Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação
ISSN: 1518-2924
Universidade Federal de Santa Catarina

Rodrigues, Marcello Mundim; Lourenço, Cíntia de Azevedo; Dias, Guilherme Ataíde

A NATUREZA DE CONJUNTOS DE DADOS CIENTÍFICOS EM REPOSITÓRIOS SUL-AMERICANOS: UM LEVANTAMENTO DE FORMATOS E EXTENSÕES

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, vol. 27, 2022, pp. 1-26
Universidade Federal de Santa Catarina

DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Disponível em: https://www.redalyc.org/articulo.oa?id=14775278011

Encontros Bibli

A NATUREZA DE CONJUNTOS DE DADOS CIENTÍFICOS EM REPOSITÓRIOS SUL-AMERICANOS: UM LEVANTAMENTO DE FORMATOS E EXTENSÕES

The nature of scientific datasets in South American repositories: a survey of formats and extensions

Marcello Mundim Rodrigues
Doutorando em Gestão e Organização do Conhecimento pela Universidade Federal de Minas Gerais
marcellomundim@ufu.br
https://orcid.org/0000-0001-7945-6673

Cíntia de Azevedo Lourenço
Doutora em Ciência da Informação pela Universidade Federal de Minas Gerais
Professora associada
Universidade Federal de Minas Gerais, Escola de Ciência da Informação, Belo Horizonte, Brasil
cintia.eci.ufmg@gmail.com
https://orcid.org/0000-0002-2172-7300

Guilherme Ataíde Dias
Pós-doutorado em Ciência da Informação pela UNESP
Professor associado III
Universidade Federal da Paraíba, Departamento de Ciência da Informação, João Pessoa, Pb, Brasil
guilhermeataide@gmail.com
https://orcid.org/0000-0001-6576-0017

A lista completa com informações dos autores está no final do artigo

RESUMO

Objetivo: identificar os repositórios de dados científicos criados e geridos por Instituições de Ensino Superior e/ou agências de pesquisa e fomento sul-americanas; identificar e descrever os formatos e extensões dos arquivos que compõem os conjuntos de dados científicos depositados nesses repositórios.
Método: oito repositórios recuperados pelo RE3DATA foram selecionados à investigação. Obteve-se uma população (N) de 1.115 conjuntos de dados científicos. A partir da Amostragem Aleatória Estratificada, chegou-se ao valor da amostra (n) igual a 258 conjuntos de dados, que corresponde a 23,15% da população (N). Os dados retirados das amostras foram condensados em tabelas e quadros.
Resultado: notou-se que a natureza dos conjuntos de dados científicos investigados se concentra em dados textuais e numéricos, salvos em arquivos de texto e em tabelas, respectivamente. Percebeu-se que os conjuntos de dados podem ser tanto homogêneos (um ou mais arquivos salvos em um único formato e extensão, ex.: formato de imagem em .jpg) ou heterogêneos (arquivos salvos em diferentes formatos e extensões, ex.: mesmo formato de imagem salvo em .jpg e .tiff) em sua composição. Apurou-se também que algumas extensões possibilitam a identificação da natureza, do domínio e do conteúdo dos dados, como observado nas extensões .gpx e .gdb, que se referem a dados de geolocalização, logo, de natureza alfanumérica.
Conclusões: há crescente necessidade de se descrever a natureza dos dados, assim como os formatos e extensões de seus arquivos. Esse tipo de metadado descritivo seria valioso a potenciais usuários, pois permitiria obter maior compreensão do contexto dos dados com foco em seu reúso.
Palavras-chave: dados científicos; conjuntos de dados; repositórios de dados; formatos e extensões; levantamento.

ABSTRACT

Objective: identifying the scientific data repositories created and managed by Higher Education Institutions and/or South American research and funding agencies; identifying and describing the formats and extensions of files that compile the scientific datasets deposited in these repositories.
Methods: eight repositories retrieved by RE3DATA were selected for investigation. A population (N) of 1.115 scientific datasets was obtained. By using Stratified Random Sampling, the resulting sample (n) value was 258 datasets, which corresponds to 23,15% of the population (N). Data surveyed from the samples were condensed into tables and charts.
Results: it was noticed that the nature of the scientific datasets investigated is centered on textual and numerical data, saved in text files and tables, respectively. Also, the datasets may be either homogeneous (one or more files saved in a unique format and extension, e.g.: image format in .jpg) or heterogeneous (files saved in different formats and extensions,

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

e.g.: same image format saved in .jpg and .tiff) in their composition. It was found that some extensions enable the identification of the nature, domain and content of the data, as observed in the .gpx and gdb extensions, which refer to geospatial data, therefore, alphanumeric data.
Conclusions: There is a growing need of describing the nature of data, as well as the formats and extensions of files. This kind of descriptive metadata would be valuable to potential users, as it would allow a greater understanding of the context of the data, focusing on data reuse.
Keywords: scientific data; datasets; data repositories; formats and extensions; survey.

1 INTRODUÇÃO

Dentro do processo evolutivo da ciência, observaram-se períodos na história humana que se destacaram pela maneira que a prática científica foi conduzida. Num primeiro momento, foram feitos experimentos e observações sobre o comportamento natural das coisas do mundo físico e passíveis de análise, o que se denominou ciência empírica. O empirismo trabalha variáveis distintas, em ambientes controlados ou não, e que busca validar ou refutar correlações, como causa-efeito.

Como consequência da experiência empírica, surgiu então um paradigma apoiado na observação e experimentação com a pretensão de testar hipóteses. Assim, hipóteses, teorias e leis foram fruto desse processo científico que se manteve e perdurou por séculos, até sofrer modificações a partir de meados do século XX.

Desde então, a humanidade usufrui da computação para gerar simulações e análises de dados em seus experimentos, observações e testes de hipóteses, contando inicialmente com o uso de computadores robustos e com baixas capacidades de armazenamento e processamento. Com o passar das décadas, essas máquinas tiveram sua capacidade de processamento melhorada de forma a assistir mais efetivamente os pesquisadores em seu fazer diário, melhorando o tempo gasto e a qualidade da análise dos dados coletados.

É a partir de meados da década de 1940, com o cenário do fim da segunda guerra mundial e começo da guerra fria, que o mundo observou o início de uma corrida armamentista e tecnológica, além de disputas de influências e de territórios entre os Estados Unidos e a antiga União Soviética, o que impactou na velocidade do desenvolvimento computacional e industrial.

Surgem também outros avanços em Ciência e Tecnologia (C&T) no período pós-segunda guerra, tais como pesquisas no uso de energia nuclear, a terceira revolução industrial, as Tecnologias de Informação e Comunicação (TICs), entre elas a *Web* e a *Internet*, ou seja, serviços estratégicos de inteligência numa disputa de poder entre potências.

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

A computação nesse cenário tem grande influência na aceleração do desenvolvimento da C&T em nível global, uma vez que se torna o pilar dos produtos e serviços da segunda fase do século XX. Por consequência e com a mesma velocidade, a informação e a comunicação se tornam menos analógicas, adentrando, ao final de um milênio, na era digital.

Novos desafios às técnicas de tratamento dos dados e da informação emergem por conta do aumento exponencial do volume informacional. Essa se torna uma oportunidade para a Biblioteconomia e a Ciência da Informação buscarem abordar questões impostas pela digitalização, digitalização e aumento do acesso à informação. As bibliotecas virtuais e digitais surgem como respostas às necessidades de acesso remoto a serviços e produtos de bibliotecas tradicionais, sendo consequência da premissa bibliotecária de acesso universal ao conhecimento humano.

Portanto, é nesse contexto em que o atual trabalho se encaixa, a partir do ponto em que reflete o lugar das bibliotecas em um universo em expansão, mais especificamente dentro do propósito acadêmico, em que se observa o desenrolar de uma fase que busca tratar tecnicamente o material científico produzido em laboratórios de pesquisa de programas de pós-graduação de universidades e institutos, assim como de instituições e agências de fomento e pesquisa científica, públicas e privadas.

