metadados, posto que um objeto de pesquisa deve se encontrar num contínuo de possíveis estados que o possibilita fornecer informações cada vez mais detalhadas – na forma de metadados - a um explorador computacional. Dessa forma, assistir as máquinas na descoberta e exploração de dados, por meio de aplicações de tecnologias e padrões no nível das plataformas de dados, se torna a prioridade máxima de uma boa gestão de dados e coloca em relevo a essencialidade do conceito de metadados. Nessa perspectiva, os Princípios FAIR enfatizam a importância dos metadados e de seus padrões na gestão de dados, focalizando o conceito de "metadado"

transversalmente nos seus 15 princípios orientadores. "A mensagem-chave dos Princípios FAIR é que metadados e padrões de metadados devem ser articulados e tornados publicamente disponíveis na maior amplitude possível" (BOECKHOUT; ZIELHUIS; BREDENOORD, 2018, p. 932).

O terceiro ponto coloca em evidência a eventual, porém desejável, no vasto e diversificado mundo da pesquisa, desconexão entre os Princípios FAIR e dados abertos,

- *FAIR É SOBRE ACESSO SOB CONDIÇÕES BEM DEFINIDAS* - "FAIR não é igual a aberto", afirmam assertivamente JACOBSEN *et al.*, (2020). O "A" no contexto do FAIR é compreendido como "Acessível sob condições bem definidas", o que o torna diferente de aberto sem restrições. Mons e colaboradores (2017, p.51) destacam que podem existir razões legitimas para blindar dados e serviços gerados com fundos públicos do acesso indiscriminado. Esses tipos de dados incluem: dados pessoais sensíveis, dados sobre geolocalização de espécie em perigo de extinção, sobre processos patenteáveis, segurança nacional, entre muitos outros. Além do mais, diversos setores, como o industrial e o médico, por razões legais, éticas, contratuais ou de competitividade, precisam de segurança apropriada para seus dados e requerem medidas adicionais de autorização e autenticação, tanto para exploradores humanos, como para agentes computacionais; na prática, a Internet de Dados & Serviços FAIR não pode funcionar sem esses mecanismos (JACOBSEN *et al*, 2020).

Considerando esses pontos como subjacentes aos processos mais elevados de FAIRificação, Mons e colaboradores (2017) destacam que os mecanismo para expressar os relacionamentos entre os diversos elementos identificadores, metadados e dados também são FAIR – isto é, seguem padrões legíveis por máquina e amplamente aceitos e são interlinkados com outros dados FAIR relacionados ou com ferramentas analíticas no ambiente da Internet de Dados e Serviços FAIR, permitindo um reuso profundo, diversificado e interpretativo por seres humanos e computadores.

8 À guisa de conclusão

A geração intensiva de dados que caracteriza a ciência contemporânea exige uma gestão peculiar para os ativos informacionais que ela gera e consome, cuja escala extrapola as medidas mais tradicionais praticadas, por exemplo, para a ges-

tão de livros e artigos acadêmicos, estejam eles na forma impressa ou digital. O que confere valor aos dados e a outros objetos digitais produzidos pela e-pesquisa é a qualidade deles se apresentarem em um estado ótimo que os coloquem continuamente prontos para o acesso e reuso – objetivos finalísticos da gestão - por seres humanos e, sobretudo, por provedores de serviços, por meio de aplicações computacionais. Estas novas condições ampliam o potencial de repropósito e de ressignificação transversal dos dados, além das fronteiras científicas, ampliando o seu potencial de aplicação também no âmbito das comunidades de prática, como nas áreas de engenharia e agricultura.

Nessa nova era científica, em que a escassez de dados e informações é menos crítica que o excesso, as dificuldades dos agentes humanos operarem na frequência e velocidade exigidas pela complexidade das ciências intensivas de dados, reforçam a necessidade de exploradores computacionais agirem de forma autônoma e inteligente, tendo como perspectiva a articulação de um ecossistema global de dados e serviços subjacentes aos dispositivos intelectuais, sociais e ciberestruturais de produção de conhecimento científico. Esse contexto exige que as plataformas de gestão de dados se ajustem às infraestruturas computacionais, aos processos de análises e workflows sofisticados e incorporem expertises que sejam capazes de lidar com os ambientes e processos tecnologicamente sofisticados da pesquisa atual.

