**ALEXANDRE RIBAS SEMELER**

# CIÊNCIA DA INFORMAÇÃO EM CONTEXTOS DE *E-SCIENCE*: BIBLIOTECÁRIOS DE DADOS EM TEMPOS DE *DATA SCIENCE*

Tese submetida ao Programa de Pós-Graduação em Ciência da Informação do Centro de Ciências da Educação da Universidade Federal de Santa Catarina, para a obtenção do Grau de Doutor em Ciência da Informação
Orientador: Prof. Dr. Adilson Luiz Pinto

Florianópolis
2017

Ficha de identificação da obra elaborada pelo autor
através do Programa de Geração Automática da Biblioteca Universitária
da UFSC.

Semeler, Alexandre Ribas
CIÊNCIA DA INFORMAÇÃO EM CONTEXTOS DE E-SCIENCE : bibliotecários de dados em tempos de Data Science / Alexandre Ribas Semeler ; orientador, Adilson Luiz Pinto , 2017.
175 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro de Ciências da Educação, Programa de Pós-Graduação em Ciência da Informação, Florianópolis, 2017.

Inclui referências.

1. Ciência da Informação. 2. Biblioteconomia de Dados. 3. Bibliotecário de Dados. 4. Data Science. 5. e-Science. I. , Adilson Luiz Pinto. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Informação. III. Título.

**ALEXANDRE RIBAS SEMELER**

**CIÊNCIA DA INFORMAÇÃO EM CONTEXTOS DE *E-SCIENCE*: BIBLIOTECÁRIOS DE DADOS EM TEMPOS DE *DATA SCIENCE***

Tese de doutorado apresentada ao Programa de Pós-Graduação em Ciência da Informação do Centro de Ciências da Educação da Universidade Federal de Santa Catarina em cumprimento a requisito parcial para a obtenção do título de Doutor em Ciência da Informação.

APROVADA PELA COMISSÃO EXAMINADORA

EM FLORIANÓPOLIS 15 DE DEZEMBRO DE 2017

Prof. Adilson Luiz Pinto, Dr. - Coordenador do Curso

Prof. Adilson Luiz Pinto, Dr.
Universidade Federal de Santa Catarina/Programa de Pós-Graduação em Ciência da Informação - Orientador

Prof. Angel Freddy Godoy Viera, Dr.
Universidade Federal de Santa Catarina/Programa de Pós-Graduação em Ciência da Informação

Profa. Elaine Rosângela de Oliveira Lucas, Dra.
Universidade Estadual do Estado de Santa Catarina/ Programa de Pós-Graduação em Gestão da Informação

Prof. Thiago Magela Rodrigues Dias, Dr.
Centro Federal de Educação Tecnológica de Minas Gerais/Departamento de Informática, Gestão e Design

**AGRADECIMENTOS**

Agradeço a meu orientador Prof. Dr Adilson Luiz Pinto pela amizade, ensinamentos e apoio incondicional em todo desenvolver do curso de doutorado.

A minha esposa, Janaína, pelo amor, compreensão e paciência ao longo desse percurso.

A Nima pela companhia e momentos de reflexão.

A todos os professores, funcionários e alunos do PPGCIN/UFSC. E a todos que de alguma forma contribuíram.

# RESUMO

A tese de doutorado "**Ciência da Informação em contextos de e-*Science*: bibliotecários de dados em tempos de *Data Science***" tem como objetivo geral investigar como as relações entre Ciência da Informação, *e-Science* e *Data Science* influenciam a Biblioteconomia de Dados. O objeto de estudo da Ciência da Informação é a informação e as transformações impulsionadas pelas tecnologias na sociedade contemporânea. A *e-Science* é entendida como a base para o pensamento das ciências orientadas ao uso intensivo de dados. A *Data Science* é o conjunto de habilidades, métodos, técnicas e tecnologias da Estatística e da Ciência da Computação utilizadas para extrair conhecimento de dados e para a criação de produtos e serviços a partir de dados. A Biblioteconomia de Dados é uma disciplina que aplica Ciência da Informação, *e-Science* e *Data Science* em Bibliotecas. Acredita-se que os repositórios digitais de dados de pesquisa são a parte da ciberinfraestrutura de *e-Science* que deve ser investigada pelos bibliotecários de dados. Nesse contexto, a tese configura-se em quatro seções. No capítulo 1, apresentam-se uma introdução ao tema, o contexto do estudo, o problema de pesquisa, os objetivos e a metodologia. No capítulo 2, contempla-se a fundamentação teórica. O objetivo desse capítulo é refletir a respeito de questões básicas, como: o que é informação? O que é tecnologia? O que são dados de pesquisa? O que é Biblioteconomia de Dados? O que é a *Data Science*? Quais as habilidades necessárias aos cientistas de dados que podem ser aproveitadas pelos bibliotecários de dados? O método apresenta-se nas seguintes etapas: coleta de dados (*web scraping*), manipulação e análise descritiva da frequência dos dados (limpeza, organização, tratamento de desambiguação e formato de dados) e visualização (criação de gráficos e grafos). No capítulo 3, apresenta-se o resultado do método que representa o cenário internacional dos Repositórios de Dados de Pesquisa em Geociências disponíveis no Registry of Research Data Repositories (re3data). Esses repositórios são fontes de dados heterogêneas que possibilitam o acesso e a preservação a uma ampla gama de tipos de dados de pesquisa em Geociências. Disponíveis em países da América do Norte e da Europa. Em setembro de 2017 haviam 604 RDPGs registrados no re3data, dos quais é possível sumarizar as seguintes características: (a) tipos de dados: documentos de escritório, dados científicos e estatísticos, imagens e dados brutos; (b) os países: USA, DEU, CAN, GBR, AUS e FRA e a língua preponderante nos dados de pesquisa é o inglês; (c) áreas do conhecimento: Ciências atmosféricas e Oceanografia, sendo as principais temáticas: o clima, as mudanças climáticas, a geologia, a hidrologia, o sensoriamento remoto e a oceanografia; (d) em sua maioria, são de acesso de aberto e utilizam licenças Creative Commons; (e) os *softwares* para desenvolvimento desses repositórios são desconhecidos, 37 utilizam *softwares* como *Ckan, DSpace, DataVerse, e-Prints* e *Fedora*. Conclui-se com uma discussão a respeito do Diagrama de Venn da Biblioteconomia de Dados, o qual representa as bases Ciência da Informação, *e-Science* e *Data Science* como fundamento para Biblioteconomia de dados.

**Palavras-chave:** Ciência da Informação. *e-Science. Data Science.* Biblioteconomia de Dados. Bibliotecário de Dados. Dados de Pesquisa. Repositório de dados de pesquisa.

# ABSTRACT

A doctoral thesis "**Information Science in e-Science Contexts: data librarians in the age of Data Science**" aims to investigate how the relationships between Information Science, e-Science and Data Science influences Data Librarianship. The object of study of the Information Science is the information and the changes by technology in contemporary society. e-Science is understood as the grounding for the thinking about the data-intensive sciences. Data Science is the set of the skills, methods, techniques and technologies of the Statistics and of the Computer Science used to extract knowledge of data and to create products and services from data. Data Librarianship is a discipline that applies Information Science, e-Science and Data Science in Libraries. It is believed that digital repositories of research data are part of the cyberinfrastructure of e-Science that should be investigated by data librarians. In this context, the thesis presents in four sections. The chapter 1, presents an introduction to the subject of the thesis, the context of the study, the research problem, objectives and methodology. The chapter 2 considers the theoretical foundation. The aim of this chapter is to reflect on basic questions such as: what is information? What is technology? What is search data? What is Data Librarianship? What is Data Science? What are the skills needed to data scientists who can be taken advantage of by data librarians? The method is presented in the following steps. Data collection (web scraping), data handling and descriptive analysis of the data frequency (cleaning, organization, processing disambiguation and data format) and visualization (creating charts and graphs). In Chapter 3, we present the result of the method that represents the international scenario of the Research Data Repositories in Geosciences available in the Registry of Research Data Repositories (re3data). These repositories are heterogeneous data sources that enable access and preservation to a wide range of types of research data in Geosciences. In September 2017 there were 604 RDPGs recorded in re3data, and we can summarize the characteristics: (a) data types: office documents, scientific and statistical data, images and raw data; (b) the countries: USA, DEU, CAN, GBR, AUS and FRA and the preponderant language in the research data is the English; (C) knowledge areas: Atmospheric Sciences and Oceanography, the main subjects: climate, climate change, geology, hydrology, remote sensing and oceanography; (d) for the most part, they are open access and use Creative Commons licenses; (e) the software for the development of these repositories are unknown, 37 use software such as Ckan, DSpace, DataVerse, e-Prints and Fedora. It concludes presents the Venn Diagram of Data Librarianship which represents the bases (Information Science, e-Science and Data Science) as fundamentals of Data Librarianship.

**Keywords:** Information Science. e-Science. Data Science. Data Librarianship. Data Librarian. Research Data. Research Data Repository

# RESUMEN

La tesis de doctorado "**Ciencia de la Información en contextos de e-Science: bibliotecarios de datos en tiempos de Data Science**" tiene como objetivo general investigar cómo las relaciones entre Ciencia de la Información, e-Science y Data Science influencian la Biblioteconomía de Datos. El objeto de estudio de la Ciencia de la Información es la información y las transformaciones impulsadas por las tecnologías en la sociedad contemporánea. La e-Science es entendida como la base para el pensamiento de las ciencias orientadas al uso intensivo de datos. La Data Science es el conjunto de habilidades, métodos, técnicas y tecnologías de la estadística y de la ciencia la computación utilizadas para extraer conocimiento de datos y para la creación de productos y servicios a partir de datos. La Biblioteconomía de datos es una disciplina que aplica Ciencia de la Información, e-Science y Data Science en Bibliotecas. Se cree que los repositorios digitales de datos de investigación son la parte de la ciberinfraestructura de e-Science que debe ser investigada por los bibliotecarios de datos. En este contexto, la tesis se presenta en cuatro secciones. En el capítulo 1 se presenta una introducción al tema, el contexto del estudio, el problema de investigación, los objetivos y la metodología. En el capítulo 2 se contempla la fundamentación teórica. El objetivo de este capítulo es reflexionar acerca de cuestiones básicas, como: ¿qué es información? ¿Qué es la tecnología? ¿Qué son los datos de investigación? ¿Qué es Biblioteconomía de Datos? ¿Qué es la Data Science? ¿Cuáles son las habilidades necesarias para los científicos de datos que pueden ser aprovechados por los bibliotecarios de datos? El método se presenta en las siguientes etapas: colección de datos (web scraping), manipulación y análisis descriptivo de la frecuencia de los datos (limpieza, organización, tratamiento de desambiguación y formato de datos) y visualización (creación de gráficos y grafos). En el capítulo 3, se presenta el resultado del método que representa el escenario internacional de los repositorios de datos de investigación en las Geociencias disponibles en el Registry of Research Data Repositories (re3data). Estos repositorios son fuentes de datos heterogéneas que posibilitan el acceso y la preservación a una amplia gama de tipos de datos de investigación en Geociencias. Disponibles en países de América del Norte y de Europa. En septiembre de 2017 había 604 RDPG registrados en el re3data, de los cuales es posible resumir las siguientes características: (a) tipos de datos: documentos, datos científicos y estadísticos, imágenes y datos brutos; (b) los países: USA, DEU, CAN, GBR, AUS y FRA y la lengua preponderante en los datos de investigación es el inglés; (c) áreas del conocimiento: Ciencias atmosféricas y Oceanografía, siendo las principales temáticas: el clima, los cambios climáticos, la geología, la hidrología, el sensoriamiento remoto y la oceanografía; (d) en su mayoría, son de acceso de abierto y utilizan licencias Creative Commons; (e) los softwares para el desarrollo de estos repositorios son desconocidos, 37 utilizan software como Ckan, DSpace, DataVerse, e-Prints y Fedora. Por fin, se presenta el Diagrama de Venn de la Biblioteconomía de datos que representa el trípode (Ciencia de la Información, e-Science y Data Science) como fundamentos de la Biblioteconomía de datos.

**Palavras clave:** Ciencia de la Información. e-Science. Data Science. Biblioteconomía de datos. Bibliotecarios de datos. Datos de Investigación. Repositorios de Datos de Investigación.

## LISTA DE FIGURAS

Figura 1 - Fluxo metodológico. 26
Figura 2 - Conjunto de dados coletados do re3data. 28
Figura 3 - Processo de criação de conjuntos de dados para visualização. 30
Figura 4 - Elementos do trabalho acadêmico. 73
Figura 5 - *Harvesting* de metadados. 99
Figura 6 - Fluxo *Tidyverse*. 104
Figura 7 - Distribuição geográfica dos RDPGs. 113
Figura 8 - Temas mais frequentes em Ciências Naturais. 120
Figura 9 - Temas mais frequentes em Ciências da Vida. 121
Figura 10 - Temas mais frequentes em Humanidades. 122
Figura 11 - Temas mais frequentes em Engenharias 123
Figura 12 - Diagrama de Venn da Biblioteconomia de Dados. 133

# LISTA DE QUADROS

Quadro 1 - Caracterização do método. ..... 25
Quadro 2 - Tipos de *data scientists*. ..... 75
Quadro 3 - Habilidades requeridas aos bibliotecários de dados. ..... 78
Quadro 4 - Habilidades requeridas aos bibliotecários de dados. ..... 84
Quadro 5 - Tipos de análises de dados. ..... 88
Quadro 6 - *Softwares* para elaboração de repositórios de dados. ..... 92
Quadro 7 - Softwares para *web scraping*. ..... 97
Quadro 8 - *Softwares* para *Data handling*. ..... 100

# LISTA DE TABELAS

Tabela 1 - Países e instituições que possuem RDPGs. ..... 109
Tabela 2 - Distribuição de áreas do conhecimento por palavras - chave. ..... 118
Tabela 3 - Relação de RDPGs: instituição, tipo de dados, países, acesso e licenças. ..... 125

# LISTA DE GRÁFICOS

Gráfico 1 – Países ..... 112
Gráfico 2 – Tipologia dos repositórios ..... 114
Gráfico 3 – Tipos de dados de pesquisa ..... 116
Gráfico 4 – Tipo de licença ..... 117
Gráfico 5 – Tipologia dos *softwares* de repositórios ..... 124