Assim, bibliotecas universitárias e/ou especializadas que desejam salvaguardar, organizar e dar acesso à produção científica de sua instituição por meio digital devem desenvolver ambientes virtuais conhecidos como repositórios digitais institucionais. Esses repositórios podem ser classificados em documentais, bibliográficos, de dados, ou híbridos.

Com o aumento da atividade humana em ambientes digitais, cresce o número de documentos que são digitalizados (cópias) e que nascem nesse formato (nato digitais), que por sua vez necessitam de adequada gestão documental na busca por sua preservação, organização e recuperação. Dessa maneira, surge a preservação digital como campo de atuação e pesquisa científica na Biblioteconomia e Ciência da Informação. Fundamentada na atividade bibliotecária, a preservação digital busca salvaguardar, organizar,

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

disponibilizar, compartilhar, disseminar e permutar ativos digitais que compõem e dão sustento ao conhecimento humano.

Vários são os recursos atualmente disponíveis que procuram reduzir os problemas de organização e acesso a documentos e informações em ambientes digitais, como os esquemas ou padrões de metadados, os quais possibilitam a universalização (padronização) da descrição do conteúdo de objetos informacionais tendo como fim a recuperação da informação, os relacionamentos e a importação de registros (interoperabilidade) de um sistema a outro, pois padrões universais de linguagem e codificação podem ser reconhecidos por sistemas inteligentes. Esses recursos servem aos propósitos de acervos digitais em ascensão como os repositórios.

Dessa forma, busca-se conhecer a natureza do conteúdo dos dados arquivados em repositórios digitais, e consequentemente, os formatos e extensões de arquivos que compõem os conjuntos de dados gerados ao longo de um processo científico. Em recente publicação, os pesquisadores Sales e Sayão (2019) apresentam uma taxonomia (Figura 1) que se refere à natureza dos dados científicos.

Figura 1 – Taxonomia quanto à natureza dos dados

Número	Medidas Resultado de levantamento Fórmula Equação Algoritmo	**Textual**	Metadado Questionário Entrevista Anotação Certificado Caderno de laboratório Transcrição Correspondência Diário Caderno de campo
Multimídia	Imagem Vídeo Áudio Animação Filme Fotografia		
Software	Base de dados Simulação Códigos	**Artefato**	Espécime Amostra Maquete Phantom/Manequim
Visualização	Tabelas Gráficos Diagrams Modelo em 3D Modelo reduzido Desenho	**Processo**	Procedimentos operacionais padronizados Workflow Protocolo Teste

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

Fonte: Sales e Sayão, 2019, p. 41.

Por dados científicos, entende-se que são as menores unidades informacionais (devido à natureza granular) coletadas, preservadas em documentos de forma analógica ou digital, estruturadas e analisadas por métodos científicos, que têm seu fim na produção e manutenção do conhecimento. Esses dados brutos possuem complexidade tal que são, muitas das vezes, apenas compreendidos por pesquisadores, máquinas e programas que os coletaram, geraram, simularam ou processaram, o que acaba se tornando um problema aos profissionais atuantes no campo da preservação digital, geralmente fora do ambiente e do domínio da pesquisa.

Na longa cauda da ciência, a diversidade da natureza dos dados coletados e agrupados é refletida em um número plural de arquivos salvos em diferentes formatos e extensões, gerando assim um ou mais conjuntos de dados científicos. Esses conjuntos podem ser definidos então como uma compilação de arquivos digitais que são gerados no processo científico-investigativo, e que por sua vez possuem dados de natureza e conteúdo heterogêneo. Ou seja, a natureza dos dados está diretamente relacionada aos formatos e extensões dos arquivos digitais que os agrupam.

> Quanto à natureza dos dados – retrata a grande diversidade e heterogeneidade de tipos de dados que podem ser originados no ambiente de pesquisa em termos de formatos, mídias, suportes, expressões, arcabouço tecnológico, etc (SALES; SAYÃO, 2019, p. 43-44).

O meio para que os dados científicos encontrem seu fim está na comunicação científica (a ponta do *iceberg*), que reporta, na maioria das vezes, resultados condensados por meio da publicação de artigos científicos em periódicos especializados, da apresentação de trabalhos em congressos e respectiva publicação em anais, entre outros. “[...] A publicação dos resultados da sua pesquisa, e a literatura publicada é apenas a ponta do *iceberg* de dados. [...] Por *iceberg* de dados quero dizer que há muitos dados que são coletados, mas não tratados ou publicados de forma sistemática.” (HEY; TANSLEY; TOLLE, 2009, p. xvii, tradução nossa).

Durante todo o processo de investigação científica, pesquisadores se utilizam de diversos meios para documentar seus achados (o uso de cadernos de pesquisa para registro de anotações em meio analógico ou eletrônico, entre outros) e arquivar seus

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

documentos contendo dados coletados (computadores, celulares, hds, nuvem, etc.). “No século XX, os dados nos quais teorias científicas eram baseadas ficavam geralmente ocultos em *notebooks* científicos individuais ou, por alguns aspectos da ‘grande ciência’, armazenados em mídia magnética que eventualmente se torna ilegível.” (HEY; TANSLEY; TOLLE, 2009, p. xi, tradução nossa). Logo, esses dados por vezes se perdem em meio à má gestão e organização ou ao fim de um ciclo investigativo.

> [...] os cadernos [...] muitas vezes ficam limitados às paredes dos laboratórios, acentuando uma cultura de segredo que, cada vez mais, precisa ser discutida e superada. Os dados que sustentam uma pesquisa [...] geralmente ficam adormecidos, armazenados em computadores ou mídias pessoais [...]. [...] cadernos de laboratório [...] contistuem a espinha dorsal da guarda de registros, gestão de dados, análises iniciais e interpretação de resultados em pesquisas (ROCHA; SALES; SAYÃO, 2017, p. 3).

Uma reflexão sobre o tratamento e organização das informações desses dados consiste na realidade de que esses geralmente se constituem em conjuntos de objetos informacionais, que integrados podem ser processados como um fundo arquivístico. Caso esse tratamento seja individualizado, como na organização da informação de dados bibliográficos, o relacionamento entre os objetos informacionais de uma determinada pesquisa pode se perder, prejudicando a integridade das informações contidas nesses conjuntos de dados gerados em determinada investigação científica.

São objetos de estudo desta investigação: dados científicos; conjuntos de dados científicos; e repositórios digitais institucionais de dados científicos. A pergunta que se pretendeu ao final responder foi: qual a natureza dos conjuntos de dados científicos arquivados em repositórios digitais institucionais oriundos do continente sul-americano? Os objetivos foram identificar os repositórios de dados científicos criados e geridos por Instituições de Ensino Superior (IES) e/ou agências de pesquisa e fomento sul-americanas; identificar e descrever os formatos e extensões dos arquivos que compõem os conjuntos de dados científicos depositados nesses repositórios.

A Ciência da Informação (CI) “é, por natureza, interdisciplinar, embora suas relações com outras disciplinas estejam mudando” (SARACEVIC, 1996, p. 42), e que tem como

Encontros Bibli

proposta a organização do conhecimento humano. Logo, lidar com o fenômeno dos dados sob essa ótica é assistir a comunidade acadêmica em seus diversos campos, assim como a sociedade civil em suas dificuldades tecnológicas e digitais, disponibilizando meios de acesso ágeis e assertivos num ambiente digital cada vez mais nebuloso e caótico.

Um dos deveres do profissional da informação nesse crescente contexto é pôr em prática os princípios do acesso aberto, que busca expandir o uso dos recursos científicos de forma a agilizar e ampliar o processo de construção do saber.

Esse movimento de abertura também equilibra a desigualdade elitista científica e dá retorno dos investimentos de ordem pública à sociedade, uma vez que se passa a exigir de pesquisadores acesso amplo a resultados e dados de pesquisas científicas financiadas pelo Estado. Para que se possa tornar tais ideais realidade, é que se justifica o investimento de recursos em ambientes e normas que se destinam a desenvolver melhores práticas com fim na acessibilidade do conhecimento construído a partir da pesquisa científica.

É a partir do texto de Borgman, Scharnhorst e Golshan (2019) que se justifica a investigação proposta, da percepção das necessidades de exploração dos assuntos e da procura por soluções criativas à abordagem dos problemas que envolvem os dados científicos.