Com esta perspectiva, o modelo proposto procurou desconstruir os blocos que compõem uma arquitetura genérica de uma plataforma de serviços de gestão de dados, articulando os vários módulos conceituais – diretrizes, políticas, serviços, ferramentas, infraestruturas etc. – na forma de peças que podem ser ajustadas de acordo com a profundidade, alcance e filosofia de cada instituição ou disciplina, proporcionando, dessa forma, uma possível escala para apoiar a mensuração do nível de maturidade dos projetos de serviços de gestão, posto que nem toda instituição aspira ter o mesmo patamar de serviço ou de maturidade. Para ordenar esses esforços, por vezes entrópicos e difusos, a aplicação do FAIR, dimensionada pelos seus 15 princípios orientadores, aponta um horizonte para a construção eficiente de arquiteturas de plataformas de gestão de dados.

Mesmo tendo em conta o enfoque generalista do modelo, é preciso considerar que na implantação de práticas e infraestruturas de gestão de dados, o contexto específico das comunidades científicas e as possibilidades da adoção devem ser observadas. A importância de cada serviço pode depender das prioridades e da geração e uso de determinados objetos de pesquisa. Essa condição implica que diferentes disciplinas encontrem soluções técnicas e necessitem de arcabouços infraestruturais e organizacionais em torno de serviços de gestão diferentes para alcançar o grau de FAIRificação requerido por suas comunidades.

Por fim, é preciso compreender, no modelo proposto, o papel integrador da biblioteca no mundo complexo da *eScience*: as bibliotecas de pesquisa são responsáveis pela conexão entre os serviços informacionais, computacionais, científicos e administrativos. Nesse domínio de grandes novidades e de novas interlocuções, para oferecer apoio efetivo à gestão de dados de pesquisa para seus diversos públicos-alvo, as bibliotecas devem ser proativas e desenvolver serviços que olhem para o futuro, mas que, todavia, valorizem os recursos humanos, tecnológicos e intelectuais consolidados por uma longa trajetória, valorizando o conhecimento de outros profissionais e trazendo-os para dentro da biblioteca, visando a implementação de serviços inovadores e harmônicos para com os interesses de suas comunidades. De fato, cada vez mais as bibliotecas de pesquisa incorporam ao seu elenco de tarefas tradicionais a oferta de serviços avançados de gestão de dados. Diante desta reconfiguração contínua, é reconfortante compreender que os avanços nos registros de pesquisas estão sendo gerenciados por uma instituição ancorada na tradição secular de acesso e preservação da memória acadêmica.

Referências

BALL, Alex. **Review of data management lifecycle models**. Bath, UK: University of Bath, 2012. Disponível em: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.224.4219&rep=rep1&type=pdf. Acesso em: 25 abr. 2021.

BOECKHOUT, Martin; ZIELHUIS, Gerhard A.; BREDENOORD, Annelien L. The FAIR guiding principles for data stewardship: fair enough? **European Journal of Human Genetics**, v. 26, n. 7, p. 931-936, 2018. Disponível em: https://www.nature.com/articles/s41431-018-0160-0.pdf. Acesso em: 25 abr. 2021.

BORGMAN, Christine L. **Big Data, Little Data, No Data**; Scholarship in the Networked World. London: The MIT Press, 2016.

BORGMAN, Christine L**.** Data: the input and output of scholarship. *In:* BORGMAN, Christine L. **Scholarship in the Digital Age**: Information, Infrastructure, and the Internet. London: The MIT Press, 2007. p. 115-147.

CHOUDHURY, Sayeed *et al.* **Research Data Curation**: A Framework for an Institution-wide Services Approach. 2018. Disponível em: https://hsrc.himmelfarb.gwu.edu/libfacpubs/35. Acesso em: 30 jul. 2021.

COATES, Heather L. Building Data Services from the Ground Up: Strategies and Resources. **Journal of eScience Librarianship,** v. 3, n.1, 2014. Disponível em: https://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1063&context=jeslib. Acesso em: 25 abr. 2021.