## SUMÁRIO

**1 CIÊNCIA DA INFORMAÇÃO EM CONTEXTOS DE E-*SCIENCE*: BIBLIOTECÁRIO DE DADOS EM TEMPOS DE *DATA SCIENCE* 15**

1.1 REPOSITÓRIOS DE DADOS DE PESQUISA ..... 19

1.2 POR QUE E COMO OS BIBLIOTECÁRIOS DE DADOS DEVEM SE PREOCUPAR COM REPOSITÓRIOS DE DADOS DE PESQUISA? ..... 22

1.3 OBJETIVOS ..... 24

1.4 MÉTODO ..... 24

**1.4.1 Coleta de dados ..... 27**

**1.4.2 Manipulação e análise de dados ..... 29**

**1.4.3 Visualização ..... 30**

**2 FUNDAMENTAÇÃO TEÓRICA ..... 33**

2.1 INFORMAÇÃO E TECNOLOGIA, ASPECTOS GERAIS EM CIÊNCIA DA INFORMAÇÃO ..... 33

**2.1.1 Ciência da Informação ..... 33**

**2.1.2 O conceito de informação ..... 40**

**2.1.3 O conceito de tecnologia ..... 42**

2.2 E-SCIENCE ..... 45

**2.2.1 Ciência orientada aos dados*:* o quarto paradigma ..... 50**

**2.2.2 Ciência da Informação em contextos de *e-Science*: imbricações ..... 53**

2.3 O CONCEITO DE DADOS ..... 56

**2.3.1 Dados de pesquisa ..... 57**

**2.3.2 Dados de pesquisa e as diferentes disciplinas ..... 61**

**2.3.3 Biblioteconomia de Dados ..... 65**

*2.4 DATA SCIENCE* ..... 68

**2.4.1 Habilidades do cientista de dados ..... 71**

**2.4.2 Habilidades do bibliotecário de dados ..... 75**

**2.4.3 Bibliotecários de dados e os dados de pesquisa ..... 79**

2.5 TECNOLOGIAS E INSTRUMENTOS PARA MANIPULAÇÃO DE DADOS 83

**2.5.1 *Softwares* e algoritmos para análise de dados ..... 83**

**2.5.2 *Software* para elaboração de repositórios de dados ..... 90**

**2.5.3 *Software* para recuperação de informação ..... 95**

**2.5.4 *Harvesting* de metadados ..... 98**

**2.5.5 *Software* para manipulação de dados (*Data Handling*)..... 100**

**2.5.6 Linguagens de programação para manipulação e análise de dados 103**

2.5.6.1 *R-project*..... 103

Fonte: GROLEMUND; WICKHAM (2017). ..... *104*

2.5.6.2 *Python* ..... 105

**3 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS..... 108**

3.1 CENÁRIO INTERNACIONAL DOS REPOSITÓRIOS DE DADOS DE PESQUISA EM GEOCIÊNCIAS (RDPGs) ..... 108

4 **CONSIDERAÇÕES ..... 132**

**REFERÊNCIAS ..... 144**

# 1 CIÊNCIA DA INFORMAÇÃO EM CONTEXTOS DE E-*SCIENCE*: BIBLIOTECÁRIO DE DADOS EM TEMPOS DE *DATA SCIENCE*

A tese de doutorado "**Ciência da Informação em contextos de e-*Science*: bibliotecários de dados em tempos de *Data Science***" investiga como as relações entre Ciência da Informação, *e-Science* e *Data Science* influenciam a Biblioteconomia de Dados.

*A priori*, acredita-se que a Ciência da Informação (CI) é um campo do conhecimento que envolve uma série de disciplinas, como Ciência da Computação, Biblioteconomia, entre outras. O objeto de estudo da CI é a informação e as transformações impulsionadas pelo consumo e pelo uso de tecnologias. A CI cada vez mais se preocupa com a natureza digital da informação. Os dados em formato digital são um novo tipo de informação gerada pelas atividades do homem e por meio da mediação tecnológica.

A CI em contextos de *e-Science* pode ser entendida como uma corrente da CI que se preocupa com as ciências orientadas ao uso intensivo de dados. O termo *e-Science* foi utilizado para se referir ao uso de tecnologia computacional durante a prática de investigação científica, o que inclui a preparação, a experimentação, a coleta de dados, a divulgação de resultados, o armazenamento, a preservação e a acessibilidade de todos os materiais utilizados e gerados durante a pesquisa científica. Pode-se afirmar que a *e-Science* se preocupa com os processos de representação da informação e do conhecimento científico por meio da utilização de tecnologias que proporcionam o uso intensivo de dados (HEY; HEY, 2006).

Os métodos da *e-Science* primeiramente foram utilizados por pesquisadores das áreas da Biologia, Física e Astronomia em projetos como o *GenBank*[1], o *Large Hadron Collider* (LHC)[2], o *Panoramic Survey Telescope and Rapid Response System* (Pan-STARRS)[3], entre outros projetos. A própria *world wide web*, inventada por Tim Berners-Lee no European Organization for Nuclear Research (CERN), foi criada para facilitar o compartilhamento de documentos e dados entre

---

[1] GENBANK. Disponível em: <https://www.ncbi.nlm.nih.gov/genbank/>. Acesso em: set. 2017.
[2] LHC. Disponível em: <https://home.cern/topics/large-hadron-collider>. Acesso em: set. 2017.
[3] PAN-STARRS. Disponível em: <https://panstarrs.stsci.edu/>. Acesso em: set. 2017.

pesquisadores. Essas pesquisas de *e-Science* surgem no final da década de 1990. Institucionalmente, a *e-Science* nasce em departamentos nacionais de pesquisa na Inglaterra e nos Estados Unidos da América. Desde então, as duas potências incentivam pesquisas focadas no desenvolvimento da ciência em escala global por meio da colaboração científica e do intercâmbio de dados e tecnologias. (HEY; HEY, 2006; JANKOWSKI, 2007; HEY; TANSLEY; TOLLE, 2009).

A *e-Science* pode ser considerada o quarto paradigma[4] das ciências, dando origem às ciências orientadas para o uso intensivo de dados. Conforme conceito cunhado por Jim Gray[5] em 2007, a *e-Science* pode ser classificada como o quarto paradigma e se configura como uma nova forma de se fazer ciência. Ela é o ponto de encontro entre a tecnologia e os cientistas. O quarto paradigma é um campo do conhecimento que se preocupa em traduzir os métodos científicos para os computadores. Como um paradigma incluso pela tecnologia, a *e-Science* mostra novas interfaces para qualquer domínio científico, o computador, a rede (Internet) e a dimensão dos dados.

A dimensão dos dados aponta uma forte tendência de estudos e práticas que vem sendo adotada pelos cientistas da informação (**ver capítulo 2.2.2**). A maioria das investigações sobre *e-Science* está orientada à prática, ao uso e ao consumo de dados. Assim, pode-se afirmar que as atuais tendências de pesquisa em CI nos contextos de *e-Science* apontam para temáticas relacionadas à coleta, à curadoria, ao processamento, ao armazenamento, à análise e à visualização de dados[6].

---

[4] Os outros três paradigmas são, respectivamente: experimental empírico, teórico e simulação.

[5]James Nicholas Gray, vencedor do Prêmio Turing de 1998, o Nobel da ciência da computação, e um dos pioneiros em aplicações de técnicas computacionais para o tratamento de grandes quantidades de dados gerados por cientistas de outras áreas, via nessa mudança uma verdadeira transformação no fazer ciência. Em sua opinião, estaríamos vivendo o início de um quarto paradigma, que redefiniria a metodologia científica de diversas áreas do conhecimento (CORDEIRO et al., 2013, p.75).

[6] O objetivo da visualização é comunicar dados de forma eficaz e precisa. A visualização funciona como uma janela por onde cientistas podem investigar, avaliar e explorar dados (BONNEAU et. al., 2014). Na Computação e na Ciência da Informação, a visualização é um domínio da representação visual que utiliza gráficos, imagens, sequências animadas, bem como a argumentação, para apresentar tipos e estruturas de dados de comportamento dinâmico, complexo e volumoso, os quais representam sistemas, eventos, processos, objetos e coisas. Embora os computadores não visualizem, eles suportam e proporcionam a simulação da visualização humana. Segundo Sochats e Morse (1997), a visualização é a transformação da informação em exibição visual que eleva a compreensão de usuários de informações. A visualização toma a forma de esboços, desenhos, gráficos e imagens que retratam fenômenos naturais, como o clima, estruturas teóricas de objetos, partículas atômicas, fluxos de diferentes formas de matéria e misturas, modelos de relacionamento em conjuntos de dados, informação volumétrica como a encontrada em imagens médicas. Também a visualização implica processos de interpretação e análise visual de grandes quantidades de dados, o que inclui os

O empoderamento das tecnologias de manipulação de dados pela ciência é uma variável de transformação comum a quase todas as teorias e práticas científicas. No que tange à CI, as pesquisas sobre *e-Science* podem se manifestar na forma da Biblioteconomia de Dados (**ver capítulo 2.3.3)**.

Os estudos sobre a Biblioteconomia de Dados têm origem na metade dos anos 2000. São o resultado do empenho de bibliotecários norte-americanos, britânicos e canadenses em criar serviços e produtos de divulgação, de consultoria, de gerenciamento, preservação e elaboração de esquemas de metadados para efetiva incorporação de dados de pesquisa em coleções de documentos (livros, artigos, relatórios, entre outros). Embora a Biblioteconomia de Dados tenha suas origens na criação de serviços de dados (Data Archives Services) no Reino Unido e (Data Library Services) nos Estados Unidos e no Canadá na década de 1960, o foco da Biblioteconomia de Dados hoje está mais relacionado à criação de produtos e/ou serviços relacionados ao uso e ao consumo de dados de pesquisa (RICE; SOUTHALL, 2016; KOLTAY, 2017; TENOPIR et al., 2017).

Nesse sentido, dados de pesquisa científica são o resultado de qualquer investigação sistemática que envolva processos de observação, experimentação ou simulação de procedimentos de pesquisa científica. Os dados de pesquisa dependem do assunto ou da disciplina científica e podem diferir segundo suas tipologias. Os dados de pesquisa precisam ser identificáveis, citáveis, visíveis, recuperáveis, interpretáveis, contextualizáveis, interoperáveis e reutilizáveis onde quesitos de consistência e procedência são considerados. Dados de pesquisa são objetos digitais e não digitais, como: documentos, questionários, avaliações, registros de casos, protocolos de estudo, planilhas, notas de laboratório, notas de campo, diários, filmes, imagens, arquivos digitais de áudio e vídeo, sequências genéticas, coordenadas geográficas, banco de dados, algoritmos, metodologias, protocolos, entre outras formas de manifestações de pesquisa (ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, 2004; XIA; WANG, 2014; SAYÃO; SALES, 2015; KOLTAY, 2017; DUDZIAK, 2016; INTERNATIONAL ASSOCIATION FOR SOCIAL SCIENCE INFORMATION SERVICES AND TECHNOLOGY, 2016; HENDERSON, 2017).

---

negócios, as finanças, a administração, a mídia digital, os *softwares*, a medicina, a geociências, a biologia etc. (SOCHATS; MORSE 1997).

Os dados de pesquisa gerados pelas mais diversas disciplinas estão em ascensão no que tange aos estudos sobre a Biblioteconomia de Dados. O bibliotecário, tradicionalmente, desenvolve suas atividades com base no armazenamento, na circulação, na preservação e na gestão de acervos documentais. Mas, no contexto da Biblioteconomia de Dados, acredita-se que ele possa aplicar suas habilidades para oferecer produtos e/ou serviços de dados de pesquisa, como no contexto relacionado aos repositórios digitais (**ver capítulo 1.2**). Os repositórios digitais de dados de pesquisa são parte da ciberinfraestrutura de *e-Science* que deve ser investigada pelos bibliotecários de dados.

O bibliotecário de dados deve investigar os sistemas e as normas que envolvem a pesquisa com dados de pesquisa em repositórios digitais, como um profissional que deve buscar conhecimentos sobre coleta, manipulação, análise e visualização de dados de pesquisa para oferecer produtos e/ou serviços ligados a gestão e curadoria de dados de pesquisa. O bibliotecário de dados deve possuir as habilidades (**ver capítulo 2.4.2**) necessárias para trabalhar com qualquer tipo de dados sejam eles observacionais, computacionais ou experimentais. Ele deve compreender como as investigações sobre tecnologia modificam as práticas e as teorias que fundamentam a Biblioteconomia.

A Biblioteconomia de Dados é uma disciplina que aplica CI, *e-Science* e *Data Science* em bibliotecas. Assim, com a finalidade de compreender como a Biblioteconomia de Dados pode contribuir para o desenvolvimento das atividades de *e-Science,* sustenta-se, nesta tese, que as bibliotecas devam se preocupar em capacitar os bibliotecários para o trabalho com as tecnologias de dados (**ver capítulo 2.5**). Esses profissionais precisam se preocupar com dados de pesquisa gerados antes, durante e depois da realização da investigação científica. Para dar conta desse problema, os bibliotecários de dados interessados em *e-Science* devem buscar fundamentação teórica e prática em *Data Science*.

Atualmente, a *Data Science* (**ver capítulo 2.4**) é vista como a teoria e a prática de extrair conhecimento de dados e se preocupa com a criação de produtos e/ou serviços a partir de dados. A *Data Science* configura-se como um campo do conhecimento que mescla técnicas da Ciência da Computação e da Estatística. Ela é uma área do conhecimento que demanda saberes específicos, geralmente relacionados ao uso de linguagens de programação e metodologias científicas para coletar, analisar e visualizar dados (BULINGAME; NIELSEN, 2012).

Acredita-se que a *Data Science* possa ser considerada como uma das principais áreas de estudos para os bibliotecários de dados

interessados no contexto dos repositórios de dados de pesquisa. Principalmente no que tange às habilidades técnicas requisitadas para o trabalho com essas plataformas.

Nesse sentido, propõem-se, nesta tese, que o bibliotecário de dados possa assimilar algumas das competências requeridas aos cientistas de dados **(ver capítulo 2.4.1)**. Os bibliotecários de dados possuem papel fundamental para o desenvolvimento da *Data Science* em bibliotecas, mas a eles cabe aplicar a ciência de uso intensivo de dados mais como uma ciência do que como apenas uma extensão de técnicas de manipulação e análise de dados (ZHU; XIONG, 2015; CHEN, 2015).

Em suma, os temas de pesquisa desta tese envolvem a *e-Science* como o novo paradigma das ciências orientadas aos dados, o que inclui habilidades, métodos, técnicas e tecnologias de *Data Science* como uma fonte para novas práticas em Biblioteconomia de Dados. O bibliotecário de dados é aquele que deve aplicar a CI, a *e-Science* e a *Data Science* em bibliotecas; esse profissional precisa resolver os problemas relacionados a coleta, manipulação, análise e visualização de dados de pesquisa em repositórios digitais.

No que segue, apresentam-se os repositórios de dados de pesquisa (**ver capítulo 2.5.2**) como contexto de estudo para os bibliotecários de dados investigarem a ciberinfraestrutura de *e-Science.*

## 1.1 REPOSITÓRIOS DE DADOS DE PESQUISA

Os diversos tipos de sistemas tecnológicos criados para dar suporte aos dados de pesquisa compõem um rico universo semântico-relacional de informação e de conhecimento que registram o saber resultante do processo de investigação científica. A diversidade de maneiras com que os dados de pesquisa fluem nesses sistemas, principalmente na *web*, acentua a necessidade de organizar, compreender, preservar e analisar a informação e o conhecimento que podem ser extraídos desse meio digital.

Um repositório de dados de pesquisa é um sistema de informação técnico e organizacional que serve para auxiliar pesquisadores na gestão e no armazenamento de dados e também deve facilitar a busca e o acesso a dados de pesquisa em uma ou diversas fontes internas e/ou externas ao repositório. Os repositórios de dados de pesquisa são parte essencial da ciberinfraestrutura de pesquisa e visam à preservação, ao acesso em longo prazo e à reutilização de dados de pesquisa (KINDLING et.al., 2017).