2 REFERENCIAL TEÓRICO

Organizações do conhecimento estão incorporando práticas destinadas ao uso dos dados em seu dia a dia, uma vez que esses são a base da pirâmide do conhecimento. Considerando a visão de mundo dos autores e o trabalho de Zins (2007), entende-se a gestão do conhecimento como a gestão do conhecimento tácito ou cognitivo; a gestão da informação como a gestão do conhecimento documentado, estruturado e explícito; e a gestão de dados como a gestão do conhecimento não estruturado e explícito. O primeiro modelo que define dados, informação e conhecimento dentro do estudo de Zins (2007) sugere que o conhecimento está no Domínio Subjetivo (DS), o qual traz entendimento que se trata de um fenômeno intrínseco, não podendo ser resultado de dados por si só, embora possa ser criado por meio da análise e interpretação dos dados (internalização), transformando-os em capital intelectual das organizações.

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

Por dado eletrônico, pode-se compreender todo dado coletado que venha a ser digitalizado, ou seja, registrado mais comumente por anotações em meio físico e transformado em digital (eletrônico analógico), ou dados nato digitais, que nascem em meio digital por coletas feitas por humanos, máquinas, sensores, robôs, entre outros (eletrônico digital). “O dado digital é todo aquele armazenado na forma de ‘zeros e uns’, independente de sua estrutura. [...] informação estruturada em planilha eletrônica é dado. Vídeos digitais, postagens em redes sociais, dados de acelerômetros em um celular [...]” entre outros (AMARAL, 2016, p. 4).

Ilharco (2004) levanta também alguns questionamentos que tangem à informação: “[...] a sociedade da informação é a sociedade de quê? O que é a informação? Quais os seus princípios de base? Às suas relações com fenômenos próximos, como a comunicação, a ação, o conhecimento, a nova tecnologia?”. (ILHARCO, 2004, p. 1). Essas são perguntas relevantes e que também podem ser utilizadas e adaptadas ao contexto dos dados, conforme proposto por Borgman (2015).

No centro do problema da curadoria de dados estão perguntas como: quais dados são dignos de preservação, por quê, para quem, por quem, e por quanto tempo? Quais responsabilidades da curadoria de dados devem recair sobre investigadores, comunidades, universidades, agências financiadoras, ou outros *stakeholders*? (BORGMAN, 2015, p. 29, tradução nossa).

Borgman, Scharnhorst e Golshan (2019) definem os *stakeholders* como sendo acadêmicos e equipes que produzem dados, agências de financiamento que provêm recursos à condução de pesquisas, universidades e outras instituições de pesquisa, produtores de políticas de pesquisa em organizações públicas e privadas, usuários desses dados, bibliotecas e arquivos que podem adquirir e gerir dados (BORGMAN; SCHARNHORST; GOLSHAN, 2019, p. 888, tradução nossa).

Dados são parte de um fenômeno tanto quanto a informação também o é, pois ela se origina a partir da análise e interpretação de dados ou conjuntos de dados em determinado formato e/ou mídia, que por sua vez são compreendidos dentro de um ou vários contextos. Dados em grande volume e variedade são conhecidos como *Big Data*, provenientes de n fontes, estruturados ou não, públicos ou privados. Wamba e outros

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

(2015) definem "[...] '*Big Data*' como uma abordagem holística para gerir, processar e analisar os 5 Vs [...] com o intuito de criar *insights* acionáveis para a entrega de valor sustentado, medindo o desempenho e estabelecendo vantagens competitivas". (WAMBA *et al*., 2015, p. 235, tradução nossa).

De acordo com Storey e Song (2017), *Big Data* se refere a grandes quantidades de dados, os quais organizações são capazes de capturar e analisar de forma significativa para que assim decisões baseadas em dados possam ser tomadas. O volume de dados tem crescido exponencialmente nas últimas décadas, ao ponto em que o gerenciamento desse ativo (dados) por meios tradicionais não seja mais possível (STOREY; SONG, 2017, p. 50, tradução nossa).

Profissionais de áreas distintas (Ciência da Computação, Estatística, Sistemas de Informação, Ciência da Informação, entre outras) estão lidando ou sendo imbuídos com a tarefa desafiadora de gerir esses conjuntos de dados, focando seus esforços na criação de uma cultura organizacional inovadora, ao desenvolver melhores práticas, mudando assim as formas de pensar e resolver problemas com vistas à vantagem competitiva. "O *Big Data* vai oferecer muitas oportunidades. Essas oportunidades virão de duas formas: vantagem competitiva ou criação de produtos e/ou serviços orientados a dados." (AMARAL, 2016, p. 11). Dados são, portanto, ativos a serem explorados.

> [...] hoje dados são produzidos massivamente em redes sociais, comunidades virtuais, blogs, dispositivos médicos, TVs digitais, cartões inteligentes, sensores em carros, trens e aviões, leitores de código de barra e identificadores por radiofrequência, câmeras de vigilância, celulares, sistemas informatizados, satélites, entre outros (AMARAL, 2016, p. 8).

Concernente à proteção de ativos intangíveis em instituições do conhecimento, Amaral (2016) coloca que "[...] o dado, enquanto existente, terá a ele associado questões de segurança, privacidade e qualidade. Ainda, dados dentro de uma organização são governados por políticas e procedimentos, mesmo que informais". (AMARAL, 2016, p. 5). Uma política institucional voltada aos dados deve se preocupar com seu ciclo de vida, pois a Ciência da Dados não se resume à coleta e análise de dados. Desse modo, pode-se usufruir da máxima capacidade dos dados, seja por uso ou reúso.

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

Dados científicos são resultado do fazer científico, considerando que essa prática se utiliza da computação como meio para um fim, fazendo computadores processarem simulações, gerando *petabytes* de dados no mundo inteiro todos os dias. A ciência não apenas simula, mas também registra experimentos em vários campos de pesquisa, mirando sempre a criação de novos conhecimentos baseados em análise e interpretação dos dados.

A gestão e curadoria de dados científicos envolvem serviços, ferramentas e infraestruturas do conhecimento que abrangem o ciclo de vida da pesquisa científica. Pesquisadores necessitam de apoio nos processos de planejamento, gestão, organização, documentação e preservação de seus conjuntos de dados, bem como em questões relacionadas a licenças e propriedade intelectual, quando na intenção do compartilhamento de seus dados.

A gestão e curadoria de dados científicos são serviços fundamentais à organização, preservação, recuperação e reúso desses dados. A gestão e curadoria trazem consigo trabalho intelectual e abordagem técnica no manuseio de dados. A gestão de dados está para a organização, assim como a curadoria está para a preservação em longo prazo, porém não limitada a ela.

Destaca-se o termo ‘competência em dados’ (*data literacy*) apresentado por Koltay (KOLTAY, 2015, p. 401, tradução nossa). Esse termo é conhecido de áreas como a Ciência da Computação, Sistemas de Informação, entre outras. No entanto, nesse ambiente computacional, *data literacy* denota a formação e o desenvolvimento (letramento) de habilidades técnicas em análises quantitativas (matemáticas e estatísticas) de dados, e a apresentação de resultados de formas mais inteligíveis à compreensão humana. Na CI se trabalha outro termo, ‘competência informacional’, que não pode ser adaptado à ‘competência em dados’, ainda que se reconheça que dado gere informação, e informação gera conhecimento.

O campo da competência informacional pode não concordar com a definição precisa de competência informacional, mas a maioria se utiliza do termo ‘competência

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

informacional' ao invés de 'instrução da biblioteca[1]' ou 'fluência informacional'. O que tem sido chamado 'competência em dados' envolve competência estatística, raciocínio quantitativo ou competência quantitativa, e *numeracy*[2], todos significando a mesma coisa (HUNT, 2004, p. 14, tradução nossa).