COX, Andrew; PINFIELD Stephen. Research data management and libraries: Current activities and future priorities. **Journal of Librarianship and**

Information Science, v. 46, n. 4, p. 299-316, 2014. Disponível em: http://lis.sagepub.com/cgi/doi/10.1177/0961000613492542. Acesso em: 01 set. 2021.
ERWAY, Ricky *et al*. **Building Blocks:** Laying the Foundation for a Research Data Management Program. Dublin: OCLC, 2016. Disponível em: https://files.eric.ed.gov/fulltext/ED589141.pdf. Acesso em: 25 abr. 2021.
FEARON JR., David *et al*. **SPEC Kit 334:** Research data management services. Washington, DC: Association of Research Libraries, 2013.
GOODMAN, Alyssa *et al*. Ten simple rules for the care and feeding of scientific data. **PLoS Computational Biology**, v. 10, n. 4, 2014. Disponível em: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003542. Acesso em: 01 set. 2021.
GRAAF, Maurits van der; WAAIJERS, Leo. **A surfboard for riding the wave**: Towards a four country action programme on research data. Copenhagen: Knowledge Exchange, 2011. Disponível em: https://www.voced.edu.au/content/ngv%3A48428. Acesso em: 25 abr. 2021.
HARVEY, Ross. **Digital Curation**: A How-to-do-it Manual. New York: Neal-Schuman Publisher, Inc., 2010.
HONG, Neil Chue *et al*. **Six recommendation to implementation of FAIR Practices**. Bruxelas: European Commission, 2020. Disponível em: https://ec.europa.eu/info/publications/six-recommendations-implementation-fair-practice_en. Acesso em: 25 abr. 2021.
JACOBSEN, Annika *et al*. FAIR principles: Interpretations and implementation considerations. **Data Intelligence**, n. 2, p. 10–29, 2020. Disponível em: http://www.inf.ufes.br/~gguizzardi/102-Annika_Jacobsen-1_GRFHSzW.pdf. Acesso em: 25 abr. 2021.
JOHNSTON, Lisa. **"Introduction to Data Curation" from Curating Research Data Volume One**: Practical Strategies for Your Digital Repository. Chicago: Association of College and Research Libraries, 2017. Disponível em: https://conservancy.umn.edu/handle/11299/185334. Acesso em: 25 abr. 2021.
JONES, Sarah; PRIOR, Grahan; WHITE, Angus. **How to develop research data management services – a guide for HEIs**. Edinburgh: Digital Curation Centre, 2013. Disponível em: https://www.dcc.ac.uk/guidance/how-guides/how-develop-rdm-services. Acesso em: 30 jul. 2021.
KOUPER, Inna *et al.* Research Data Services Maturity in Academic Libraries. *In*: JOHNSTON, Lisa R. (ed.). Curating Research Data*: Practical Strategies for Your Digital Repository. Chicago: Association of College and Research Libraries,* 2017. p. 153-170. Disponível em: https://experts.illinois.edu/en/publications/research-data-services-maturity-in-academic-libraries. Acesso em: 25 abr. 2021.