> Os repositórios de dados de pesquisa são uma evolução dos repositórios documentais [...] os repositórios de dados [de pesquisa] são ambientes implementados em universidades com infraestrutura apropriada para dar suporte aos pesquisadores no gerenciamento e na disponibilização [de dados de pesquisa] e, dessa forma, contribuir para que outros pesquisadores possam reutilizá-los. [...] os repositórios de dados [de pesquisa] auxiliam no gerenciamento de dados e é fundamental que sejam planejados e estruturados desde o início de sua implementação (MONTEIRO, 2017, p. 19-34).

Para Pampel et el. (2013), os repositórios de dados de pesquisa podem ser institucionais, disciplinares, multidisciplinares e/ou resultantes de projetos de pesquisa específicos. Comumente, todos os tipos de repositórios de dados de pesquisa enfatizam o acesso e a preservação aos dados de pesquisa.

Os repositórios de dados de pesquisa institucionais são administrados por uma instituição, como uma universidade ou instituição de pesquisa. No nível universitário, seu escopo é multidisciplinar. Um exemplo desse tipo é o Edinburgh DataShare[7]; com sede no Reino Unido, ele disponibiliza conjuntos de dados produzidos pela University of Edinburgh (PAMPEL et el., 2013; SAYÃO; SALES, 2016).

Outro tipo de repositório de dados de pesquisa apresentado por Pampel et el. (2013) é o disciplinar. Como exemplo, cita-se o PANGAEA[8] – Data Publisher for Earth & Environmental Science, o qual se define como uma biblioteca de acesso aberto destinada a arquivar, publicar e distribuir dados georreferenciados sobre o sistema terrestre. É desenvolvido e mantido pela Alfred Wegener Institute for Polar and Marine Research (AWI) e MARUM – Center for Marine Environmental Sciences da University of Bremen, na Alemanha. Esses repositórios são voltados para o arquivamento de domínios específicos de pesquisa. Conforme Sayão e Sales (2016, p. 102), "[...] os repositórios disciplinares ou temáticos [...] são extremamente variados e heterogêneos, refletindo a multiplicidade de disciplinas e a diversidades

[7] DATASHARE. Disponível em: <http://datashare.is.ed.ac.uk>. Acesso em: ago. 2017.
[8] PANGAEA. Disponível em: <http://www.pangaea.de>. Acesso em: ago. 2017.

de dados gerados no contexto da pesquisa científica mundial".

Paralelamente às abordagens disciplinares e institucionais, existem os repositórios de dados de pesquisa multidisciplinares. Sayão e Sales (2016, p. 103) explicam que esses repositórios são aqueles "[...] cujas políticas correspondentes aceitam submissões de coleções de dados de várias áreas do conhecimento e que sejam provenientes de diferentes instituições de pesquisa". Um exemplo citado por Pampel et al. (2013) é o FigShare[9], o qual permite que pesquisadores publiquem todos os seus dados de forma citável, pesquisável e compartilhável e é desenvolvido pela Digital Science e a Macmillan Publishers Company, empresa internacional com sede nos Estados Unidos e no Reino Unido.

Por fim, há os repositórios de dados de pesquisa com foco em dados resultantes de projetos específicos, como o Scientific Drilling Database (SDDB)[10], desenvolvido pelo Scientific Continental Drilling Program (ICDP), disponibilizando conjuntos de dados sobre Geociências em acesso aberto (PAMPEL et al., 2013; SAYÃO; SALES, 2016).

A distribuição global dos repositórios de dados de pesquisa é catalogada pelo Registry of Research Data Repositories (re3data), que é o registro internacional de repositórios de dados de pesquisa. O re3data abrange diferentes disciplinas acadêmicas e registra esse tipo de repositório desde 2012. Ele apresenta uma relação de repositórios que serve para o armazenamento permanente e o acesso a conjuntos de dados de pesquisa em que órgãos de financiamento, editores e instituições acadêmicas promovem uma cultura de compartilhamento, acesso e visibilidade aos dados de pesquisa (RE3DATA, 2012)[11].

Fundado pelo German Research Foundation (DFG), o re3data congrega iniciativas da Library and Information Services (LIS) of the GFZ German Research Centre for Geosciences, da Library of the Karlsruhe Institute of Technology (KIT), da School of Library and Information Science (BSLIS) na Humboldt-Universität, em Berlin e das Libraries of the Purdue University, na Alemanha. O re3data.org indexa, internacionalmente, cerca de 1900 repositórios de dados de pesquisa desde o outono de 2012. Ele oferece a pesquisadores, organizações de financiamento, bibliotecas e editores uma visão geral sistemática da paisagem heterogênea dos repositórios de dados de pesquisa (PAMPEL et al., 2013; COSTA; BRAGA, 2016; KINDLING et. al. 2017).

---

[9] FIGSHARE. Disponível em: <http:// figshare.com>. Acesso em: ago. 2017.
[10] SDDB. Disponível em: <http://www.scientificdrilling.org>. Acesso em: ago. 2017.
[11] RE3DATA. Disponível em: <http://www.re3data.org/about>. Acesso em: ago. 2017.

Os repositórios de dados de pesquisa possibilitam o acesso a uma ampla gama de tipos de dados de pesquisa e sua preservação. Com base no re3data, Costa e Braga (2016, p. 81) descrevem esses repositórios como:

> [...] usualmente hospedados nos países da América do Norte e Europa. [...] são repositórios de dados de pesquisa temáticos, focados no armazenamento de textos e imagens, com pouca representatividade no que diz respeito a certificação e padrões de metadados. [...] De forma geral, a maioria dos repositórios de dados distribuem seus conteúdos sob licença de *Copyright*, embora seus dados estejam caracterizados como de acesso aberto. [...]

Conforme o contexto explanado, existe um ambiente consolidado de repositórios de dados de pesquisa que estão registrados nas mais diversas áreas do conhecimento e distribuídos globalmente. No que segue, discorre-se sobre o problema de pesquisa que irá nortear esta investigação.

## 1.2 POR QUE E COMO OS BIBLIOTECÁRIOS DE DADOS DEVEM SE PREOCUPAR COM REPOSITÓRIOS DE DADOS DE PESQUISA?

Os repositórios de dados de pesquisa são parte da ciberinfraestrutura de *e-Science* que os bibliotecários de dados devem dominar. Assim, uma das propostas desta tese será destacar quais são os conhecimentos e as habilidades necessários para coleta, manipulação, análise e visualização de dados disponíveis em repositórios de dados de pesquisa. Portanto, supõe-se que algumas dessas habilidades (**ver capítulo 2.4.1)** são aquelas exigidas dos cientistas de dados.

A *Data Science* fornece novas habilidades, técnicas, tecnologias e métodos que podem ser utilizados pelos bibliotecários de dados para que trabalhem com coleta, manipulação, análise e visualização de dados de pesquisa. Tal exigência justifica a necessidade de se conhecer o cenário internacional de repositórios de dados de pesquisa, seja para depositar dados, seja para buscar dados para realizar uma nova investigação.

Nesse sentido, algumas perguntas iniciais podem ser elencadas. Quantos repositórios de dados estão registrados no re3data? Quais são as principais instituições? Em quais os países e em quais línguas os dados de pesquisa estão disponíveis? Qual a tipologia desses

repositórios? Quais tipos de dados são depositados? Quais as temáticas e as áreas do conhecimento? Quais os tipos de acesso e as licenças utilizadas? Qual o tipo de *software*? Não é objetivo deste estudo analisar todo o *output* dos repositórios de dados de pesquisa registrados no re3data, mas sim propor uma pesquisa sobre o domínio das Geociências[12] e avaliar como os RDPGs elaborados com *Ckan*, *Dspace*, *Dataverse*, *E-Prints* e *Fedora* (**ver capítulo 2.5.2**) estão representados no re3data. As respostas para esses problemas são apresentadas no **capítulo 3 da tese**.

A complexidade e o ineditismo desses problemas são amplificados se pensarmos que os bibliotecários de dados devem assimilar habilidades teórico-técnicas junto ao tripé CI, *e-Science* e *Data Science* para elaboração de serviços de gestão, curadoria, preservação e acesso a dados de pesquisa por meio do uso das tecnologias digitais de dados como os repositórios.

Após apresentados o tema, o contexto, o problema de pesquisa, traçam-se os objetivos dessa pesquisa.

[12] A escolha sobre do domínio das Geociências compõe o interesse pessoal do autor da tese. O proponente é bibliotecário no Instituto de Geociências da Universidade Federal do Rio Grande do Sul (IGEO/UFRGS), instituição que apoia e fomenta os estudos de doutorado junto ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Santa Catarina (PGCIN/UFSC). O trabalho no IGEO/UFRGS facilita o acesso às informações sobre Geociências e proporciona uma melhor convivência com o tema.

## 1.3 OBJETIVOS

O objetivo geral desta tese é explorar como as relações entre Ciência da Informação, *e-Science* e *Data Science* podem consolidar-se como bases para Biblioteconomia de Dados. A fim de alcançar esse objetivo, buscam-se subsídios nos seguintes objetivos específicos:

a) delimitar os conceitos de informação, tecnologia, dados, dados de pesquisa, bibliotecários de dados, habilidades técnicas para bibliotecários de dados, sistemas de gerenciamento e elaboração de repositórios de dados de pesquisa, *softwares* e linguagens de programação utilizados para manipulação de dados;
b) experimentar processos e tecnologias de pesquisa utilizadas por cientistas de dados para manipulação de dados de pesquisa extraídos de RDPG;
c) realizar uma análise descritiva sobre a composição internacional dos RDPGs;
d) identificar o corpus de temáticas, países, licenças, *softwares,* instituições e tipos de dados de pesquisa mais utilizados pela área de Geociências;

Após apresentados os objetivos, é preciso propor uma metodologia que dê conta dos pressupostos teóricos e práticos até aqui discutidos.

## 1.4 MÉTODO

O método de investigação desta tese, além de fundamentar-se em uma pesquisa bibliográfica (**capítulo 2),** busca a realização de um estudo exploratório e descritivo a respeito do cenário internacional dos repositórios de dados de pesquisa em geociências (RDPG) disponíveis no re3data (**capítulo 3**).

Em suma, a caracterização do método aplicado neste estudo pode ser visualizada no Quadro 1.

Quadro 1 - Caracterização do método.

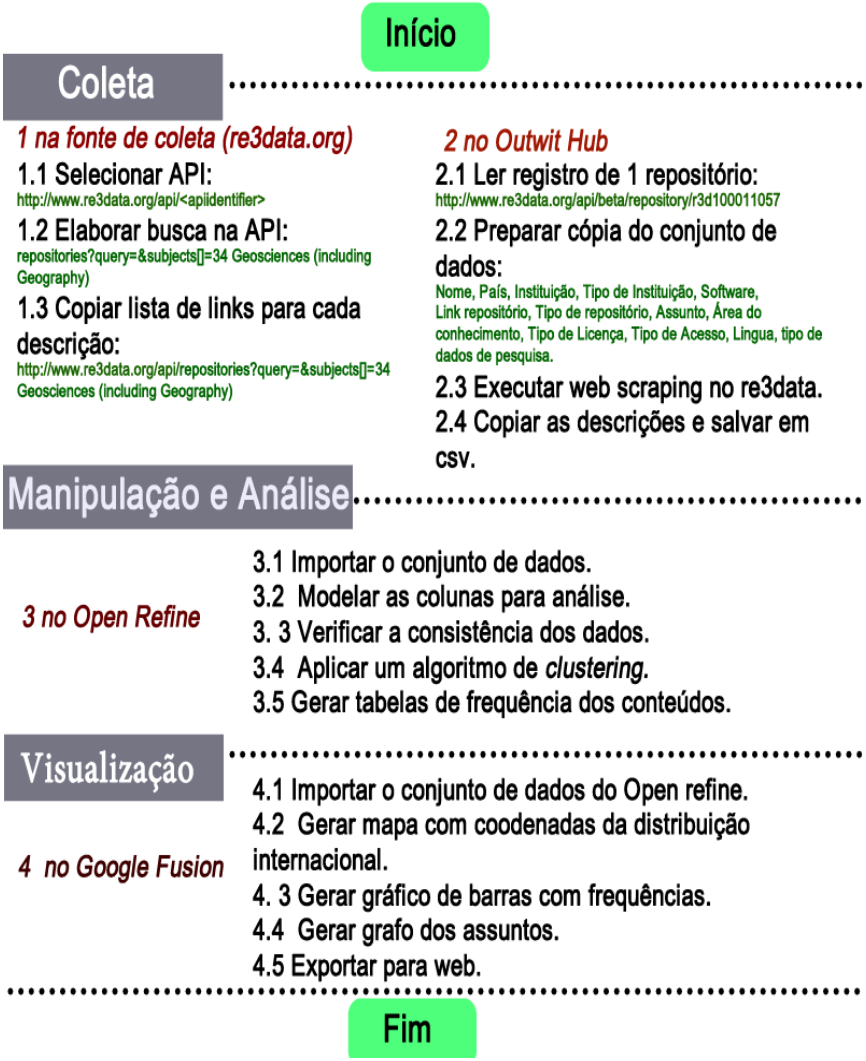
| CARACTERIZAÇÃO DO MÉTODO | |
|---|---|
| **TIPO DE PESQUISA** | Teórica, Exploratória e Descritiva |
| **ESTRATÉGIA** | Uso de *softwares*, linguagens de programação e fontes de dados da *web* para validar as competências necessárias aos bibliotecários de dados |
| **MÉTODOS DE ANÁLISE** | Pesquisa bibliográfica<br>Análise exploratória e descritiva (*clustering* e frequências) |
| **NATUREZA** | Qualitativa e quantitativa |
| **PANORAMA GERAL DA INVESTIGAÇÃO** | Repositórios de dados de pesquisa |
| **AMOSTRA** | Geociências |
| ***CORPUS*** | Repositórios de dados de pesquisa em Geociências |
| **INSTRUMENTOS DE COLETA** | *Web scraping* do esquema de metadados do re3data |
| ***SOFTWARES*** | *Output hub*<br>*Open refine*<br>*Google Fusion* |
| **FONTES DE COLETA DE DADOS** | re3data.org |

Fonte: Dados da Pesquisa (2017).

O desenvolvimento desta pesquisa exigiu a aplicação de conhecimentos inerentes a técnicas e tecnologias utilizadas para a análise descritiva de dados, para recuperação de informação, para manipulação e análise dos dados textuais e para visualização. Também é preciso ressaltar que os procedimentos metodológicos exigem a compreensão de alguns conceitos tratados na fundamentação teórica – dados de pesquisa (**ver capítulo 2.3.1**), repositórios **(ver capítulo 1.2 e 2.5.2)**, *Data Science* (**capítulo 2.5.1**) e *softwares* para manipulação de dados **(ver capítulo 2.5.5)** e linguagens de programação para análise de dados (**capítulo 2.5.6**). A metodologia irá pautar-se neles.