O termo 'competência em dados' na Ciência da Informação necessita ser definido a partir de conceitos e técnicas praticadas por profissionais com *expertise* em organização e preservação de dados digitais (científicos, governamentais, mercadológicos, financeiros, etc.). Isso significa conhecer e mapear o fluxo de trabalho científico e necessidades da comunidade alvo; depositar e descrever dados de forma padronizada com uso de vocabulários controlados e padrões de metadados universais e de domínio; atribuir identificadores persistentes a objetos digitais (conjuntos de dados); vincular publicações e dados a outros relacionados; conhecer leis de proteção a dados e propriedade intelectual; dominar ferramentas e *software* de preservação digital; entre outros.

Repositórios de documentos bibliográficos digitais no Brasil são uma realidade, enquanto repositórios de dados científicos se encontram em processo de estudo e discussão, embora muito ainda tenha que ser estabelecido. Profissionais que buscarem trabalhar com dados encontrarão solo fértil em laboratórios de pesquisa, agências de fomento, unidades de informação e em organizações do conhecimento, assim como em equipes heterogêneas de Ciência de Dados, passando dessa maneira a encarar o desafio de possuir habilidades computacionais aliadas à *expertise* de domínio.

3 PROCEDIMENTOS METODOLÓGICOS

Precedendo a fase da coleta de dados, que ocorreu entre 2019 e 2020, a estratégia de busca para encontrar repositórios brasileiros foi o uso dos termos vernáculos 'repositório' e 'dados científicos' na ferramenta de pesquisa do *Google*, pois se pretendia obter um

[1] O termo original utilizado pela autora é *library instruction*, que também poderia ser traduzido como 'instrução bibliotecária', porém se confundiria com o profissional dessa área.
[2] Como o termo *literacy* significa letramento, o termo *numeracy* denota competência com números e matemática.

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

primeiro contato com possíveis repositórios de dados científicos associados a instituições brasileiras.

Os resultados foram insatisfatórios devido à dificuldade de identificação desses repositórios entre as diversas páginas recuperadas pelo *Google*. Foi então que se obteve conhecimento da base RE3DATA e da possibilidade da recuperação de repositórios digitais de dados científicos por meio de *browsing*[3] entre as categorias: Assunto, Tipo de Conteúdo (ex.: imagem, código de fonte, entre outros) e País.

A partir daí, fez-se levantamento dos repositórios brasileiros na referida base, utilizando-se de estratégia de busca 'por país' com objetivo de conhecer rapidamente os repositórios indexados. Após o primeiro contato, foi preciso estabelecer critérios de seleção para que se sustentasse critérios científicos.

À época, oito repositórios de dados científicos foram encontrados indexados como brasileiros na base RE3DATA, porém apenas a metade foi selecionada a partir dos critérios seguintes: a) repositórios geridos por grupos ou instituições brasileiras em ambientes controlados (repositórios institucionais, não bancos de dados e/ou arquivos pessoais de pesquisadores); b) possibilidade de busca de registros (conjuntos de dados depositados em repositórios) por browsing e; c) repositórios ativos e de acesso aberto.

Foram selecionados: 1) PPBio Data Repository (Repositório de Dados de Levantamentos Biológicos); 2) Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) Dataverse Network; 3) Centro de Documentação e Acervo Digital da Pesquisa (CEDAP) Research Data Repository e; 4) Base de Dados Científicos da Universidade Federal do Paraná (UFPR).

Feita a análise dos repositórios, identificou-se problema nos *links* de acesso ao repositório CEDAP, de responsabilidade da Universidade Federal do Rio Grande do Sul (UFRGS). Para além disso, reconheceram-se inconsistências de acesso aos registros e seus conjuntos de dados. Após inúmeras tentativas, decidiu-se por excluir o repositório do estudo, considerando o levantamento feito insuficiente para fins comparativos dentro da proposta de pesquisa.

[3] Técnica de busca exploratória onde se procura de unidade a unidade em um catálogo ou acervo, até que se encontre o desejado ou se esgote todas as opções.

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

Portanto, decidiu-se por aumentar o escopo da investigação em nível continental. Num novo acesso à base RE3DATA, foram encontrados 12 repositórios de dados divididos entre outros 4 países sul-americanos: Argentina, Chile, Colômbia e Peru. À vista dos critérios de seleção anteriormente estabelecidos, com exceção ao primeiro que agora se expande aos demais países do continente, selecionaram-se outros 5 repositórios de origem chilena, colombiana e peruana. Nenhum repositório argentino atendeu aos critérios estabelecidos.

Foram incluídos nessa fase: o CIAT Dataverse (Colômbia); o Portal de Datos Abiertos de la Pontificia Universidad Católica del Perú (Peru); o Repositorio de datos del Ministerio de Educación del Perú (Peru); o Repositorio Institucional USIL (Peru); e o Repositorio de Datos de Investigación de la Universidad de Chile.

Obteve-se a partir daí uma população (N) de 1.115 conjuntos de dados científicos. Esse valor é resultado da soma dos conjuntos identificados em cada repositório. A partir da Amostragem Aleatória Estratificada, obteve-se o valor da amostra (n) igual a 258 conjuntos de dados, que corresponde a 23,15% da população (N). Investigou-se os conjuntos da amostra de forma aleatória, a partir do peso de cada valor por repositório. O repositório do PPBio possuía 403 conjuntos de dados, tendo sua amostra calculada em 145,7 conjuntos; o do USIL hospedava 256 conjuntos (amostra: 58,8 conjuntos); CIAT, 189 conjuntos (amostra: 32); IBICT, 139 conjuntos (amostra: 17,3); PUC Peru, 44 conjuntos (amostra: 1,7); MEC Peru, 44 conjuntos (amostra: 1,7); UFPR, 31 conjuntos (amostra: 0,9); Universidade do Chile, 9 conjuntos (amostra: 0,1). Os números de conjuntos foram arredondados em números inteiros. Assim, a amostra da UFPR foi um (1) conjunto, e o da Universidade do Chile, zero (0), sendo dessa forma eliminado estatisticamente.

A coleta dos dados foi registrada em planilhas nas categorias: Número do Depósito (número não fixo nos repositórios); URL ou URI do Depósito; Identificador Persistente do Depósito; extensões encontradas nos arquivos observados. As categorias criadas com as

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

extensões dos arquivos encontrados tinham o intuito de: 1) obter retorno estatístico por formato e extensão incidente; e 2) registrar, investigar e descrever cada um deles.

3.1 Observações aos procedimentos metodológicos

Em uma segunda fase do levantamento de dados com fim ao alcance de outros objetivos específicos não mencionados, exclui-se da investigação o repositório de dados científicos do IBICT. No entanto, sua exclusão não altera os resultados aqui apresentados. Para mais informações em relação ao processo metodológico, consultar Rodrigues, Dias e Lourenço (2022, p. 309-311).

4 ANÁLISE DOS RESULTADOS

A análise desses dados foi executada em uma etapa que identifica e descreve os formatos e extensões dos arquivos que compõem os conjuntos de dados.

4.1 Análise exploratória dos formatos e extensões dos arquivos avaliados

4.1.1 Repositórios brasileiros

Ao conjunto que apresentou mais de um arquivo com a mesma extensão foi contabilizado apenas uma ocorrência em tabela, desse modo, os números registrados são referentes à incidência de cada extensão por depósito. A Tabela 1 apresenta as extensões encontradas, assim como a estatística resultante de sua incidência nos repositórios brasileiros em questão.

Tabela 1 – Extensões dos arquivos em conjuntos de dados científicos nos repositórios brasileiros

Repositórios brasileiros	.txt	.doc	.docx	.pdf	.csv	.xlsx	.gpx	.gdb	.data	.zip	.rar	.xml	.rdf	**Total por repositório**
PPBio *Data Repository*	119	2	N/A	12	26	2	1	1	2	N/A	N/A	146	N/A	311
	38,26%	0,64%	N/A	3,86%	8,36%	0,64%	0,32%	0,32%	0,64%	N/A	N/A	46,95%	N/A	100%
IBICT *Dataverse Network*	N/A	1	4	13	N/A	N/A	N/A	N/A	N/A	1	1	N/A	N/A	20
	N/A	5,00%	20,00%	65,00%	N/A	N/A	N/A	N/A	N/A	5,00%	5,00%	N/A	N/A	100%
B. de Dados Científicos da UFPR	N/A	N/A	N/A	N/A	N/A	1	N/A	N/A	N/A	N/A	N/A	N/A	1	2
	N/A	N/A	N/A	N/A	N/A	50%	N/A	N/A	N/A	N/A	N/A	N/A	50%	100%
Total de extensões por conjunto	119	3	4	25	26	3	1	1	2	1	1	146	1	333
	35,74%	0,90%	1,20%	7,51%	7,81%	0,90%	0,30%	0,30%	0,60%	0,30%	0,30%	43,84%	0,30%	100,00%

Fonte: elaborado pelos autores, 2020.