LEONARDI, Paul M. Digital materiality? How artifacts without matter, matter. **First Monday**, v. 15, n. 6-7, 2010. Disponível em: https://journals.uic.edu/ojs/index.php/fm/article/view/3036. Acesso em: 25 abr. 2021.
LORD, Philip; MACDONALD, Alison. **E-Science curation report**. **Data curation for e-science in the UK**: An audit to establish requirements for future curation and provision. Twickenham: JCSR, 2003. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.96.5156&rep=rep1&type=pdf. Acesso em: 25 abr. 2021.
MARTINEZ-URIBE, Luis. **Research data management services**: findings on the consultation with service providers. Oxford: Oxford Digital Repositories Steering Group, 2019.
MAYERNIK, Mathews S. *et al*. The data conservancy instance: infrastructure and organizational services for research data curation. **D-Lib Magazine**, v. 18, n. 9-10, Sept./Oct. 2012. Disponível em: http://www.dlib.org/dlib/september12/mayernik/09mayernik.html. Acesso em: 25 abr. 2021.
MONS, Barend *et al.* Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. **Information Services & Use**, v. 37, n. 1, p. 49-56, 2017.
MUSHI, Gilbert Exaudi; PIENAAR, Heila; DEVENTER, Martie van. 2020. Identifying and Implementing Relevant Research Data Management Services for the Library at the University of Dodoma, Tanzania. **Data Science Journal**, v.19, n. 1, p. 1–9, 2020. Disponível em: https://datascience.codata.org/articles/10.5334/dsj-2020-001/. Acesso em: 25 abr. 2021.
NATIONAL RESEARCH COUNCIL. **Preparing the workforce for digital curation**. Washington, D.C.: The National Academies Press, 2015.
NIELSEN, Hans Jorn; HJORLAND, Bierger. Curation research data: the potential roles of libraries and information professionals. **Journal of Documentation**, v. 70, n. 2, 2014. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/JD-03-2013-0034/full/html. Acesso em: 30 jul. 2021.
PINFIELD, Stephen; COX, Andrew M.; SMITH, Jen. Research data management and libraries: Relationships, activities, drivers and influences. **PLoS One**, v. 9, n. 12, p. e114734, 2014. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0114734. Acesso em: 25 abr. 2021.
REED, Robyn B. Diving into data: Planning a research data management event. **Journal of Escience Librarianship**, v. 4, n. 1, 2015. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4517608/. Acesso em: 25 abr. 2021.
SALES, Luana Farias *et al*. GO FAIR Brazil: a challenge for brazilian data science. **Data Intelligence**, v. 2, n. 1-2, p. 238-245, 2020. Disponível em: https://

direct.mit.edu/dint/article/2/1-2/238/10004/GO-FAIR-Brazil-A-Challenge-for-Brazilian-Data. Acesso em: 25 abr. 2021.
SALES, Luana Farias; SAYÃO, Luís Fernando. A ciência invisível: revelando os dados da cauda longa da pesquisa. *In:* ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 19. (XIX ENANCIB). **Anais**... Marilia: UNESP, 2018.
SAYÃO, Luís Fernando. Modelos teóricos em ciência da informação - abstração e método científico. **Ciência da informação**, v. 30, n. 1, p. 82-91, 2001.
SAYÃO, Luís Fernando; SALES, Luana Farias. Afinal, o que é dado de pesquisa? **BIBLOS**, v. 34, n. 2, 2020. Disponível em: https://www.seer.furg.br/biblos/article/view/11875. Acesso em: 12 maio 2021.
SAYÂO, Luís Fernando; SALES, Luana Farias. **Guia de gestão de dados de pesquisa para pesquisadores e bibliotecários**. Rio de janeiro: CNEN, 2015.
SOLOMONIDES, Anthony. Research Data Governance, Roles, and Infrastructure: Methods and Applications. *In:* RICHESSON, Rachel L.; ANDREWS, James E. (ed.). **Clinical Research Informatics**. Cham: Springer, 2019. p. 291-310.
STRASSER, Carly. **Research data management**. Baltimore: NISO, 2015. Disponível em: https://wiki.lib.sun.ac.za/images/2/24/PrimerRDM-2015-0727.pdf. Acesso em: 25 abr. 2021.
TANG, Rong; HU, Zhan. Providing Research Data Management (RDM) Services in Libraries: Preparedness, Roles, Challenges, and Training for RDM Practice. **Data and Information Management**, v. 3, n. 2, p. 84-102, 2019.
TENOPIR, Carol; BIRCH, Ben; ALLARD, Suzie. **Academic libraries and research data services**: Current practices and plans for the future. Chicago, IL: Association of College and Research Libraries, 2012. Disponível em: https://trace.tennessee.edu/utk_dataone/20/. Acesso em: 25 abr. 2021.
TORINO, E.; ROA-MARTÍNEZ, S. M.; VIDOTTI, S. A. B. G. Dados de pesquisa: disponibilização ou publicação?. In: SHINTAKU, M.; SALES, L. F; COSTA, M. (org). **Tópicos sobre dados abertos para editores científicos**. Botucatu, SP: ABEC, 2020. p. 183-201. DOI: 10.21452/ 978-85-93910-04-3.cap15.
WILKINSON, Mark D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data**, v. 3, n. 1, p. 1-9, 2016. Disponível em: https://www.nature.com/articles/sdata201618.pdf. Acesso em: 25 abr. 2021.
WILSON, James A. J. *et al.* An institutional approach to developing research data management infrastructure. **The International Journal of Digital Curation**, v. 6, n. 2, 2011. Disponível em: http://ijdc.net/index.php/ijdc/article/view/198. Acesso em: 25 abr. 2021.