Assim, o estudo está dividido nas seguintes etapas: coleta de dados (*web scraping*), manipulação e análise descritiva da frequência dos dados (limpeza, organização, tratamento de desambiguação e formato de dados) e visualização (criação de gráficos e grafos). Conforme figura 1:

Figura 1 - Fluxo metodológico.




Fonte: Dados da Pesquisa (2017).

A seguir, são esclarecidos os procedimentos práticos de coleta e análise do *corpus* de dados utilizados neste estudo. Na coleta de dados, descrevem-se as técnicas e as ferramentas de *softwares* utilizadas para operacionalização da pesquisa.

### 1.4.1 Coleta de dados

A fonte de coleta de dados é o re3data, que em (set. 2017) indexava cerca de 1905 repositórios de dados de pesquisa. Os metadados[13] com as descrições recuperadas no re3data servem para delimitar a distribuição internacional dos RDPGs. Neles é possível selecionar informações para uma análise descritiva dos RDPGs. A coleta dos dados buscou selecionar informações sobre a composição internacional dos RDPGs. Para tal, utilizou-se a API (http://www.re3data.org/api/<apiidentifier>) disponível no re3data. O re3data oferece a recuperação de todo o seu conteúdo ou de parte dele via API. Atualmente, a plataforma oferece uma implementação de pesquisa aberta simples, bem como uma versão de *interface representational state transfer* (REST) para extração de dados (PAMPEL et al., 2013; COSTA; BRAGA, 2016; RE3DATA, 2017; KINDLING et. al., 2017).

Assim, por meio da utilização da API de interface REST do re3data, selecionaram-se os RDPGs. A consulta no re3data resultou em uma lista de *links*[14] contendo o endereço de cada uma das descrições dos RDPGs. O esquema utilizado pelo re3data para descrever os repositórios está na versão 3.0 de dezembro de 2015[15]. Ele representa a estrutura dos dados que serão analisados. O padrão está disponível em XML e contém propriedades de metadados sobre o escopo geral, o conteúdo, a infraestrutura, os padrões técnicos e de qualidade para repositórios de dados de pesquisa (RE3DATA, 2017).

Para operacionalizar a coleta das descrições dos RDPGs, utilizou-se o *software OutWitHub* (**ver capítulo 2.5.5)**. Com ele, foi possível coletar cada uma das descrições da lista de *links* de RDPGs. Na prática, a coleta foi a execução de uma aplicação de *web scraping* no

---

[13] O termo metadados, segundo Caplan (2003, p. 10), refere-se "[...] a todos os dados utilizados para auxiliar na identificação, descrição e localização de recursos eletrônicos em rede [...]". A autora (2003, p. 14) explica que "[...] o prefixo meta é comumente usado para significar algo sobre algo, por exemplo, uma metalinguagem usada para descrever outra linguagem [...]". Eles são informações que se encontram acopladas, embutidas ou diretamente associadas na maioria dos objetos digitais. Servem para descrição, administração e mapeamento da estrutura e do conteúdo desses objetos digitais (CAPLAN, 2003; MOORE; MOREAU, 2001).

[14] Disponível em: <http://www.re3data.org/api/beta/repositories?query=&subjects[]=34%20Geosciences%20(including%20Geography)>. Acesso em: maio 2016.

[15] ESQUEMA RE3DATA. Disponível em: <http://gfzpublic.gfz-potsdam.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:1397899>. Acesso em: ago. 2017.

servidor de APIs do re3data. O *web scraping* dos dados coletados no re3data possibilitou a seleção de informações essenciais sobre os RDPGs. Após o *web scraping*, foi feita uma cópia dos dados em formato XML. Essa cópia contém todos os campos descritivos, conforme o esquema de metadados do re3data; os dados extraídos podem ser visualizados na Figura 2.

Figura 2 - Conjunto de dados coletados do re3data.

Find in page key

| Active | Name of Scraper | Apply If Page URL Contains |
|---|---|---|
| ☑ | conjunto de dados(re3data) | http://www.re3data.org |

| | OK | Description | Marker Before | Marker After |
|---|---|---|---|---|
| 1 | ☑ | Nome | <r3d:repositoryName language="eng"> | </r3d:repositoryName> |
| 2 | ☑ | Pais | <r3d:institutionCountry> | </r3d:institutionCountry> |
| 3 | ☑ | Instituição | <r3d:institutionName language="eng"> | </r3d:institutionName> |
| 4 | ☑ | Tipo de Instituição | <r3d:institutionType> | </r3d:institutionType> |
| 5 | ☑ | Descrição | <r3d:description language="eng"> | </r3d:description> |
| 6 | ☑ | Tipo de Software | <r3d:softwareName> | </r3d:softwareName> |
| 7 | ☑ | Link | <r3d:repositoryURL> | </r3d:repositoryURL> |
| 8 | ☑ | Tipo de Reposiório | <r3d:type> | </r3d:type> |
| 9 | ☑ | Assunto | <r3d:keyword> | </r3d:keyword> |
| 10 | ☑ | Área do conhecimento (... | <r3d:subject subjectScheme="DFG"> | </r3d:subject> |
| 11 | ☑ | Tipo de Licença | <r3d:dataLicenseName> | </r3d:dataLicenseName> |
| 12 | ☑ | Tipo de acesso aos dados | <r3d:dataAccessType> | </r3d:dataAccessType> |
| 13 | ☑ | Linguagem | <r3d:repositoryLanguage> | </r3d:repositoryLanguage> |
| 14 | ☑ | Tipo de dados | <r3d:contentType contentTypeScheme... | </r3d:contentType> |

Fonte: Dados da Pesquisa (2017).

O resultado da requisição de coleta ao re3data pode ajudar a responder às perguntas iniciais em relação às quantidades e à categorização dos RDPGs, como: quantos repositórios de dados de pesquisa registrados no re3data são sobre Geociências? Quais são as principais instituições? Quais os países e em quais línguas os dados de pesquisa estão disponíveis? Qual a tipologia desses repositórios? Quais os tipos de dados depositados? Quais as temáticas e as áreas do conhecimento? Quais os tipos de acesso e as licenças utilizadas? Qual o tipo de *software*? Essas são algumas das perguntas que podem ser lançadas durante a interpretação de dados.

### 1.4.2 Manipulação e análise de dados

Após coletados, os dados necessitam ser manipulados para atenderem aos requisitos de preparação para análise. Nesse ponto, é preciso a incorporação de algum *software* de refinamento de dados que realize a limpeza, a recodificação, caso necessário, a mesclagem de conjuntos de dados, a importação e a exportação dos dados de um tipo de *software* para outro sem perder conteúdo.

A geração de conjuntos de dados analíticos pode ser feita com a utilização do *OpenRefine* (**ver capítulo 2.5.5**), que serve para carregar, limpar, reconciliar, categorizar e converter dados de um formato para outro. A utilização do *OpenRefine* segue o fluxo:

a) importar o conjunto de dados;

b) modelar as colunas para análise;

c) verificar a consistência dos dados;

d) aplicar um algoritmo de *clustering* para procurar palavras que estejam escritas de maneira diferente, como "*New York"* e *"new York"*. O *Open Refine* permite a utilização do método *key-colision* e a função *fingerprint* para encontrar grupos de valores diferentes que podem ser representações alternativas da mesma entidade;

e) gerar conjuntos de dados com a frequência de ocorrência dos conteúdos contidos no conjunto de dados;

f) converter o conjunto de dados para análise e para visualização.

Somente após o refinamento é possível realizar a sumarização dos dados e criar diferentes tipos de tabelas e visualizações. É relevante lembrar que o processo de refinamento permite a sumarização pautada na descrição das frequências e na mineração[16] dos textos que compõem o conjunto de dados coletados.

Após realizar tais procedimentos de pesquisa, é possível realizar a sumarização dos dados por meio da criação de diferentes tipos de gráficos.

---

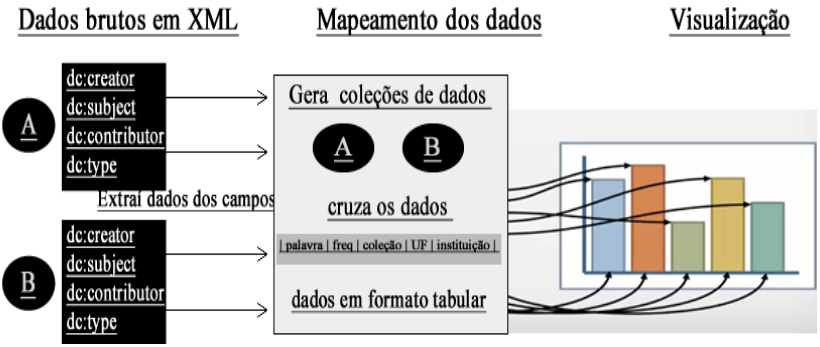
[16] Inicialmente, o termo "mineração de dados"' tinha uma conotação negativa, especialmente entre os estatísticos. Termos como *data snooping*, *fishing* e *data dredging* referem-se a técnicas *ad-hoc* para extrair conclusões a partir de dados sem uma base estatística sólida. No entanto, ao longo do tempo, a disciplina de mineração de dados tornou-se madura e, atualmente, caracteriza-se como método científico sólido com muitas aplicações práticas. As tarefas de mineração de dados típicas são: *classification* (e.g., *constructing a decision tree*), *clustering*, *regression*, *summarization* e *association rule learning*. Todas essas são baseadas em técnicas de dados tabulares simples em que as linhas correspondem às instâncias e as colunas correspondem às variáveis (VAN DER AALST, 2014).

### 1.4.3 Visualização

Após a coleta, a manipulação e a análise, é preciso tratar e preparar os dados para visualização. Esse processo consiste em ajustar os dados para que atendam aos requisitos da visualização. Por exemplo, para criar um gráfico de barras, será preciso gerar dados analíticos sobre a frequência dos resultados da coleta os quais serão a base para o desenho da visualização. Esse processo serve para geração de conjuntos de dados secundários e analíticos que serão utilizados para modelar a visualização.

A Figura 3 explica o fluxo de transformação dos metadados para algum formato de dados para visualização, sendo um movimento de dados primários para dados secundários.

Figura 3 - Processo de criação de conjuntos de dados para visualização.




Fonte: Dados da Pesquisa (2017).

A visualização funciona como estratégia de redução e interpretação dos resultados oriundos das análises da qualidade dos metadados e apresenta informações que facilitam a compreensão dos dados. É na visualização que se escolhe a representação visual mais adequada aos dados.

Na visualização, os dados são utilizados para responder às perguntas de pesquisa, testar hipóteses e explorar relacionamentos. A visualização é uma ferramenta que permite apresentar grandes quantidades de dados em uma pequena quantidade de espaço (BONNEAU et. al., 2014).

Os gráficos são usados por uma série de razões, podendo ser encontrados em jornais, revistas, televisão, etc. Eles ajudam a comunicar informações. Existem tipos diferentes de gráficos e cada um tem algo diferente sobre aquilo que o torna útil de uma forma única. Para se criar

um gráfico, é preciso primeiro conhecer o tipo de informação que se deseja transmitir. Um gráfico informa as tendências de uma série de valores em relação a um determinado espaço de tempo ou compara duas ou mais situações. Cada tipo de gráfico é adequado para uma diferente situação a ser analisada (MSTE, 1997).

No contexto desta investigação, propõe-se a geração de gráficos simples (barras[17]) com o resultado da transformação dos metadados obtidos durante a coleta para, assim, realizar experimentação com diferentes formas de visualização. Alguns dos tipos de visualização gráfica comuns são: gráfico de linha, gráfico de dispersão, gráfico de barras e histogramas. A representação de dados e informação na forma de gráficos é mais ampla; no entanto, esses tipos são suficientes para a aplicação prática desse método.

Outra forma simples de representar dados são os grafos. Os grafos são objetos práticos para estudar as relações entre os objetos. Além de sua essência matemática, o grafo assume uma estrutura de dados potencializada pelas tecnologias, em especial pela visualização da informação[18]. Os grafos permitem representar modelos complexos por meio de estruturas simples, o que compreende um grande número de situações, expressando os relacionamentos e as dependências entre os elementos. O objetivo da teoria dos grafos é resolver os problemas

---

[17] Gráfico de barras é um tipo de gráfico usado para mostrar e comparar números, frequência ou outra medida que possa ser utilizada com diferentes categorias de dados discretas. É um tipo de gráfico flexível. Existem diversas variações do gráfico de barras padrão, incluindo gráficos de barras horizontais, agrupadas ou gráficos de componentes e gráficos de barras empilhadas (MSTE, 1997; SAYÃO, 200?).

[18] A prática com a Visualização da Informação (InfoVis), desde o seu início, na segunda parte do século XVIII[18], até os dias de hoje, baseia-se em dois princípios fundamentais: a redução e o uso de variáveis espaciais. O primeiro princípio, a redução, utiliza-se de primitivas gráficas, como pontos, linhas retas, curvas e formas geométricas simples, para compreender os objetos e as relações entre eles. O segundo princípio é o uso de variáveis espaciais (posição, tamanho, forma e movimento) para representar as principais diferenças nos dados e revelar padrões e relações. Na primeira parte do século XIX, foram inventados quase todos os tipos de gráficos comumente encontrados hoje em programas estatísticos e gráficos. Percebe-se que o desenvolvimento de novas técnicas para a redução visual, paralelo à trajetória reducionista da ciência moderna do século XIX (física, química, biologia, linguística, psicologia e sociologia), propõe que ambos os mundos, natural e social, devam ser entendidos em termos de elementos simples (moléculas, átomos, fonemas etc.) e suas regras de interação De todas as técnicas de visualização, percebe-se algo em comum entre elas: todas usam variáveis espaciais (posição, tamanho, forma e, mais recentemente, curvatura das linhas e movimentos) para representar as principais diferenças nos dados e revelar padrões e relações entre eles. As variáveis espaciais são o segundo princípio fundamental da InfoVis, percebidas desde o gráfico de linha (1711), o gráfico de barras (1786) até os gráficos de pizza (1801), e a sua ubiquidade hoje está em todos os *softwares* gráficos de representação de dados, como E*xcel, Numbers, Tableau, Google Docs e Open Office*, entre outros *softwares* de manipulação de dados (MANOVICH, 2011).

simplificando sua representação **(**GONZALES-AGUILAR **et. al., 2017).**

A ferramenta utilizada para gerar a visualização dessa tese foi o Google Fusion. Além de gráficos e grafos, ele reconhece dados de latitude e longitude do Google Maps e outros tipos de dados e/ou serviços oferecidos pelo Google (**ver capítulo 2.5.5**).

# 2 FUNDAMENTAÇÃO TEÓRICA

O escopo deste capítulo é desenvolver os objetivos específicos (a) delimitar os conceitos de informação, tecnologia, dados, dados de pesquisa, bibliotecários de dados, habilidades técnicas para bibliotecários de dados, sistemas de gerenciamento e elaboração de repositórios de dados de pesquisa, *softwares* e linguagens de programação utilizados para manipulação de dados; e (b) experimentar processos e tecnologias de pesquisa utilizadas por cientistas de dados para coleta, manipulação, análise e visualização de dados de pesquisa extraídos de RDPGs.