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

A maioria (8 em 13) dos formatos e extensões encontrados são parte da rotina de pessoas que utilizam computadores com certa frequência, como os arquivos de natureza textual (.txt, .doc, .docx, e .pdf), visual alfanumérica (.csv e .xlsx), ou aqueles que fazem compactação de outros arquivos (.zip e .rar). Não foi objetivo deste trabalho investigar o conteúdo de arquivos compactados encontrados durante o levantamento.

As demais extensões apresentadas na Tabela 1 precisaram ser investigadas para que se pudesse identificar e descrever a sua natureza. Dessa forma, o Quadro 1 apresenta a definição de cada extensão para melhor entendimento dos conteúdos salvos nos arquivos analisados.

Quadro 1 – Definições de extensões não usuais

Extensão	Definição
.gpx	Um arquivo GPX é um arquivo de dados de GPS salvos no formato GPS *Exchange*, o qual é um padrão aberto que pode ser livremente usado por programas de GPS. Contém dados de localização representados por longitude e latitude que pode incluir pontos de notificação, rotas e trajetos. Arquivos GPX são salvos em formato XML, de forma que permite dados de GPS serem mais facilmente importados e lidos por múltiplos programas e serviços *Web*.
.gdb	Um arquivo GDB é um arquivo de base de dados criado por *MapSource*, um aplicativo editor de rotas GPS e de planejamento de viagens. Contém pontos de notificação, rotas, e trajetos que podem ser transferidos a um dispositivo de navegação Garmin. Arquivos GDB são similares aos universalmente transferíveis arquivos .gpx.
.data	Um arquivo DATA é um arquivo de dados usado por *Analysis Studio*, um programa de mineração de dados e análise estatística. Contém dados minerados em um texto simples, formato tabelar delimitado, incluindo um cabeçalho de arquivo do *Analysis Studio*. Arquivos DATA são comumente usados para armazenar dados para análise de dados offline quando não conectado a um servidor *Analysis Studio*, contudo pode também ser usado em modo *online*.
.xml	Um arquivo XML é um arquivo de dados XML. É formatado quase como um documento .HTML, mas usa *tags* personalizadas para definir objetos e os dados dentro de cada objeto. Arquivos XML podem ser pensados como uma base de dados baseada em texto.
.rdf	Um arquivo RDF é um documento escrito em linguagem RDF, que é usada para representar informação sobre recursos na *Web*. Contém informação sobre um *Website* em um formato estruturado chamado metadado. Arquivos RDF podem incluir um mapa de site, um registro de atualizações, descrições de páginas, e palavras-chave.

Fonte: Sharpened Productions, 2020, tradução nossa.

As extensões de arquivos com maior frequência no repositório do PPBio estão em formato de texto (.txt) com 38,26%, e em formato XML com incidência em todos os conjuntos investigados (146). Arquivos .xml guardam os dados, metadados e suas *tags* anotadas referentes aos conjuntos depositados no repositório, o que facilita a importação de registros a partir de seu uso (interoperabilidade). O repositório do PPBio foi aquele com o maior volume de conjuntos de dados científicos investigado (403), e conta basicamente com um acervo textual e numérico tabular. Os conjuntos de dados associados ao PPBio estão disponibilizados na infraestrutura tecnológica provida pela iniciativa DataONE, comunidade internacional que compartilha acervos de dados científicos nos campos da

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

Biologia e das Ciências da Natureza. O DataONE utiliza a aplicação Metacat (escrita em Java e mantida por servidor Apache Tomcat) que gerencia acervos das áreas supracitadas.

Metacat é um catálogo flexível de metadados de código-fonte aberto e um repositório de dados com foco em dados científicos, particularmente advindos da Ecologia e das Ciências do Meio Ambiente. Ele adota a linguagem XML como uma sintaxe comum para representar o vasto número de padrões de conteúdo de metadados que são relevantes à ecologia e outras ciências. Ademais, o Metacat é uma base de dados XML genérica que permite armazenamento, consulta, e recuperação de documentos XML arbitrários sem conhecimento prévio do esquema XML. Está sendo utilizado extensivamente ao redor do mundo para gerir dados do meio ambiente (DATAONE, 2020, tradução nossa).

No que diz respeito ao processo de inserção e descrição de dados, o usuário desse sistema pode inserir metadados por meio do Morpho, *software* editor de metadados também utilizado pelo DataONE em assistência à gestão e curadoria de acervos digitais no campo da Ecologia.

Morpho é um programa que pode ser usado para inserir metadados, que são então armazenados em um arquivo que obedece à especificação da *Ecological Metadata Language* (EML). Informação sobre pessoas, lugares, métodos de pesquisa, e atributos de dados estão entre os metadados coletados. Dados podem ser armazenados com os metadados no mesmo arquivo. Ele permite ao usuário criar um catálogo local de dados e metadados que podem ser consultados, editados e visualizados. O Morpho também faz interface com o servidor Metacat chamado *Knowledge Network for Biocomplexity*[4] (KNB), que permite cientistas enviarem, baixarem, armazenarem, consultarem e visualizarem dados e metadados públicos (DATAONE, 2020, tradução nossa).

Adiante, o repositório do IBICT teve incidência maior de arquivos .pdf, que são arquivos em formato de texto protegidos de qualquer alteração, como devem ser publicações científicas, trabalhos acadêmicos e documentos oficiais. São esses os casos encontrados nos arquivos acessados. Cabe a crítica ao arquivamento de trabalhos completos em repositórios de dados científicos que não descreve o endereço onde esses

[4] Rede de Conhecimento à Biocomplexidade (RCB).

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

foram primariamente publicados na *Web* (ausência de URI ou identificador persistente que os localize, por exemplo).

Por conseguinte, não faz sentido se utilizar de repositórios de dados científicos para depositar apenas arquivos de textos referentes à comunicação científica, duplicando algumas vezes os arquivos sem os interconecta-los. Exemplo disso seria uma instituição que possui um repositório direcionado à gestão e curadoria de documentos bibliográficos e também um repositório de dados científicos, onde uma mesma publicação é depositada em ambos repositórios, sendo que o segundo também recebe – ou deveria receber – o depósito dos dados relativos a ela. Sem a devida indicação de correlação e coexistência entre os depósitos por meio de *link*, haverá duplicação despercebida por parte de seus gestores.

Até onde se pôde apurar antes de sair do ar (durante período investigativo), o IBICT manteve seu repositório de dados científicos no *software* Dataverse, uma comunidade que compartilha repositórios de todo o globo nas diversas áreas do conhecimento humano e é um projeto de responsabilidade da Universidade de Harvard (Cambridge, Massachusetts).

Dataverse é um aplicativo *Web* de código aberto para compartilhar, preservar, citar, explorar, e analisar dados científicos. Facilita o processo de depósito, publicação e recuperação dos dados, tornando-os disponíveis a terceiros, e permite replicar trabalhos mais facilmente. Pesquisadores, periódicos, autores de dados, editores, distribuidores de dados, e instituições afiliadas recebem crédito acadêmico e visibilidade na *Web*. Um repositório Dataverse é uma instalação deste *software*, que então hospeda múltiplos arquivos virtuais chamados Dataverses. Cada dataverse contém conjuntos de dados, e cada conjunto contém metadados descritivos e arquivos de dados (incluindo documentação e código que acompanham os dados). Como um método de organização, os dataverses podem também conter outros dataverses (DATAVERSE, 2020, tradução nossa).