Agradecimentos

Ao CNPq, por financiar esta pesquisa por meio de nossas bolsas de produtividade.

Aos alunos da turma 1/2021 da disciplina "Gestão de Informação para produção do conhecimento", dos Cursos de Mestrado e Doutorado em Ciência da Informação, do PPGCI IBICT-UFRJ, que através de seus trabalhos e discussões nos motivaram e nos alimentaram com informações importantes para o desenvolvimento do presente trabalho.

Agradecimentos

Agradeço à Carla Maria Martellote Viola, por ter aceitado o desafio de organizar esta obra em homenagem ao Prof. Luís Fernando Sayão. Para mim, sempre foi difícil separar nossa história de pesquisa, de amor e de amizade. Carla me ajudou até aqui a tornar esta obra um pouco mais neutra, mas não seria natural, se esses agradecimentos não viessem cheios de amor, carinho e admiração.

Deixo, então, aqui registrado o meu agradecimento ao meu companheiro de pesquisa e de vida, Prof. Luís Fernando Sayão, pelo conhecimento, pelo incentivo, pelas novas ideias e pelo amor que divide comigo todos os dias.

Agradeço ao nosso Programa de Pós-Graduação, que ao lançar o Edital Comemorativo dos 50 anos do Programa, criou a categoria "**PHr - Pesquisas históricas em revista"**, cujo objetivo era a organização de "coletâneas autorais" de artigos de docentes com 30 ou mais anos de atuação no Programa, me permitindo ousar organizar esta obra, que há muito tempo estava em minha mente e em meu coração, homenageando aquele que para mim, durante muito tempo, foi apenas mestre e orientador.

Agradeço aos editores das revistas Ciência da Informação, Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Informação & Sociedade: Estudos, PontoDeAcesso, Revista Eletrônica de Comunicação, Informação e Inovação em Saúde, Revista USP e Transinformação, por possibilitarem o compartilhamento dessas informações.

Agradeço ao querido Prof. Carlos Henrique Marcondes, meu professor por diversas vezes, primeiro parceiro intelectual do Sayão, e que também dividiu com ele o conhecimento, por muitos anos.

Agradeço ao nosso Comitê Científico, que imediatamente aceitou o nosso convite e abraçou a ideia, nos ajudando, inclusive, com a organização dos capítulos. Agradecimento em especial à Professora Lena Vânia Ribeiro Pinheiro, que tão carinhosamente apresentou o autor.

Agradeço também ao nosso Grupo de Pesquisa BRIET, em especial à Marcelle Costal, Castro, à Melina Brito dos Santos, à Tayná Regly e à Dilza Fonseca da Motta pelo trabalho tão minucioso no Comitê Editorial.

Agradeço ainda àquela que tão gentilmente, há anos, revisa todos os nossos textos, compartilhando conosco a alegria de cada novo conhecimento adquirido: nossa querida Teodora Marly Gama das Neves. Por fim, agradeço ao CNPq, que vem financiando nossas pesquisas através de bolsa de produtividade Pq1.

Sobre o autor

Luís Fernando Sayão

Doutor em Ciência da Informação pela UFRJ/IBICT (1994). Mestre em Ciência da Informação pela (UFRJ/IBICT). Graduação em Física pela Universidade Federal do Rio de Janeiro (1978).

Instituição

Comissão Nacional de Energia Nuclear
Centro de Informação Nuclear
Programa de Pós-Graduação em Biblioteconomia da UNIRIO - PPGB
Programa de Pós-Graduação em Memória e Acervos da Fundação Casa de Rui Barbosa

Dados biográficos

Tecnologista sênior desde 1980 na Comissão Nacional de Energia Nuclear. É conselheiro do CONARQ - Conselho Nacional de Arquivos, docente permanente do Programa de Pós-graduação em Ciência da Informação do convênio IBICT-UFR. Docente Colaborador no Programa de Pós-Graduação em Biblioteconomia da UNIRIO - Universidade Federal do Estado do Rio de Janeiro e no Programa de Pós-Graduação em Memória e Acervos da Fundação Casa de Rui Barbosa. Bolsista de Produtividade do CNPq Pq2. Vice-líder do Grupo de Pesquisa BRIET – Biblioteconomia, Recuperação, Interoperabilidade, E-science e Tecnologias.