## 2.1 INFORMAÇÃO E TECNOLOGIA, ASPECTOS GERAIS EM CIÊNCIA DA INFORMAÇÃO

O conceito de informação e o de tecnologia são fundamentais para compreensão do que é CI. Assim, torna-se necessário descrever as relações existentes entre esses conceitos. Apoiando-se na relação epistemológica entre informação e tecnologia, busca-se identificar as bases de ligação entre a CI, a *e-Science* e a *Data Science*.

Tendo em vista essa proposição, acredita-se em um *continuum* tecnológico que irá relacionar o surgimento da CI com os desdobramentos dos conceitos de informação e tecnologia atualmente. Ressalta-se a CI enquanto disciplina que investiga as propriedades da informação e da tecnologia, sendo ela preocupada com o corpo de conhecimentos relacionados à organização, ao armazenamento, à recuperação, à transformação e à utilização da informação por meio da tecnologia, sendo a CI uma ciência interdisciplinar derivada de campos como a Ciência da Computação e Biblioteconomia (BORKO, 1968).

Assim, apresentam-se as origens da CI e os seus desdobramentos em relação aos conceitos de informação e tecnologia.

### 2.1.1 Ciência da Informação

A epistemologia ou teoria do conhecimento preocupa-se com questões sobre o escopo e os limites do que e de como conhecemos algo. As suas fontes teóricas procuram respostas no problema da justificação do saber. Ela lida com argumentos céticos, ou seja, com a negação de que possamos ter conhecimento sobre alguma coisa, seja científica, seja de senso comum (GRECO, 2008).

A epistemologia em Ciência da Informação ou a teoria do conhecimento sobre a CI preocupa-se com questões sobre o cerne conceitual do campo e suas definições. Simples indagações podem ser usadas para definir o campo: o que é CI? Quais suas origens? Quais são os conceitos fundamentais? Quem os cunhou? O que devemos conhecer para compreender e/ou explicar a CI? Tendo em vista a epistemologia aplicada ao campo da CI, propõe-se neste estudo uma revisão de conceitos amplamente adotados pela comunidade de cientistas da informação, sendo objetivo específico delimitar as bases para definição dos termos informação e tecnologia.

Independentemente da estrutura que venha a ser adota para definir a CI, torna-se relevante compreender que os problemas informacionais são, em sua natureza, internacionais ou globais. A evolução da CI nos vários países ou regiões acompanhou diferentes acontecimentos ou prioridades distintas, mas a justificativa e os conceitos básicos são os mesmos globalmente. Cada escola de pensamentos sobre CI assume um carácter relevante e específico, pois propõe definições precisas e fundadas na prática dos cientistas da informação de cada país em específico (SARACEVIC, 1996; PINHEIRO; LOUREIRO, 1998).

Fundamentando-se na evolução histórica descritiva das tecnologias de informação e comunicação, o campo da CI, no Brasil e no mundo, muitas vezes revela-se como uma Ciência Social Aplicada e pós-moderna. Lena Vânia Pinheiro e José Mauro Matheus Loureiro (1998) desenvolvem a ideia de que a CI tem seu próprio estatuto científico, como Ciência Social interdisciplinar, e possui como origem os processos de comunicação da informação que se desenvolvem em diferentes territórios: científicos, tecnológicos, educacionais, sociais, artísticos e culturais; portanto, múltiplos contextos e condições experimentais.

Já Gernot Wersig (1993) vê a CI como um protótipo de ciência pós-moderna. O autor busca relatar seus processos comunicativos, cognitivos e tecnológicos a fim de compreender a interação do homem com a informação, com base na investigação sociotécnico-cultural da CI, trazendo à tona o debate sobre os diversos discursos das ciências, da filosofia e das práticas com a tecnologia. Assim, talvez seja possível pensar a CI como uma ciência pós-moderna e interdisciplinar, como a ecologia (WERSIG, 1993).

Pinheiro e Loureiro (1998) descrevem acontecimentos que vão das ideias de Vannevar Bush (1945), com sua máquina de associações (Memex), ao surgimento do Instituto Brasileiro de Bibliografia e

Documentação (IBBD), em 1954, depois denominado, em 1976, IBICT, e a criação de cursos de pós-graduação em CI no Brasil. Sempre pautada na interdisciplinaridade e na influência da tecnologia sobre os estudos da informação, a autora ratifica a ligação da CI com as tecnologias, conectando-se às ideias de Tefko Saracevic (1996) sobre a interconexão da CI com os desenvolvimentos da Sociedade da Informação.

As narrativas clássicas sobre CI surgem nos USA na década de 1960, em reuniões de cientistas do Georgia Institute for Technology. As reuniões da American Society for Information Science and Technology (ASIST) constituem as primeiras iniciativas para definir a CI enquanto atividade profissional e disciplina científica nos USA. Focada nos estudos sobre a informação científica e a tecnologia, a ASIST definiu CI em seus encontros na década de 1960, conforme Saracevic (1996) e os professores da London University City Bawden e Robinson (2012) explicam. A primeira definição sobre CI foi publicada por Harold Borko, em 1968:

> Ciência da Informação é a disciplina que investiga as propriedades e o comportamento informacional, as forças que governam os fluxos de informação e os significados do processamento da informação, visando à acessibilidade e à usabilidade ótima. A Ciência da Informação está preocupada com o corpo de conhecimentos relacionados à origem, à coleção, à organização, ao armazenamento, à recuperação, à interpretação, à transmissão, à transformação e à utilização da informação. Isso inclui a pesquisa sobre a representação da informação em ambos os sistemas, tanto naturais quanto artificiais, o uso de códigos para a transmissão eficiente da mensagem, bem como o estudo do processamento e de técnicas aplicadas aos computadores e a seus sistemas de programação. É uma ciência interdisciplinar derivada de campos relacionados, tais como a Matemática, a Lógica, a Linguística, a Psicologia, a Ciência da Computação, a Engenharia da Produção, as Artes Gráficas, a Comunicação, a Biblioteconomia, a Administração e outros campos científicos semelhantes. Tem ambos os componentes de ciência pura, visto que investiga seu objeto sem considerar sua aplicação, e um componente de

> ciência aplicada, visto que desenvolve serviços e produtos (BORKO, 1968, p. 3, tradução nossa)[19].

O núcleo conceitual da CI investiga questões teórico-práticas ligadas aos comportamentos da informação e preocupa-se com as interações que ocorrem entre a informação e a tecnologia. Globalmente, difunde-se e reúne-se sob as tradições e as práticas sobre o registro documental e o nascimento da computação contemporânea. A transmutação da informação do orgânico para o digital/eletrônico revelou-se como um objeto de estudos complexo para CI. As disciplinas de Documentação/Bibliografia, a Recuperação da Informação, a Teoria da informação e a Cibernética estão claramente impressas nas palavras de Borko e revelam estar nas origens do conceito de CI.

A corrente europeia da CI explica sua origem na Documentação, precisamente nas experiências do advogado belga Paul Otlet, em seu *Tratado de Documentação* (1934), e nos estudos sobre a Bibliografia Estatística. Assim, percebe-se que a CI surge das transformações ocorridas na prática da Documentação e da Bibliografia:

> O surgimento de uma nova disciplina ou profissão tem sido tipicamente reconhecido, desde o Século 19, por meio da (criação) de organizações profissionais (documentalistas). A primeira organização desse tipo na área de Ciência da Informação gerou um órgão internacional que coordenou as atividades do movimento da documentação. O Institut International de Bibliographie (IIB), mais tarde conhecido como Federation for Information and Documentation (FID), foi criado pelos dois pioneiros belgas da documentação, Paul Otlet e Henri La Fontaine,

---

[19] *"Discipline that investigates the properties and behavior of information, the forces governing the flow of information, and the means of processing information for optimum accessibility and usability. It is concerned with that body of knowledge relating to the origination, collection, organization, storage, retrieval, interpretation, transmission, transformation, and utilization of information. This includes the investigation of information representations in both natural and artificial systems, the use of codes for efficient message transmission, and the study of information processing devices and techniques such as computers and their programming systems. It is an interdisciplinary science derived from and related to such fields as mathematics, logic, linguistics, psychology, computer technology, operations research, the graphic arts, communications, library science, management, and other similar fields. It has both a pure science component, which inquires into the subject without regard to its application, and an applied science component, which develops services and products"* (BORKO, 1968, p. 3).

> em 1895. Durando até o novo milênio, pode reivindicar ser a primeira associação reconhecida de Ciência da Informação (BAWDEN; ROBINSON, 2012, p. 10, tradução nossa).[20]

Assim, o surgimento da Ciência da Informação, anos mais tarde, em 1968, nos USA, mostra um rompimento dos cientistas da informação norte-americanos com as áreas da Documentação e da Bibliografia. Esses cientistas preferiram a informação científica, e não o documento, para representar suas atividades de pesquisa. Outra definição de CI proposta por Saracevic aponta a tecnologia como relacionada a serviços de informação:

> […] a ciência e a prática de lidar com a efetiva coleta, o armazenamento, a recuperação e o uso de informação. Esta preocupada com informação registrada e com os conhecimentos, as tecnologias e os serviços que facilitem a sua gestão e utilização (SARACEVIC, 2010, p. 2570, tradução nossa)[21].

O supracitado Tefko Saracevic também descreve as principais características do campo da CI: a interdisciplinaridade, a forte ligação com a tecnologia e com os desenvolvimentos da Sociedade da Informação.

> […] Primeiro, a CI é, por natureza, interdisciplinar, embora suas relações com outras disciplinas estejam mudando. A evolução interdisciplinar está longe de ser completada. Segundo, a CI está inexoravelmente ligada à tecnologia da informação. O imperativo

---

[20] "*The establishment of a new discipline or profession has typically been recognized, since the 19th century, by the setting up of a professional body to represent it. The first such body in the area that was to spawn information science was an international body for the co-ordination of the activities of the documentation movement. The Institut International de Bibliographie (IIB), later renamed as the International Federation for Information and Documentation (FID), was established by the two Belgian pioneers of documentation, Paul Otlet and Henri La Fontaine, in 1895. Lasting until the new millennium, it can claim to be the first recognizable information science* association" (BAWDEN; ROBINSON, 2012, p. 10).

[21] "*[...] the science and practice dealing with the effective collection, storage, retrieval and use of information. It is concerned with recordable information and knowledge, and the technologies and related services that facilitate their management and use"* (SARACEVIC, 2010, p. 2570).

> tecnológico determina a CI, como ocorre também em outros campos. Em sentido amplo, o imperativo tecnológico está impondo a transformação da sociedade moderna em sociedade da informação [...]. Terceiro, a CI é, juntamente com muitas outras disciplinas, uma participante ativa e deliberada na evolução da sociedade da informação. A CI teve e tem um importante papel a desempenhar por sua forte dimensão social e humana, que ultrapassa a tecnologia (SARACEVIC, 1996, p. 42).

Tal cenário lançado por Tefko Saracevic reforça a proposição clássica da CI que tem em mente a tecnologia como base epistemológica para suas investigações. No tratamento desse problema, a CI tem interesse particular em usufruir das tecnologias. Embora Saracevic não se preocupe em desenvolver questões mais específicas sobre a tecnologia, ele trata a CI como um campo multifacetado e busca compreender a realidade das questões informacionais com base no comportamento e nas interações entre o homem, a informação e a tecnologia.

Esse panorama influencia pesquisadores brasileiros, como Pinheiro e Loureiro (1998), que acreditam que o conjunto de disciplinas que compõem a CI só podem ser pensadas com base na interdisciplinaridade. Os autores procuram suas respostas em Saracevic e buscam compreender como outros campos do conhecimento contribuem para a CI, investigando conceitos, princípios, técnicas, métodos e teorias que influenciam o campo, como a Teoria da Informação, a Cibernética e a Teoria Geral dos Sistemas (PINHEIRO; LOUREIRO 1998).

Nesse contexto, a CI surge da interseção entre a Recuperação de Informação, a Teoria da Informação, a Organização da Informação e a Cibernética. As origens dos estudos sobre a recuperação de informação em registros documentais impulsionam o desenvolvimento de pesquisa em CI desde os anos de 1950.

> […] o trabalho com a recuperação de informação foi responsável pelo desenvolvimento de inúmeras aplicações bem-sucedidas (produtos, sistemas, redes, serviços). Mas, também, foi o responsável por duas outras coisas: primeiro, pelo

> desenvolvimento da CI como um campo onde se interpenetram os componentes científicos e profissionais. […] segundo, a recuperação da informação influenciou a emergência, a forma e a evolução da indústria informacional (SARACEVIC, 1996, p. 45).

A questão investigativa da década de 1950 foi a recuperação de informação. Nas palavras de Gilchrist (2014), no artigo *The information Paradox*, percebe-se uma tendência atual nesse contexto de estudos da CI:

> A teoria da informação de Shannon é a teoria da transmissão de sinais, independentemente do seu possível significado; a recuperação de informação, cunhada por Calvin Mooers, foi considerada por Robert Fairthorne como a recuperação de referências (hoje o Google recupera *web sites*); a revolução de informação é realmente a revolução da comunicação, e assim por diante (GILCHRIST, 2014, p. 13, tradução nossa)[22].

A CI desenvolveu uma espécie de pensamento técnico com base na Recuperação de Informação (RI). As primeiras experiências de RI foram realizadas com cartões perfurados na Cranfield University e originaram um tipo de pensamento pragmático pautado na probabilidade de recuperar informação em bases de dados. Essas visões impulsionaram o surgimento da RI e das métricas de avaliação da relevância da informação nos processos de RI. Nesse sentido, ressalta-se a importância da RI para definição das práticas que irão consolidar os estudos sobre a CI.

As primeiras métricas de RI – *Precision* e *Revocation* – cunhadas por Alen Kent (entre os anos de 1951 e 1955) e desenvolvidas na Cranfield University (CLEVERDON, 1960) foram propostas para medir a qualidade dos resultados que são recuperados em sistemas de recuperação da informação (SRI). A precisão é a métrica que mede como um SRI recupera somente os documentos relevantes, é a fração

---

[22] "*Shannon's "Information theory" is actually the theory of the transmission of signals, independent of their possible meaning; "Information retrieval", coined by Calvin Mooers was judged by Robert Fairthorne to be "Reference retrieval" (and even today Google retrieves web-sites); the "Information revolution" is actually the "Communications revolution", and so on"* (GILCHRIST, 2014, p. 13).

dos documentos recuperados que são relevantes para quem os busca. Já a revocação mede como um SRI recupera todos os documentos relevantes, é a fração dos documentos relevantes que foi recuperada. Tais métricas são simples e exploradas pelos cientistas da informação até os dias de hoje (BAEZA-YATES; RIBEIRO-NETO, 2013).

A ideia de recuperação de informação será retomada no capítulo 2.5.3 desta tese, na seção de *softwares* científicos por meio da abordagem sobre máquinas de busca como o *web scraping* e o *harvesting* de metadados. Por ora, acredita-se ser fundamental esclarecer o conceito de informação, fundamental para que se possa compreender a CI em relação ao objeto de estudo desta tese.