O último repositório analisado nesta etapa foi a Base de Dados Científicos da Universidade Federal do Paraná (UFPR). Observou-se que ela faz parte do Repositório Digital Institucional da UFPR, onde se encontram depósitos de documentos variados produzidos pela universidade, tais como Teses e Dissertações, Trabalhos de Especialização, Trabalhos de Graduação, entre outros. Ambos os repositórios compartilham o mesmo ambiente e são geridos por meio do mesmo *software* de código-

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

fonte aberto chamado DSpace. O “DSpace foi desenvolvido para possibilitar a criação de repositórios digitais com funções de armazenamento, gerenciamento, preservação e visibilidade da produção intelectual, permitindo sua adoção por outras instituições em forma consorciada federada” (IBICT, 2019).

Como o DSpace não se destina à criação de comunidades de compartilhamento de repositórios e conjuntos de dados científicos entre diferentes instituições como o DataONE e o Dataverse, a sua utilização reduz as possibilidades de acesso a dados científicos em potencial dentro de um cenário de consulta exploratória. Apesar do DSpace não ter sido originalmente projetado para lidar com dados científicos, o mesmo pode ser customizado para tal fim, se assim for necessário.

Além do repositório de dados científicos da UFPR possuir o menor acervo dentre os repositórios brasileiros qualificados à investigação, ele também se confunde com o repositório institucional durante a navegação. Pôde-se verificar que 24 dos seus 30 conjuntos de dados depositados possuem arquivos .rdf que contêm *tags* e textos referentes ao uso da licença *Creative Commons*.

4.1.2 Demais repositórios sul-americanos

Os critérios apresentados na primeira etapa de análise dos dados foram replicados aqui, de forma a padronizar as ações para que se pudesse comparar os resultados obtidos. Assim, construiu-se a Tabela 2 nos mesmos moldes da anterior, sendo ela a referência das análises desta subseção.

Tabela 2 – Extensões dos arquivos em conjuntos de dados científicos nos demais repositórios sul-americanos

Demais Repositórios sul-americanos	.txt	.docx	.pdf	.csv	.xls	.xlsx	.xlt	.pptx	.jpg	.png	.tif	.tab	.dta	.zip	.sav	.do	Total por Repos.
CIAT (Colômbia)	1	3	9	1	19	2	1	1	1	1	1	23	2	7	N/A	2	74
	1,35%	4,05%	12,16%	1,35%	25,68%	2,70%	1,35%	1,35%	1,35%	1,35%	1,35%	31,08%	2,70%	9,46%	N/A	2,70%	100%
P. *Datos Abiertos* PUC (Peru)	N/A	N/A	2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	2	N/A	N/A	N/A	N/A	4
	N/A	N/A	50,00%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	50,00%	N/A	N/A	N/A	N/A	100%
R. *Datos* MEC (Peru)	N/A	N/A	2	2	N/A	1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	5
	N/A	N/A	40%	40%	N/A	20%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	100%
USIL (Peru)	N/A	N/A	2	N/A	1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	1	1	54	N/A	59

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

	N/A	N/A	3,39%	N/A	1,69%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	1,69%	1,69%	91,53%	N/A	100%
Total formatos e extensões	1	3	15	3	20	3	1	1	1	1	1	25	3	8	54	2	142
	0,70%	2,11%	10,56%	2,11%	14,08%	2,11%	0,70%	0,70%	0,70%	0,70%	0,70%	17,61%	2,11%	5,63%	38,03%	1,41%	100%

Fonte: elaborado pelos autores, 2020.

Torna-se relevante observar que entre os demais países investigados, apenas a Colômbia e o Peru possuem repositórios qualificáveis a este estudo. Também se deve notar que o Peru detém 3 dos 6 repositórios investigados em toda a América do Sul. Em contrapartida, o Brasil se responsabiliza por dois repositórios que juntos detêm o maior volume de dados científicos arquivados. Entende-se que o volume não está diretamente relacionado à qualidade da gestão desses repositórios, tampouco diz respeito à situação financeira de uma instituição, uma vez que dados não são adquiridos por meio de compra, como livros e outros materiais. Entretanto, o volume de dados em um repositório pode indicar determinada maturidade de seus serviços.

A informação de que o repositório do PPBio (Brasil) tem o maior volume de conjuntos de dados compartilhados pode indicar maior envolvimento dos pesquisadores por meio da conscientização do compartilhamento de seus dados, ou compartilhamento obrigatório via política de financiamento (como exigido dos pesquisadores financiados pela FAPESP).

Entre os conjuntos investigados depositados no repositório da Universidad San Ignacio de Loyola, notou-se a incidência majoritária da extensão .sav, que é descrita no Quadro 2. Outras extensões não usuais encontradas também estão presentes no quadro mencionado.

Quadro 2 – Outras definições de extensões não usuais

Extensão	Definição
.tif	Um arquivo TIF é um arquivo de imagem salvo em um formato gráfico de alta qualidade. É usualmente utilizado para armazenar imagens com muitas cores, tipicamente fotos digitais, e inclui suporte para camadas e múltiplas páginas.
.tab	Arquivo de texto que contém uma tabela de dados na qual colunas são separadas por abas; pode ser importado pela maioria dos programas de planilha, que formatarão os dados em células.
.dta	Arquivos que contêm a extensão .dta são comumente associados com uma variedade de aplicativos, para uma série de formatos de arquivos de dados. Os arquivos DTA mais comuns são normalmente armazenados em formato binário ou textual. O programa *Turbo Pascal* usa o formato de arquivo DTA para os arquivos de dados referenciados pelo aplicativo. O *software Stata* [programa de estatística] também usa o formato de arquivo DTA para arquivos de conjuntos de dados salvos.
.sav	Arquivo de dados criado pelo *Statistical Package for the Social Sciences*[5] (SPSS), um aplicativo usado para análise estatística; salvo em um formato binário proprietário e contém um conjunto de dados assim como um dicionário que o descreve; salva dados por 'casos' (linhas) e 'variáveis' (colunas).

[5] Pacote Estatístico para as Ciências Sociais.

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

.do	Um arquivo DO é um programa *Java* baseado na *Web* executado por um servidor *Web* que suporta Java, como *Tomcat* ou IBM *WebSphere*. É tipicamente mapeado ao componente *Controller* do *framework Struts*, que processa o arquivo. Arquivos DO são utilizados para gerar páginas *Web* dinâmicas.
.png	Um arquivo PNG é um arquivo de imagem armazenado no formato *Portable Network Graphic* (PNG). Contém um *bitmap* de cores indexadas e usa compactação sem perdas, similar a um arquivo .gif, mas sem limitações de direitos autorais. Arquivos PNG são comumente usados para armazenar gráficos para imagens da *Web*.

Fonte: Sharpened Productions, 2020; Bitberry Software APS, 2020, tradução nossa.

Essa ocorrência chama atenção, pois a extensão .sav foi encontrada apenas no repositório da USIL (multidisciplinar) e detém a maior porcentagem em frequência entre todas as variáveis analisadas nessa etapa. Outra observação necessária é que, assim como a UFPR, a universidade peruana (USIL) está se utilizando do DSpace para gerir não apenas seus conjuntos de dados, mas todos os outros documentos digitais institucionais. Em sequência, o CIAT (disciplinar) foi o repositório observado com a maior variação entre as extensões encontradas, ultrapassando o PPBio nesse quesito. Percebe-se que os arquivos de dados científicos desse repositório são majoritariamente estruturados em formatos alfanuméricos, possuindo 60,81% de seus arquivos salvos em formatos tabulares (.csv, .xls, .xlsx, .xlt, .tab). Os repositórios do CIAT e da PUC Peru fazem parte da comunidade do *Dataverse*.

Devido ao menor peso que obtiveram dentro da metodologia utilizada, poucos conjuntos dos repositórios da PUC e MEC do Peru foram analisados, e eles são em totalidade arquivos de natureza textual e tabelar (.pdf, .csv, .xlsx, .tab). Observa-se em alguns repositórios a inexistência de padronização quanto ao uso dos programas e extensões que se destinam ao processamento e arquivamento de dados de uma mesma natureza, como nos casos da PUC Peru e CIAT, ou nos casos do PPBio e IBICT. Por exemplo, no repositório do CIAT, os depósitos de arquivos de natureza tabelar foram feitos em 5 diversificadas extensões, 3 de programas de código fonte fechado e acesso pago e 2 de programas de código fonte aberto e acesso gratuito.