E-mail: luis.sayao@cnen.gov.br
CV: http://lattes.cnpq.br/3422623122948 38
ORCID: http://orcid.org/0000-0002-6970-0553

Sobre as organizadoras

Luana Farias Sales

Doutora em Ciência da Informação pelo Programa de Pós-Graduação do IBICT/UFRJ (2011-2014). Mestre em Ciência da Informação pelo convênio UFF/IBICT (2004-2006), Graduação em Biblioteconomia e Documentação pela Universidade Federal Fluminense (2003).

Instituição

Instituto Brasileiro de Informação em Ciência e Tecnologia - IBICT

Programa de Pós Graduação em Ciência da Informação - PPGCI

Programa de Pós-Graduação em Biblioteconomia da UNIRIO – PPGB

Dados biográficos

Analista em C & T do MCTIC/IBICT, atuando como docente do Programa de Pós-graduação em Ciência da Informação do convênio IBICT-UFRJ e Coordenadora da Rede de Implementação do GO-FAIR Brasil. Docente colaboradora no Programa de Pós-Graduação em Biblioteconomia (UNIRIO). Bolsista de Produtividade do CNPq Pq2. Líder do Grupo de Pesquisa BRIET: – Biblioteconomia, Recuperação, Interoperabilidade, E-science e Tecnologias.

E-mail: luanasales@ibict.br

CV: http://cnpq.br/9090064478702633

ORCID: http://orcid.org/0000-0002-3614-2356

Carla Maria Martellotte Viola

Doutoranda (PPGCI-IBICT/UFRJ/2019) e Mestra em Ciência da Informação (PPGCI-IBICT/UFRJ/2018), graduada em Comunicação Social/Propaganda e Publicidade (FACHA/1985) e em Direito (Universidade Santa Úrsula/1997). Pós-graduada em Gênero e Direito (EMERJ/2018-2019), em Gestão Estratégica da Comunicação (IGEC/FACHA/2011) e Direito do Consumidor Responsabilidade Civil (AVM/Candido Mendes/2013) com complementação em Didática do Ensino Superior.

Instituição

Instituto Brasileiro de Informação em Ciência e Tecnologia - IBICT
Programa de Pós Graduação em Ciência da Informação – PPGCI – IBICT/UFRJ
Programa de Pós-Graduação em Biblioteconomia da UNIRIO – PPGB

Dados biográficos

Advogada, e Publicitária. Integrante do grupo de pesquisa Perspectivas Filosóficas em Informação - Perfil-i (IBICT/UFRJ), pesquisadora-colaboradora do projeto de pesquisa FARMi, especialmente no eixo InfoGend que articula investigações sobre igualdade de gênero, direitos das mulheres e acesso à informação do IBICT/UFRJ, integrante do grupo de pesquisa BRIET: Biblioteconomia, Representação, Interoperabilidade, E-science e Tecnologia (IBICT/UFRJ), conselheira titular do Conselho de Usuários da Região Sudeste da OI TELEMAR (2020-2024) e delegada da Comissão de Direito Digital da 16ª subseção da OAB/RJ.
E-mail: viola.carla@gmail.com
CV: http://lattes.cnpq.br/3133945606177771
ORCID: http://orcid.org/0000-0001-5053-9491

Realização

Cooperação

Financiamento

Esta obra é parte da Coleção PPGCI 50 Anos e foi composta em Minion pelo Programa de Educação tutorial da Escola de Comunicação da UFRJ em setembro de 2021.

Esta obra reúne os artigos mais citados, de acordo com Google Acadêmico, do professor Dr. Luís Fernando Sayão, escritos durante sua trajetória pela Ciência da Informação e por suas diversas passagens pelo PPGCI-IBICT/UFRJ, como aluno, consultor, pesquisador e professor.

EM COOPERAÇÃO

UNESCO
United Nations
Educational, Scientific and
Cultural Organization