### 2.1.2 O conceito de informação

O conceito de informação é fundamental para se compreender a CI enquanto disciplina científica. O lugar epistemológico da informação relaciona-se ao ponto de vista de quem a estuda; sua aplicação conceitual depende de contexto específico.

Nesse sentido, Silva e Ribeiro (2012) propõem um conceito de informação para ser aplicado junto a disciplinas de Documentação como a Biblioteconomia e a Arquivologia, entre outros aspectos aplicados a sistemas de informação tecnológica. Para esses autores, a informação é um conjunto estruturado de representações mentais e emocionais codificadas (sinais e símbolos) que são modeladas pela interação social. A informação é algo capaz de ser registrado em qualquer formato de suporte (físico ou digital) por conseguinte, comunicada de uma maneira assíncrona e multidirecional de forma interativa (SILVA; RIBEIRO, 2012).

O uruguaio Rafael Capurro (2007) recupera a ideia do termo desde a etimologia da palavra *infomatio* (dar forma) até os usos nas atuais abordagens que utilizam a tecnologia como base. Para esse autor, o que torna a informação importante nos dias de hoje é a sua natureza digital, e é de extrema relevância a interpretação de seus aspectos básicos. Para ele, filósofo de corrente hermenêutica, a CI reflete uma visão do conceito de informação e torna necessária a realização de uma abordagem investigativa de dupla via, objetiva e subjetiva, sendo o conceito de informação envolto em interpretações diversas, configurando-se como um elo entre reflexões pragmáticas e filosóficas sobre uma visão da informação como fenômeno ligado a ambientes tecnológicos (CAPURRO, 2007).

O núcleo filosófico das investigações sobre a informação, como descrito por Luciano Floridi (2010), busca classificar a informação como algo ligado ao impacto social da cultura tecnológica. No livro intitulado *Luciano Floridi's Philosophy of Technology*, os autores Silva e Ribeiro (2012) argumentam que a revolução tecnológica das últimas décadas aumentou a participação da sociedade no que tange aos fenômenos informacionais ligados aos meios digitais. Tal situação provoca profundas modificações no campo da CI e de disciplinas correlatas, como a Biblioteconomia (SILVA; RIBEIRO, 2012).

Outra perspectiva sobre o conceito de informação em CI é a de Michael Buckland (1991). O autor desenvolve sua compreensão sobre o conceito de informação da seguinte forma: (a) Informação-como-processo acontece quando alguém é informado sobre aquilo que conhece, e o que conhece é modificado, é o próprio ato de informar; (b) Informação-como-conhecimento é usada para denotar aquilo que é percebido na informação-como-processo, diz respeito ao conhecimento comunicado referente a algum fato particular, assunto ou evento; (c) Informação-como-coisa é atribuída a objetos, assim como dados para documentos que são considerados como informação porque se relacionam a algo informativo, tendo qualidade de conhecimento comunicado ou comunicação informativa (BUCKLAND, 1991).

O propósito de se pensar a informação como coisa é o de: (a) esclarecer seu significado em relação aos outros usos do termo; (b) estabelecer a regra fundamental da informação em sistemas de informação; e (c) trazer ordem teórica a campos heterogêneos como o da CI (BUCKLAND, 1991).

As investigações sobre a informação no campo de estudo da CI sugerem a realização de investigações críticas a respeito da natureza conceitual dos princípios básicos da informação registrada, incluindo a dinâmica de utilização da informação pelas ciências e a aplicação de metodologias e teorias para resolver problemas gerados por meio do uso do computador (FLORIDI, 2004, 2010, 2012).

Tomando tal prerrogativa como base para construção de sentido ao conceito de CI, Capurro (2007) também vê a informação como conhecimento comunicado, que precisa ser dito, explicado e compreendido. Para ele, a informação não deve ser considerada como algo isolado, mas vê-la em relação a outros conceitos e outros objetos. O autor até sugere que ver uma informação como coisa é ver a informação como um conceito subjetivo, informação como signo, isto é, como um objeto dependente de interpretação de um agente cognitivo, sendo o

significado da informação determinado pelos contextos sociais e culturais (CAPURRO, 2007).

Morville (2014) descreve a informação como a matéria-prima da comunicação. A informação permite a comunicação através do tempo e do espaço. O autor explica que ela envolve a troca de símbolos, ideias, mensagens e significados entre as pessoas. Como tal, é caracterizada pela ambiguidade, pela redundância, pela ineficiência, pelo erro e pela beleza indescritível. A comunicação é o *backbone* da sociedade, desde o surgimento das antigas tribos às nações modernas e contemporâneas. Nesse contexto, a informação é um dos principais ingredientes sociais, pois ela permite a cooperação em escala local e global, interligando algumas dezenas de membros ou uma economia global interligada por bilhões de membros (RESMINI; ROSATI, 2011; MORVILLE, 2014).

O *everywhere* tecnológico impulsiona o comportamento pervasivo da informação e transforma a existência dos seres humanos, que cada vez mais é mediada pela tecnologia. Isso não se limita às ciências, mas abrange todos os domínios do conhecimento. Sob essa premissa, apresenta-se o conceito de tecnologia.

### 2.1.3 O conceito de tecnologia

Então, o que é tecnologia? Como conceituá-la para compreender a CI em contextos de *e-Science*?

> A tecnologia é parte notória do mundo contemporâneo. [...] pode significar tanto a nossa satisfação pelos aparelhos que tornam nossa vida mais cômoda, o nosso entusiasmo ante as possibilidades que o computador e a internet nos abrem, quanto nosso temor às armas cada vez mais potentes e sofisticadas [...] aquilo que denominamos como tecnologia se apresenta como uma realidade polifacetada: não apenas em forma de objetos e conjuntos de objetos, mas também como sistemas, como processos, como modos de proceder, como certa mentalidade (CUPANI, 2013, p. 11-12).

O mais comum é pensar a tecnologia como um dispositivo técnico, um computador ou *smartphones*, *hardware*, que é uma maneira de vê-la como algo concreto. Nas palavras de Cupani (2013, p. 16), "[...] a tecnologia se manifesta antes de tudo na forma de objetos, como um

artefato material fabricado pelo homem, cuja função depende de uma materialidade específica […]". Sob esse viés, é preciso ressaltar que os dispositivos técnicos são em parte independentes do usuário e necessitam habilidades e aprendizados especiais para o seu uso e/ou manuseio (DUSEK, 2006).

Nas palavras de Cupani (2013, p. 14) "[...] a utilização de objetos técnicos requer um comportamento específico, uma regra. Tanto a produção quanto a utilização dos artefatos tecnológicos supõe a aquisição de habilidades". Conforme tal abordagem, a tecnologia pode ser vista enquanto regra. A tecnologia como regra envolve padrões de significado e relacionamento, como as regras políticas, as leis e os métodos científicos, os quais são sistematicamente desenvolvidos em domínios específicos (DUSEK, 2006).

Stephen Kline (2006), ao se perguntar o que é tecnologia, explica que a palavra é usada para representar coisas, ações, métodos e sistemas, pois, simbolicamente, pode representar procedimentos de trabalho e o progresso. Kline discute usos ou possíveis respostas para o que é tecnologia. O primeiro uso considera a tecnologia como todos os objetos não naturais; então, todos os tipos de *hardware* e/ou artefatos. Isso leva o autor a um segundo fator, ligado ao sistema sociotécnico dos objetos tecnológicos, no qual todos os objetos podem ser um tipo particular de tecnologia, ser um sistema completo que recebe *inputs* e retorna *outputs,* seja para pessoas, máquinas, recursos, processos, políticas ou economia etc (KLINE, 2006).

Conforme Cupani (2013, p. 18), a tecnologia pode ser uma forma específica de conhecimento "[...] que implica maneiras específicas de conhecer o mundo material, maneiras que incorporam o saber científico, mas que possuem igualmente características próprias". Esse fator ressalta o uso da tecnologia como metodologia racionalizada, algo que gera e responde a questões relativas ao que a tecnologia é realmente. Pode ser um método, uma técnica ou o próprio conhecimento materializado em processos de informação, habilidades e/ou procedimentos (KLINE, 2006).

Como visto, a tecnologia pode ser definida enquanto objeto, regra e conhecimento. Outra forma de pensá-la é como sistema. Esse ponto de vista sugere que os dispositivos técnicos possuem um sistema harmônico entre o *hardware* e o *software,* mas não descarta pensar a tecnologia no contexto de quem a usa.

Nesse contexto, explica Cupani (2013, p. 19), "[...] a tecnologia pode ser volição […] um modo de manifestação mais individualizado ou subjetivo […] como vontade de controle ou poder, como vontade de

liberdade, como afã de realizar um ideal humano […]". Outro fator, nesse contexto, está na combinação de *hardwares* e pessoas, a tecnologia como uma extensão das capacidades humanas. Nesse contexto, a tecnologia transforma e exterioriza a natureza do conhecimento; os relacionamentos entre ciência e tecnologia modificam as regras de experimentação e exploração científica (KLINE, 2006). Assim, a tecnologia define-se como algo polifacetado:

> [...] fabricação e uso de artefatos [...]; uma forma de conhecimento humano endereçada a criar uma realidade conforme nossos propósitos [...]; conhecimento que funciona, *know-how* [...]; implementações práticas da inteligência [...]; a humanidade trabalhando [*at work*] [...]; a colocação da Natureza à disposição do homem como recurso [...]; o campo de conhecimento relativo ao projeto de artefatos e à planificação da sua realização, operação, ajustamento, manutenção, monitoramento, à luz do conhecimento científico [...]; o modo de vida próprio da Modernidade; a totalidade dos métodos a que se chega racionalmente [...] em todo campo da atividade humana [...]; a estrutura material da Modernidade [...] em quatro dimensões: como objetos; como modo de conhecimento; como uma forma específica de atividade e como volição (determinada atitude humana perante a realidade) (CUPANI, 2013, p. 15-16).

As variáveis tecnológicas transformam os sistemas sociotécnicos[23] pelo fato de estenderem as capacidades humanas, mas isso não é um privilégio da tecnologia da idade *high-tech*; pelo contrário, os padrões da sociedade tecnológica são adotados e adaptados.

Segundo Vallverdú (2009), a *e-Science* oferece a oportunidade de se pensarem as transformações epistemológicas ocorridas nos conceitos de informação e de tecnologia de uma maneira livre de questões positivistas e deterministas.

---

[23] Inclui um ou mais sistemas técnicos, pessoas como partes inerentes do sistema e também conhecimento de como o sistema deve ser usado, o que significa que esses sistemas têm processos operacionais definidos, são regidos pelas organizações e podem ser afetados por leis e políticas regulamentadoras.

## 2.2 E-SCIENCE

Nas palavras de Jankowski (2007, p. 549, tradução nossa), "[...] *e-Science* é um dos muitos termos usados para descrever as transformações recentes ocorridas na atividade científica"[24]. Ela surge como forma de traduzir métodos e disciplinas científicas para o computador, proporcionando a colaboração científica em escala global e o uso intensivo de dados.

*A e-Science* está associada a estudos realizados em ambientes de rede, utiliza-se de ferramentas como a internet para envolver pesquisadores em ambientes de colaboração global. Utiliza-se de recursos como: (a) a colaboração internacional entre pesquisadores; (b) o aumento do uso de computadores interconectados e em alta velocidade, aplicando arquitetura em GRID[25]; (c) a visualização de dados; (d) o desenvolvimento de ferramentas e procedimentos baseados na internet; (e) a construção de estruturas organizacionais virtuais para a realização de pesquisas; (f) a distribuição eletrônica e a publicação dos resultados (JANKOWSKI, 2007).

Institucionalmente, a *e-Science* nasce com a criação de escritórios nacionais de pesquisa no Reino Unido e nos Estados Unidos. Essas nações mostram-se preocupadas com a curadoria e a análise do grande volume e variedade de dados que são gerados durante o processo de se fazer ciência. O termo *e-Science* foi utilizado pela primeira vez na Europa, em 1999, pelo diretor do Escritório de Ciência e Tecnologia do Reino Unido, John Taylor[26], o qual cunhou o termo referindo-se ao uso da tecnologia para realização de investigação científica em escala global (HEY; TANSLEY; TOLLE, 2009).

---

[24] *"[...] e-Science, is one of many terms used to describe recent transformations in the scientic enterprise"* (JANKOWSKI, 2007, p. 549).

[25] Um tipo de sistema distribuído e paralelo que possibilita o compartilhamento, seleção e agregação dinamicamente, em tempo de execução, de recursos autônomos geograficamente distribuídos, de acordo com a sua disponibilidade, capacidade, performance, custo e requerimentos,do usuário, de qualidade de serviço.

[26] *"In the UK, at the end of the 1990s, John Taylor became Director General of Research Councils at the Office of Science and Technology (OST) in the UK – roughly equivalente to Director of the National Science Foundation (NSF) in the USA. Taylor had been Director of Hewlett-Packard (HP) Laboratories in Europe and HP's vision for the future of computing has long been that IT resources will become a new "utility". Rather than purchase IT infrastructure, users will pay for IT services as they consume them, in the same way as the conventional utilities such as electricity, gas and water – and now mobile telephones. In his role at the OST as overseeing the funding of UK scientific research, Taylor realized that many areas of science could benefit from a common IT infrastructure to support multi-disciplinary and distributed collaborations. He articulated a vision for this type of distributed, collaborative science and introduced the term "e-Science"* (HEY, HEY, 2006, p. 517).

O National e-Science Centre, no Reino Unido, define-a como:

> Ciência em larga escala, que será cada vez mais realizada por meio de colaborações globais distribuídas pela Internet. Normalmente, um aspecto de tais colaborações são empreendimentos científicos que exigem acesso a coleções *big data*, a recursos de computação em grande escala e a visualização de alta performance, tendo os cientistas como usuários individuais (PIORUN, 2013, p. 23, tradução nossa)[27].

A *e-Science* é a colaboração global entre áreas-chave da ciência e também é a infraestrutura que permite essa colaboração. Nesse sentido, a *e-Science* não é apenas uma nova disciplina científica, mas é o conjunto de ferramentas e tecnologias necessárias para apoiar a ciência em rede e a sua infrastructura. Um componente fundamental da ciberinfrastructura de *e-Science* é permitir aos pesquisadores configurar as suas próprias organizações virtuais de pesquisa de maneira segura e interconectada, ligando centros de pesquisa e investigadores em todo o planeta (NATIONAL SCIENCE BOARD, 2005; HEY; HEY, 2006).

A *cyberinfrastructure* é o sinônimo norte-americano para o termo britânico *e-infrastructure* e está contida nas iniciativas de *cyberscience* do National Science Foundation (NSF), nos Estados Unidos. A *e-Science* destina-se a capacitar os cientistas a fazerem suas pesquisas de modo mais eficiente e colaborativo, utilizando-se, para isso, da *Internet* e do computador (NATIONAL SCIENCE BOARD, 2005; HEY, HEY, 2006).

Conforme explica Jankowski (2007), *cyberscience* e *cyberinfrastructure* possuem significado semelhante e referem-se respectivamente a:

> (a) *cyberscience:* são todas as atividades de investigação acadêmicas e científicas no espaço virtual gerado pelos computadores em rede e pelas tecnologias de informação e comunicação em geral; (b) *cyberinfrastructure*: refere-se à

[27] "*Large-scale science that will increasingly be carried out through distributed global collaborations enabled by the Internet. Typically, a feature of such collaborative scientific enterprises is that they will require access to very large data collections, very large scale computing resources and high performance visualization back to the individual user scientists"* (PIORUN, 2013, p. 23).