Faz-se necessário apontar que na constituição de uma política de gestão de dados científicos, deve-se buscar orientação por padrões internacionais como as 5 Estrelas para Dados Abertos, que estabelecem critérios destinados a publicações, onde “dados devem ser publicados em licença aberta e formatos não proprietários”. (FIVESTARDATA, 2019). Assim, as instituições evitariam transtornos legais com o acesso à informação e propriedade intelectual. Com o objetivo de distinguir as extensões encontradas em formatos proprietários e não proprietários, criaram-se os Quadros 3 e 4, respectivamente.

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

Quadro 3 – Extensões de formatos proprietários encontrados nos repositórios analisados

Formato proprietário	**.doc**	**.docx**	**.ppt**	**.pptx**	**.xls**	**.xlsx**	**.xlt**	**.gdb**	**.data**	**.rar**	**.sav**
Tipo de arquivo	Documento *Microsoft Word*	Documento *Microsoft Word Open XML*	Apresentação de *PowerPoint*	Apresentação de *PowerPoint Open XML*	Planilha *Excel*	Planilha *Microsoft Excel Open XML*	*Excel Template*	Arquivo de base de dados *InterBase*	Arquivo de dados *Analysis Studio Offline*	Arquivo comprimido *WinRAR*	Arquivo de dados *SPSS*
Desenvolvedor	*Microsoft*	*Microsoft*	*Microsoft*	*Microsoft*	*Microsoft*	*Microsoft*	*Microsoft*	*Borland*	*Appricon*	Eugene Roshal	IBM
Categoria	Arquivos de texto	Arquivos de texto	Arquivos de dados	Arquivos de dados	Arquivos de planilhas	Arquivos de planilhas	Arquivos de dados	Arquivos de bancos de dados	Arquivos de dados	Arquivos compactados	Arquivos de dados
Formato	Binário	Zip	Binário	Zip	Binário	Zip	Binário	N/A	Texto	Binário	Binário

Fonte: Sharpened Productions, 2020, tradução nossa.

Quadro 4 – Extensões de formatos não proprietários encontrados nos repositórios analisados

Não proprietário	**.txt**	**.pdf**	**.csv**	**.tab**	**.do**	**.gpx**	**.jpg**	**.png**	**.tif**	**.zip**	**.xml**	**.rdf**
Tipo de arquivo	Arquivo de texto simples	Arquivo de formato de documento portátil	Arquivo de dados separados por abas	Arquivo de valores separados por vírgula	*Java ServLet*	Arquivo de troca por GPS	Imagem JPEG	Gráfico de rede portátil	Arquivo de imagem marcada	Arquivo compactado	Arquivo XML	Arquivo de estrutura de descrição de recurso
Desenvolvedor	N/A	*Adobe Systems*	N/A	N/A	N/A	N/A	*Joint Photographic Experts Group*	*PNG Development Group*	N/A	Phil Katz	N/A	N/A
Categoria	Arquivos de texto	Arquivos de *layout* de página	Arquivos de dados	Arquivos de texto	Arquivos *Web*	Arquivos GIS	Arquivos de imagem Raster	Arquivos de imagem Raster	Arquivos de imagem Raster	Arquivos compactados	Arquivos de dados	Arquivos de configurações
Formato	Texto	Binário	Texto	Texto	N/A	XML	Binário	Binário	Binário	Zip	XML	Texto

Fonte: Sharpened Productions, 2020, tradução nossa.

Neles, pôde-se notar uma breve descrição das extensões e a sua devida responsabilidade intelectual por meio da aba ‘desenvolvedor’. Os quadros devem ser observados como exemplos e orientações a pesquisadores e gestores de repositórios no processo de manutenção de acervos digitais sob sua responsabilidade.

4.2 Discussão dos resultados

A começar pelo primeiro objetivo estabelecido para esta investigação, partiu-se do ponto comum entre pesquisas de caráter exploratório, onde se obteve o primeiro contato com o universo a ser investigado. Os pressupostos partiam da ideia de que as abordagens aos assuntos pesquisados poderiam ser incipientes em nível continental. De certa forma,

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

os achados corroboram tal impressão, uma vez que o número de projetos de repositórios de dados científicos se demonstrou baixo quando comparado à quantidade de instituições de pesquisa existentes em toda a América do Sul. No entanto, sob uma ótica otimista, pode-se dizer que os três países em questão (Brasil, Colômbia e Peru) estão à frente dos demais vizinhos nesse quesito.

Em relação aos conjuntos de dados científicos investigados, notou-se que sua natureza se concentra em dados textuais e numéricos, salvos em arquivos de texto e em tabelas, respectivamente. Como observado por Sales e Sayão (2019, p. 41), tabelas têm natureza visual, entretanto, seus dados são resultados de levantamentos, fórmulas, equações, entre outros, ou seja, de natureza numérica.

Por meio das análises dos formatos e extensões dos arquivos de dados, percebeu-se que os conjuntos de dados podem ser tanto homogêneos (um ou mais arquivos salvos em um único formato e extensão, ex.: formato de imagem em .jpg) ou heterogêneos (arquivos salvos em diferentes formatos e extensões, ex.: mesmo formato de imagem salvo em .jpg e .tiff) em sua composição. Apurou-se também que algumas extensões possibilitam a identificação da natureza, do domínio e do conteúdo dos dados, como observado nas extensões .gpx e .gdb, que se referem a dados de localização geográfica, logo, de natureza alfanumérica. Há crescente necessidade de se descrever a natureza dos dados, assim como os formatos ou extensão de seus arquivos. Esse tipo de metadado descritivo seria valioso a potenciais usuários, pois permitiria obter maior compreensão do contexto dos dados com foco em seu reúso.

Os achados corroboram as afirmações de Borgman, Scharnhorst e Golshan (2019, p. 889) em que arquivos de dados digitais não são entidades monolíticas. Alguns coletam apenas dados de certos tipos e formatos, como sequências de genômas para a pesquisa biológica ou dados de *surveys* para as ciências econômicas e sociais. Outros são mais genéricos, coletando documentos textuais, imagens estáticas ou em movimento, áudio e outros tipos de dados (BORGMAN; SCHARNHORST; GOLSHAN, 2019, p. 889, tradução nossa).

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

CONSIDERAÇÕES FINAIS

Os programas por trás dos repositórios investigados que servem à gestão e curadoria de dados científicos são o *Morpho* (*DataONE*), o *DSpace*, e o *Dataverse*. Como percebido e declarado na literatura, cada um serve a um propósito e a uma ou mais comunidades. O *Morpho* se destina à comunidade compartilhada entre as Ciências Biológicas e Ecologia, foco em dados científicos; o *DSpace* (multidisciplinar), à preservação digital institucional, foco em documentos bibliográficos; o *Dataverse* (multidisciplinar), a comunidades compartilhadas entre diversos campos de pesquisa, foco em dados científicos.

Os repositórios encontrados e aptos à investigação foram o PPBio Data Repository (Brasil); a Base de Dados Científicos da Universidade Federal do Paraná (Brasil); o CIAT Dataverse (Colômbia); o Portal de Datos Abiertos de la Pontificia Universidad Católica del Perú (Peru); o Repositorio de datos del Ministerio de Educación del Perú (Peru); e o Repositorio Institucional USIL (Peru). Os demais repositórios eliminados passaram por algum problema antes da fase de análise, o que justifica a lista supracitada.

Uma limitação no desenvolvimento desta investigação esteve no desconhecimento de fontes de informação ao levantamento de repositórios de dados científicos. Não se pode afirmar que os repositórios encontrados à época eram os únicos existentes no continente sulamericano, porém foram os únicos recuperados na superfície do *wiceberg*[6]. Outra dificuldade esteve presente nos processos de coleta e análise dos dados, que idealmente deveriam atingir todo o universo amostral.