> infraestrutura distribuída de computadores, informações e tecnologias de comunicação. (JANKOWSKI, 2007, p. 552, tradução nossa)[28].

O advento da *e-Science* na Europa e nos USA move-se no sentido de desenvolver uma infraestrutura tecnológica poderosa, visando apoiar o trabalho colaborativo e multidisciplinar das ciências. A *e-infrastructure*, no Reino Unido, ou *cyberinfrastructure*, nos USA, geralmente está relacionada à computação científica, aos sistemas de armazenamento, aos *softwares,* aos serviços de dados, às informações baseadas em evidências, ao conhecimento publicado, às práticas sociais, aos sistemas de comunicação e às políticas institucionais de fomento à pesquisa científica dessas nações (HEY; TANSLEY; TOLLE, 2009).

A *cyberinfrastructure* é mais do que qualquer *hardware* físico utilizado em infraestruturas de computação ou ferramentas de *softwares* desenvolvidas para um determinado projeto, não está limitada a uma determinada disciplina. A *cyberinfrastructure* funciona como núcleo de construção de pesquisas relacionadas ao uso intensivo de dados (NATIONAL SCIENCE BOARD, 2005).

O American Council of Learned Societies relata que transformações ocorridas no campo das Ciências se devem, em sua maioria, aos avanços tecnológicos. A NSF nos USA define *cyberinfrastructure* como:

> [...] o termo *cyberinfrastructure* denota a camada de informações, conhecimentos, padrões, políticas, ferramentas e serviços que são compartilhados amplamente pelas comunidades de pesquisa, mas desenvolvidos para fins acadêmicos específicos: a *cyberinfrastructure* é algo mais específico do que a própria rede, é algo mais geral do que uma ferramenta ou um recurso desenvolvido para um projeto específico [...] Inclui diversas coleções de fontes primárias, como bibliotecas, arquivos e museus; bibliografias, auxiliares de pesquisa, sistemas de citações e concordâncias que tornam essa informação

---

[28] *"[...] (a) cyberscience: is all scholarly and scientific research activities in the virtual space generated by the networked computers and by advanced information and communication technologies in general; (b) cyberinfrastructure: refers to a n infrastructure of distributed computer, information, and communication technologies"* (JANKOWSKI, 2007, p. 552).

> recuperável; os padrões incorporados nos sistemas de catalogação e classificação; as revistas e a imprensa universitária que distribuem a informação; os editores, os bibliotecários, os arquivistas e os curadores que vinculam a operação dessa estrutura com os estudiosos que a utilizam (AMERICAN COUNCIL OF LEARNED SOCIETIES, 2006, p. 1-6, tradução nossa)[29].

As investigações no campo da *e-Science* foram estimuladas junto às áreas de ciências exatas (Astrofísica, Genética, Oceanografia, Climatologia, Física Nuclear), sendo que a principal preocupação dessas áreas foi com o imenso volume de dados gerados durante a pesquisa científica. A *e-Science* assume a transformação das ciências em direção ao uso intensivo de grandes volumes de dados.

> O projeto Pan-STARRS vai capturar 2,5 *petabytes* de dados a cada ano, quando entrar em operação (http://pan-starrs.ifa.haway.edu). O grande Colisor de Hádrons (GCH) do CERN vai gerar de 50 a 100 PB de dados por ano, com cerca de 20 PB armazenados e processados numa federação mundial de matrizes nacionais interligando 100 mil CPUs [...] Cientistas e instituições precisam de um modelo e um conjunto de melhores práticas para levar arquiteturas de hardware equilibradas e ao correspondente software capazes de lidar com esses volumes de dados [...] (FAY, 2009, p. 9).

Assim, percebe-se que, além da preocupação com o imenso volume de dados, de forma geral, as características fundamentais da *e-Science* são a digitalidade-aumentada*, networked* e a orientação a dados:

---

[29] *"[...] the term cyberinfrastructure is meant to denote the layer of information, expertise, standards, policies, tools, and services that are shared broadly across communities of inquiry but developed for specific scholarly purposes: cyberinfrastructure is something more specific than the network itself, but it is something more general than a tool or a resource developed for a particular project, [...] [...]It includes diverse collections of primary sources in libraries, archives, and museums; the bibliographies, searching aids, citation systems, and concordances that make that information retrievable; the standards that are embodied in cataloging and classification systems; the journals and university presses that distribute the information; and the editors, librarians, archivists, and curators who link the operation of this structure to the scholars who use it [...]"* (AMERICAN COUNCIL OF LEARNED SOCIETIES, 2006, p. 1-6).

> A porção da **digitalidade-aumentada** que engloba a *e-Science* compreende as habilidades de como os e-cientistas analisam, gerenciam, acessam e compartilham dados digitalmente. [...] exclusivamente dados criados por meio do uso de bancos de dados, redes, repositórios digitais, arquivos digitais e conjuntos de dados acessíveis *online* e digitalmente. A ***Networked*** é uma característica da *e-Science*, que compreende a natureza multidisciplinar do campo que cria inicialmente ligações não intencionais entre conjuntos de dados. Os e-cientistas utilizam dados e materiais em rede para formular novas informações por meio de comparação, cruzamento e manipulação de dados. Isso cria outras formas de comunicação científica, uma forma atrelada a como a rede se constrói a partir do uso de dados; como na forma em que os artigos se conectam através de citações. Portanto, o *e-Science* prospera quando os conjuntos de dados são compartilhados e acessíveis. Uma última característica [...] da *e-Science* é **a orientação a dados**. A exploração da *e-Science* é conseguida com a manipulação de dados e, portanto, os conjuntos de dados são usados como uma forma primária de experimentação. Através da manipulação e da comparação cruzada de conjuntos de dados, os pesquisadores são capazes de encontrar padrões e desenvolver novos inquéritos em todas as disciplinas (SHCMILLEN, 2015, p. 4, tradução e grifo nosso)[30].

[30] "*The **digitally-enhanced** ration of eScience embodies how eScientists analyze, manage, and gain access to, and share the data— digitally. Without the technological advances of our time the amount of data we have access to and are able to create would significantly decrease. Almost exclusively through online databases, networks, shared digital repositories, and digital files, datasets are shared and therefore digitally accessible. [...] **Networked** is a very important characteristic of eScience because this is what makes eScience so fascinating, the networked and multidisciplinary nature of the field that creates initially, unintended links among datasets. eScientists utilize networked data and material to formulate new information through cross-comparison and manipulation. This creates another form of scholarly communication that is networked and builds from one use to another; like how articles are connected through citations and websites. Therefore, eScience thrives when datasets are shared and accessible. The final characteristic [...] of **eScience is data-driven**. eScience exploration is achieved through data manipulation thus datasets are used as the primary form of experimentation. By*

Essas três características da *e-Science* transformam o modo de se fazer ciência.

### 2.2.1 Ciência orientada aos dados*:* o quarto paradigma

Os modelos tradicionais de investigação científica transformam-se com a *e-Science*. A ciência com uso intensivo de dados consiste em três atividades básicas: captura, curadoria e análise de grandes volumes de dados. Os métodos convencionais compreendem a observação/exploração empírica, a teorização e a simulação (HEY; TANSLEY; TOLLE, 2009).

> A *e-Science* [...] tem sido reconhecida como o quarto paradigma de exploração da ciência. [...] a *e-Science* refere-se à digitalidade aumentada, ao *networked* e à ciência orientada aos dados [...] A pesquisa em *e-Science* requer uma variedade de recursos tecnológicos para o gerenciamento de dados e serviços ao cliente para garantir a estabilidade, a integridade e a colaboração científica (SHCMILLEN, 2015, p. 3-6, tradução nossa)[31].

Estudos recentes sobre o uso intensivo de dados amplificam-se quando são mediados pelo computador, principalmente se incluírem redes como a *web*. Isso ocasiona o surgimento de diferentes métodos aplicados para coleta e/ou uso de dados, ou seja, um novo modo de fazer ciência, baseado nos instrumentos e nos dispositivos técnicos criados pela computação contemporânea, os quais têm como base o uso intensivo de dados.

A ciência com uso intensivo de dados institui uma nova dimensão para os cientistas, o universo dos dados, causando uma revolução no processo de pensamento científico. Samuel Thomas Khun (2000, p. 3) considera que "[…] um paradigma são as realizações científicas universalmente reconhecidas que, durante algum tempo, fornecem

---

*manipulating and cross-comparing datasets, researchers are able to find patterns and develop new inquiries across disciplines* (SHCMILLEN, 2015, p. 4).

[31] *"e-Science is a term that encompasses many variations of the science fields and has been recognized as the fourth paradigm of science exploration. [...] e-Science refers to a digitally-enhanced, networked, data-driven science [...] e-Science research requires a variety of technological, data management, and customer services-related features to ensure stability, scientific integrity, and collaboration* (SHCMILLEN, 2015, p. 3-6).

problemas e soluções modelares para comunidade de praticantes de uma ciência". Inspirado na ideia de paradigma, Jim Gray (2007) afirma que a *e-Science* compreende a evolução da ciência empírica para teórica, de teórica para simulação e da simulação para a ciência com uso intensivo de dados.

> […] as tecnologias criaram novas possibilidades (ou tipos) de pesquisa. […] dados experimentais, coletados por meio de instrumentos ou gerados por simulação, são processados por sistemas de *software* complexos, e só então a informação (ou o conhecimento) resultante é armazenada em computadores. Os cientistas só analisam os dados no final do processo. Trata-se de uma mudança importante no processo de pensamento científico, que está substituindo o processo de **"formulação de hipótese >> experimentação >> análise de resultados"** por **"formulação de hipótese >> busca da resposta no banco de dados"** […] (CORDEIRO et. al, 2013, p. 76, grifo nosso).

A ideia de quarto paradigma foi apresentada por Jim Gray em uma palestra ministrada na Califórnia em 2007 para o Conselho Nacional de Pesquisas dos Estados Unidos. Nela, ele busca discorrer sobre os desafios de uma nova forma de se fazer ciência:

> […] a ciência nasceu há milhares de anos de forma *empírica*, descrevendo fenômenos naturais. Nos últimos séculos, ela passou a incorporar uma importante componente *teórica*, utilizando modelos e generalizações. Nas últimas décadas, surgiu uma forte tendência ligada às tecnologias computacionais, o que permite a realização de experiências sofisticadas envolvendo simulações de fenômenos complexos. Nos últimos anos, observa-se o aparecimento de um *quarto paradigma*, a exploração de grandes quantidades de dados, muitas vezes chamado de *e-Science* (e-ciência), que unifica teoria, experimentos e simulação, ao mesmo tempo em que lida com uma quantidade enorme de informação (CORDEIRO et. al, 2013, p. 73).

Jim Gray, *fellows*[32] da Microsoft, amplifica a ideia de *e-Science*, pois sugere o uso do termo como sinônimo de um quarto paradigma das ciências. Acredita-se que as tecnologias digitais revolucionaram os métodos científicos, o que é fundamentado por Hey e Hey (2006, p. 515) ao comentarem que "[...] as ciências da computação fizeram emergir novos paradigmas científicos que modificaram a maneira de ser da pesquisa científica".

> […] a computação está deixando de ser apenas uma ferramenta de apoio a novas pesquisas para se tornar parte fundamental das ciências com que interage e de seus métodos científicos. A sinergia entre Ciência da Computação e as outras áreas do conhecimento criou um novo modo de se fazer ciência […] (CORDEIRO et. al, 2013, p. 72).

São exemplos dessa transformação da computação as observações feitas por áreas como Meteorologia, Medicina e Biologia. Essas áreas utilizam dispositivos técnicos (radares, telescópios, sensores e scanners) que captam as mutações do mundo em tempo real, gerando ambientes para simulação e redução do mundo. Essa prerrogativa é um fundamento básico para análise de um emaranhado de dados híbridos que ajudam a visualizar e medir a temperatura, o clima, os batimentos cardíacos, as vontades de consumo, entre outras coisas que podem ser modeladas via algoritmo computacional e *hardware*.

Delimitada a origem, o conceito e as características fundamentais da *e-Science*, percebe-se que ela é um campo de conhecimentos multidisciplinares e que possui um amplo espectro de possibilidades de ser estudada. Sua principal característica é ser orientada a dados produzidos digitalmente e em rede.

Por ora, entende-se a *e-Science* como um conceito agregado à ideia de informação, tecnologia e dados, sendo utilizada para representar coisas, ações, métodos e sistemas sob a dicotomia do *hardware* e do *software* em âmbito científico. Acredita-se que a *e-Science* esteja preocupada com a natureza e com a significância do uso das tecnologias de dados em relação às práticas e às teorias científicas contemporâneas, aplicadas pelas mais diversas áreas do conhecimento para promover a

---

[32] São pesquisadores de reconhecida capacidade de inovação, contratados por universidades ou centros de pesquisa exclusivamente para pesquisar e propor novas ideias (BELL, 2009). Não há cargo similar na academia brasileira. A posição de *fellows* na área técnica de um centro de pesquisa está entre as do topo de carreira de alguns docentes.

colaboração científica em escala global por meio do uso intensivo de dados.

### 2.2.2 Ciência da Informação em contextos de *e-Science*: imbricações

A pesquisa sobre Ciência da Informação em contextos de *e-Science* geralmente se desenvolve ligada à criação de produtos e/ou serviços de dados de pesquisa implantados nos USA e no Reino Unido. Nesses países, as bibliotecas universitárias empenham-se para gerar produtos e/ou serviços focados em sua comunidade de investigadores, especialmente no que tange à seleção de recursos, à criação de metadados, à preservação digital, à organização e à gestão de dados de pesquisa (PIORUN, 2013).

Em 2006 a Association for Research Libraries (ARL) criou um grupo de trabalho que dedica suas atividades investigativas para identificar os desenvolvimentos de produtos e/ou serviços de *e-Science* ligados ao contexto das bibliotecas. O objetivo era compreender como as bibliotecas podem contribuir para as atividades de *e-Science.*

Nesse sentido, o grupo sustenta que as bibliotecas universitárias têm papel fundamental na gestão e na preservação de conjuntos de dados gerados durante os procedimentos de pesquisa. Para o grupo de trabalho, a *e-Science* pode ser definida como o domínio da ciência que se preocupa com os novos métodos baseados em computação intensiva com base em dados digitais. Por exemplo, um relatório do grupo da ARL mostra que algumas bibliotecas universitárias, nos USA, desenvolvem serviços de referência com base no acesso a dados de pesquisa e ofertam consultoria a pesquisadores que se preocupam em preservar e compartilhar dados brutos de suas pesquisas e repositórios digitais. Essas bibliotecas estão envolvidas em projetos de investigação, elaboração e gestão de dados de pesquisa e estão proporcionando espaço de trabalho aos investigadores (física e virtualmente). O grupo da ARL também identificou que existem parcerias para criação e gestão de serviços de preservação, geração de políticas, criação de programas para o acesso aberto à informação científica e desenvolvimento de *softwares* centrados no utilizador e orientados à mineração e à descoberta de dados de pesquisa (SOEHNER; STEEVES; WARD, 2010; PIORUN, 2013).