Futuras pesquisas poderiam ampliar a investigação levantando dados de outros países e continentes visando a comparação entre as realidades do Brasil e do mundo. Paralelamente, poderia-se investigar *stakeholders* de dados científicos, buscando mapear competências, melhores práticas, desafios e oportunidades, entre outros. Há crescente necessidade de se pensar soluções inovadoras à descrição e indexação com foco na recuperação desses objetos digitais.

[6] Web Iceberg.

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 27, p. 01-26, 2022.
Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2022.e85148

Encontros Bibli

Os resultados encontrados por meio do desenvolvimento dessa investigação podem ser úteis a grupos específicos de leitores, cujos interesses se apoiam na busca da compreensão dos fenômenos aqui abordados, como pesquisadores e professores. Sendo essa uma expectativa comum, posto que se trata de um trabalho científico, espera-se alcance externo às paredes de laboratórios de pesquisa, onde os achados sejam de valia a profissionais imbuídos a enfrentar os desafios emergentes no que concerne à gestão e curadoria de dados científicos.

Há crescente necessidade de se descrever a natureza dos dados, assim como os formatos e extensões de seus arquivos. Esse tipo de metadado descritivo seria valioso a potenciais usuários, pois permitiria obter maior compreensão do contexto dos dados com foco em seu reúso.

REFERÊNCIAS

AMARAL, F. **Introdução à ciência de dados**: mineração de dados e Big Data. Rio de Janeiro: Alta Books, 2016. 320 p.

BITBERRY SOFTWARE APS. **File.org**: dta. [*S. l.*], 2020. Disponível em: https://file.org/extension/dta. Acesso em: 21 fev. 2020.

BORGMAN, C. L. **Big data, little data, no data:** scholarship in the networked world. Cambridge; London: The MIT Press, 2015.

BORGMAN, C. L; SCHARNHORST, A.; GOLSHAN, M. S. Digital data archives as knowledge infrastructures: mediating data sharing and reuse. **Journal of the Association for Information Science and Technology**, [*S. l.*], v. 70, n. 8, 2019. DOI: https://doi.org/10.1002/asi.24172. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/epdf/10.1002/asi.24172. Acesso em: 17 jun. 2020.

DATAONE. **Software tools catalog**. [*S. l.*], [2020]. Disponível em: https://www.dataone.org/software_tools_catalog. Acesso em: 19 fev. 2020.

DATAVERSE. **Dataverse project**: about. [*S. l.*], [2020]. Disponível em: https://dataverse.org/about. Acesso em: 19 fev. 2020.

FIVESTARDATA. **5 Estrelas para dados abertos**. [*S. l.*], 2019. Disponível em: https://5stardata.info/pt-BR/. Acesso em: 16 set. 2019.

HEY, T.; TANSLEY, S.; TOLLE, K. (ed.). **The fourth paradigm**: data-intensive scientific discovery. Redmond, Washington: Microsoft Research, 2009.

HUNT, K. The challenges of integrating data literacy into the curriculum in an undergraduate institution. **IASSIST**, Denmark, v. 28, n. 2-3, p. 12-16, 2004. DOI: https://doi.org/10.29173/iq791. Disponível em: https://iassistquarterly.com/index.php/iassist/article/view/791. Acesso em: 21 ago. 2019.

IBICT. **Sistema para construção de repositórios institucionais digitais (DSpace)**. Rio de Janeiro; Brasília, 2019. Disponível em: http://www.ibict.br/tecnologias-para-informacao/DSpace. Acesso em: 08 out. 2019.

ILHARCO, F. Filosofia da Informação: alguns problemas fundadores. *In*: II Congresso Ibérico de Ciências da Comunicação, 2004, Portugal. **Anais** […]. Portugal, 2004. Disponível em: https://www.cccc2004.ubi.pt. Acesso em: 26 set. 2019.

KOLTAY, T. Data literacy: in search of a name and identity. **Journal of Documentation**, [*S. l.*], v. 71, n. 2, p. 401-415, 2015. DOI: 10.1108/JD-02-2014-0026. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/JD-02-2014-0026/full/pdf?title=data-literacy-in-search-of-a-name-and-identity. Acesso em: 24 ago. 2019.

ROCHA, L. L.; SALES, L. F.; SAYÃO, L. F. Uso de cadernos eletrônicos de laboratório para as práticas de ciência aberta e preservação de dados de pesquisa. **PontodeAcesso**, Salvador, v. 11, n. 3, p. 2-16, dez. 2017. DOI: http://dx.doi.org/10.9771/rpa.v11i3.24945. Disponível em: https://portalseer.ufba.br/index.php/revistaici/article/view/24945/15542. Acesso em: 20 set. 2018.

RODRIGUES, Marcello Mundim; DIAS, Guilherme Ataíde; LOURENÇO, Cíntia de Azevedo. Repositórios de dados científicos na América do Sul: uma análise da conformidade com os Princípios FAIR. **Em Questão**, Porto Alegre, v. 28, n. 2, e-113057, abr./jun. 2022. DOI: http://dx.doi.org/10.19132/1808- 5245282.113057.

SALES, L. F.; SAYÃO, L. F. Uma proposta de taxonomia para dados de pesquisa. **Conhecimento em Ação**, Rio de Janeiro, v.4, n. 1, p. 31-48, 2019. Disponível em: https://revistas.ufrj.br/index.php/rca/article/view/26337. Acesso em: 13 ago. 2020.

SARACEVIC, T. Ciência da Informação: origem, evolução e relações. **Persp. Ci. Inf.**, Belo Horizonte, v. 1, n. 1, p. 41-62, 1996. Disponível em: https://brapci.inf.br/_repositorio/2010/08/pdf_fd9fd572cc_0011621.pdf. Acesso em: 14 out. 2019.

SHARPENED PRODUCTIONS. **Fileinfo**: the files extension database. [*S. l.*], 2020. Disponível em: https://fileinfo.com/. Acesso em: 18 set. 2019.

STOREY, V. C.; SONG, I. Big data technologies and management: what conceptual modelling can do. **Data & Knowledge Engineering**, [*S. l.*], v. 108, p. 50–67, 2017. DOI: https://doi.org/10.1016/j.datak.2017.01.001. Disponível em:

https://www.sciencedirect.com/science/article/abs/pii/S0169023X17300277. Acesso em: 25 jun. 2018.

WAMBA, S. F. *et al*. How 'big data' can make big impact: findings from a systematic review and a longitudinal case study. **Int. J. Production Economics**, [*S. l.*], v. 165, p. 234-246, 2015. DOI: https://doi.org/10.1016/j.ijpe.2014.12.031. Disponível em: https://www.sciencedirect.com/science/article/pii/S0925527314004253. Acesso em: 25 jun. 2018.

ZINS, C. Conceptual approaches for defining data, information, and knowledge. **Journal of the American Society for Information Science and Technology**, [*S. l.*], v. 58, n. 4, p. 479-493, 2007. DOI: https://doi.org/10.1002/asi.20508. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20508. Acesso em: 25 jun. 2018.

NOTAS

CONTRIBUIÇÃO DE AUTORIA

Concepção e elaboração do manuscrito: M. M. Rodrigues.
Coleta de dados: M. M. Rodrigues.
Análise de dados: M. M. Rodrigues.
Discussão dos resultados: M. M. Rodrigues.
Revisão e aprovação: G. A. Dias, C. A. Lourenço.

Caso necessário veja outros papéis em: https://casrai.org/credit/

CONJUNTO DE DADOS DE PESQUISA

Escolha uma das opções e apague as demais.

1) Todo o conjunto de dados que dá suporte aos resultados deste estudo foi publicado no artigo e na seção "Materiais suplementares".

PUBLISHER
Universidade Federal de Santa Catarina. Programa de Pós-graduação em Ciência da Informação. Publicação no Portal de Periódicos UFSC. As ideias expressadas neste artigo são de responsabilidade de seus autores, não representando, necessariamente, a opinião dos editores ou da universidade.

EDITORES
Edgar Bisset Alvarez, Ana Clara Cândido, Patrícia Neubert e Genilson Geraldo.

HISTÓRICO
Recebido em: 12-08-2021 – Aprovado em: 04-05-2022 - Publicado em: 25-05-2022.