No entanto, é possível explanar que as pesquisas sobre CI em contextos de *e-Science* no âmbito mundial:

> […] encontram-se em estágio inicial, sendo que no Brasil, especificamente, encontra-se numa fase

> embrionária. Pesquisas precisam ser desenvolvidas para identificar como os dados produzidos em estudos conduzidos nas universidades federais, nos institutos federais, centros de pesquisa ou laboratórios de pesquisa estão sendo tratados. O bibliotecário, principalmente aquele que exerce suas atividades em universidades ou instituições científicas e tecnológicas, precisa se perguntar: como organizar esses dados produzidos nas universidades e centros de pesquisa? Quais serão as habilidades requeridas para o profissional da informação que deseja trabalhar com organização de dados de pesquisa (COSTA; CUNHA, 2014, p. 203).

O conceito de dados aponta uma forte tendência de estudos e práticas profissionais que está se desenvolvendo no cerne das investigações realizadas por bibliotecários e cientistas da informação. Assim, pode-se afirmar que as atuais tendências de pesquisa em Ciência da Informação em contextos de *e-Science* revelam estudos sobre a proveniência, as melhores práticas, o planejamento, a curadoria entre outras atividades científicas relacionadas à teoria e à *práxis* com dados de pesquisa, principalmente quando esses dados se relacionam com os processos de pesquisa científica em meio digital. Basicamente, o campo da *e-Science* está relacionado com a coleta, a curadoria, o processamento, o armazenamento, a análise e a visualização de dados de pesquisa. A maioria das investigações sobre a *e-Science* está orientada na prática com o uso e com o consumo de dados.

Por fim, cita-se outro relatório publicado em 2009 pelo governo dos USA. Produzido pelo Interagency Working Group on Digital Data do National Science and Technology Council (NSTC), o texto configura-se como uma estratégia para promover a preservação e o acesso a dados de pesquisa digitais. Intitulado *Harnessing the Power of Digital Data for Science and Society*, o trabalho inclui o primeiro conjunto de princípios para orientar estratégias do governo federal dos Estados Unidos para o uso de dados em projetos de *e-Science*. O relatório prevê o surgimento de um universo de dados de pesquisa em formato digital no qual o governo tem papel crucial na criação, na coleta, na documentação, na análise, na preservação e na divulgação de dados. Tal situação sugere o investimento em investigação sobre dados digitais, como pesquisas catalisadoras para o progresso da *e-Science* nos USA.

O relatório parte dos pressupostos de que as características-chave para o desenvolvimento de pesquisas com dados é:

> [...] os produtos da ciência e as novas pesquisas surgem em ambientes digitais *born-digital*; os volumes da demanda para o uso de dados são movidos pelo ritmo das inovações tecnológicas digitais; todos os setores da sociedade são partes interessadas na preservação digital de dados; [...] as ciências são globais e prosperam cada vez mais nas dimensões digitais; os dados de pesquisa são bens nacionais e globais; nem todos os dados científicos de pesquisa precisam ser preservados e nem todos os dados preservados precisam ser preservados indefinidamente; [...] a preservação de dados de pesquisa em meio digital é de responsabilidade tanto do governo quanto do setor privado, pois traz benefícios à sociedade como um todo; a preservação de dados de pesquisa deve ser uma preocupação em longo prazo e deve incluir problemas relacionados ao acesso, à interoperabilidade e ao gerenciamento completo do ciclo de vida dos dados; e estratégias do ciclo de pesquisa (INTERAGENCY WORKING GROUP ON DIGITAL DATA, 2009, p. 9, tradução nossa)[33].

Adquirir *know-how* sobre o uso de dados é uma das principais tarefas da CI. Dessa forma, conclui-se que os estudos sobre dados são uma das principais tendências de pesquisa em CI. Então, compreender a CI no contexto da *e-Science* exige conhecimentos sobre o conceito de dados. Dessa forma, torna-se necessário o desenvolvimento de uma

---

[33] *"[...] the products of science and the starting point for new research are increasingly digital and increasingly "born-digital"; exploding volumes and rising demand for data use are driven by the rapid pace of digital technology innovations; all sectors of society are stakeholders in digital preservation and access; and a comprehensive framework for cooperation and coordination to manage the risks to preservation of digital data is missing.[...] science is global and thrives in the digital dimensions; digital scientific data are national and global assets; not all digital scientific data need to be preserved and not all preserved data need to be preserved indefinitely; [communities of practice are an essential feature of the digital landscape;] preservation of digital scientific data is both a government and private sector responsibility and benefits society as a whole;long-term preservation, access, and interoperability require management of the full data life cycle; [and dynamic strategies are required]* (INTERAGENCY WORKING GROUP ON DIGITAL DATA, 2009, p. 9).

introdução ao conceito de dados de pesquisa e as suas tipologias relacionadas às diferentes disciplinas científicas.

## 2.3 O CONCEITO DE DADOS

A palavra dados (*data*)[34] surge na metade do século XVII; plural do latim *datum*, significa uma única peça de informação. O *Oxford English Dictionary on-line*[35] define dados como fatos e/ou estatísticas reunidas para referência ou análise de algo, as quantidades (informação estatística), os caracteres e os símbolos de operações lógicas executadas por um computador (*software*). Também é o registro dessas operações armazenado, preservado e transmitido sob forma de sinal elétrico (digital) que possa ser gravado em qualquer tipo de mídia magnética, óptica ou mecânica (*hardware*). A palavra dados se aplica a uma série de objetos que são originários de ambientes digitais, como textos, números, imagens, vídeos, áudios, *softwares*, algoritmos, equações, animações, modelos e simulações. (NATIONAL SCIENCE BOARD, 2011; KOLTAY, 2017; OED, 2015).

Nesse sentido, os dados são peças de informação digital distintas, geralmente formatadas de maneira particular. Os dados digitais são considerados um tipo de informação de natureza digital. Os dados podem ser considerados:

> [...] representações formais da natureza em sistemas computacionais [...] Os dados podem ser símbolos e representações de informação e conhecimento representáveis [em formato digital]; no entanto, eles não devem ser equivalentes a informações ou a conhecimentos (ZHU; XIONG, 2015, p. 3, tradução nossa)[36].

---

[34] Em latim, *data* é o plural de *datum*. Historicamente e em áreas científicas especializadas, o termo *data*, em inglês, é tratado como plural, tendo um verbo correspondente no plural, como na frase: "dados foram coletados e classificados". Em se tratando, no entanto, de uso moderno não científico, *data* geralmente não é empregado como um termo plural. Em vez disso, é tratado como um substantivo incontável (*mass noun* ou *uncountable noun*), semelhante a uma palavra como informação, levando um verbo correspondente no singular (SEMIDÃO, 2014, p. 71).

[35] Disponível em: <https://en.oxforddictionaries.com/definition/data>. Acesso em: maio 2016.

[36] *"Data is the formal representation of nature in computer systems; information is the phenomena of nature, society, and thinking activities; and knowledge is experience gained through practice. Data can be regarded as symbols and representations of information and knowledge; however, they should not be equivalent to information and knowledge"* (ZHU; XIONG, 2015, p. 3).

Em decorrência do desenvolvimento das tecnologias digitais, as entidades do mundo natural são cada vez mais armazenadas na forma de dados e compartilhadas em redes como a *Internet*. Os dados são inseridos, gerados e criados no computador, tornam-se cada vez mais diversos, complexos e fora do controle humano (ZHU; XIONG, 2015).

Os dados são tão complexos e fluidos que virtualmente todos os tipos de informação digital podem ser algum tipo específico de dados. É possível se referir a tudo em meio digital como dados (MANTRA; 2012-13; RICE; SOUTHALL, 2016).

Para os fins deste estudo, os dados podem e serão considerados como todo objeto criado em formato digital (*digital-born*) ou convertido para formato digital (digitalizados) que possa ser usado para geração de *insights* de conhecimento.

Os dados podem ser primários (aqueles dados que não são manipulados ou alterados por pesquisadores), uma vez manipulados por pesquisadores, ou seja, depois de filtrados, analisados ou organizados, os dados primários tornam-se dados secundários ou transformados (NATIONAL RESEARCH COUNCIL, 1999; PIORUN, 2013).

No caso da pesquisa científica, os dados são potencialmente utilizados como recurso primário para geração de resultados de pesquisa. Assim, percebe-se que o conceito atribuído ao termo dados depende do contexto em que ele é produzido e/ou em que fase do ciclo de vida dos dados[37] a informação e/ou conhecimento estarão presentes na pesquisa (MANTRA, 2012-13).

### 2.3.1 Dados de pesquisa

O termo dados de pesquisa é formalmente definido, em âmbito internacional, em 2007, pela Organization Economic Co-operation and Development (OECD), por meio da declaração de acesso aos dados científicos com financiamento. Segundo National Science Board (2011), Rodrigues et. al. (2010), Panpel et. al (2013), Organization for Economic Co-Operation and Development (2004), Tibbo e Jones (2017) e Costa e Braga (2016), entre outros autores que tratam da temática, dados de pesquisa são "[...] registros factuais usados como fontes

[37] *"The research data lifecycle can be depicted in multiple different ways from simple to complex. It provides a useful way to think through the various different stages that data go through during a research project. Key phases of the research data lifecycle created by the Data Documentation Initiative include: Discovery and Planning, Initial Data Collection, Data Preparation and Analysis, Publication and Sharing, and Long-term Management"* (TIBBO; JONES 2017, on-line).

primárias na investigação científica, e que são geralmente aceitos pela comunidade científica como necessários para validar os resultados de investigação"[38]. Para o National Science Board (NSF) (2011), essa definição inclui dados analisados e os metadados que descrevem como esses dados foram gerados. Os dados analisados são limitados a informações digitais que descrevem os resultados de pesquisa financiada, incluindo imagens digitais, tabelas publicadas e tabelas usadas para criação de visualizações gráficas (PANPEL et. al., 2013; NATIONAL SCIENCE BOARD, 2011).

Outra definição de dados de pesquisa inclui a ideia de que esses dados compreendem:

> Informação registrada necessária para apoiar ou validar as observações, descobertas ou resultados de um projeto de pesquisa. […] o que é coletado, observado ou criado de forma digital, para fins de análise e para produzir resultados de pesquisas originais […] dados de pesquisa são: dados, registros, arquivos ou outras evidências, independentemente do seu conteúdo ou forma (por exemplo, material impresso ou digital), que compreendem observações de pesquisa, achados ou resultados, incluindo materiais primários e dados já analisados (RICE; SOUTHALL, 2016, p. 20, tradução nossa)[39].

Dados de pesquisa são evidências e/ou insumos de pesquisa que são coletados, observados, registrados e/ou criados para fins de análise e que podem produzir resultados de pesquisa para um estudo científico. Os dados de pesquisa são um produto para pesquisa, bem como um insumo para a pesquisa. Os dados de pesquisa devem poder ser coletados sistematicamente, estruturados e documentados para servir

---

[38] *"[...] the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues"*

[39] *"The recorded information necessary to support or validate a research project's observations, findings or outputs' […] 'That which is collected, observed, or created in a digital form, for purposes of analysing to produce original research results' […] 'Research data: The data, records, files or other evidence, irrespective of their content or form (e.g. in print, digital, physical or other forms), that comprise research observations, findings or outcomes, including primary materials and analysed"* (RICE; SOUTHALL, 2016, p. 20).

como insumos para futuras pesquisas. Os dados da pesquisa devem ser validados e compartilhados entre a comunidade científica. Dados de pesquisa são situacionais, da mesma forma que podem ser considerados válidos para responder a algumas questões de pesquisa, podem não ser. A mesma informação pode ser dados de pesquisa para um pesquisador em um ponto do tempo e em outro ponto do tempo não, dependendo da forma como esse pesquisador usa a informação ou material para análise de pesquisa (KELLAM; THOMPSON, 2016; HENDERSON, 2017; TIBBO; JONES, 2017; PRYOR, 2012; MANTRA, 2012-13).

Como exemplo, Tibbo e Jones (2017) citam definições de dados de pesquisa fornecidos e utilizados por instituições que fomentam pesquisas científicas nos USA, como a National Institutes of Health (NIH), que vê os dados de pesquisa como material fatual registrado e aceito comumente por uma comunidade científica como evidência para documentar e apoiar resultados de pesquisa (NATIONAL INSTITUTES OF HEALTH, 2003). Outra instituição nos USA, a NSF descreve dados de pesquisa como algo determinado por uma comunidade de interesse por meio do processo de revisão por pares e programas de gerenciamento (NATIONAL SCIENCE BOARD, 2011). Por fim, Tibbo e Jones (2017), apresentam a definição da National Endowment for the Humanities que define dados como materiais gerados ou coletados durante a realização de pesquisas científicas. (TIBBO; JONES 2017; NATIONAL INSTITUTES OF HEALTH, 2003; NATIONAL SCIENCE BOARD, 2011).

Os dados de pesquisa podem diferir em suas tipologias. Dependendo do assunto ou da disciplina científica, as definições utilizadas podem incluir uma vasta tipologia de objetos digitais e não digitais, como: documentos (texto, palavras), memorandos, notas, avaliações, registros de casos, protocolos de estudo, planilhas, notas de laboratório, notas de campo, diários, questionários, transcrições, filmes, fitas de áudio e vídeo, arquivos digitais de áudio e vídeo, proteínas, sequências genéticas, DNA, cristalografia, coordenadas geográficas, *slides*, artefatos, espécimes, organismos, coleções de objetos digitais adquiridos e/ou gerados durante o processo de pesquisa, conteúdo em banco de dados (vídeo, áudio, texto, imagens), modelos, algoritmos, *scripts* de uma aplicação de computador (entrada, saída, arquivos de *log* para *software* de análise, *software* de simulação, esquemas), metodologias, fluxos de trabalho, informações técnicas, procedimentos e protocolos (DUDZIAK, 2016; HENDERSON, 2017).

Dados de pesquisa podem ser caracterizados de várias formas, conforme sua natureza, origem ou de acordo com seu status no fluxo de

trabalho da investigação científica. Segundo o *Guia de Gestão de Dados de Pesquisa* elaborado pela Comissão de Gestão de Energia Nuclear do governo brasileiro, esses dados são categorizados:

> [...] segundo sua natureza: (a) observacionais: gerados por meio de observações diretas que podem ser associadas a lugares e/ou tempo específico, por sua natureza instantânea qualificam-se como registros históricos, pois não podem ser coletados uma segunda vez e, portanto, devem ser submetidos a processos de curadoria que os preservem para sempre; (b) computacionais, são o resultado da execução de modelos computacionais ou de simulações; (c) experimentais, são provenientes de situações controladas em laboratórios. [...] podem ser números, imagens, vídeos, áudio, *software*, algoritmos, equações, animações, modelos e simulações. Por fim, conforme o fluxo da pesquisa, (a) dados brutos, coletados diretamente de instrumentos científicos de pesquisa; (b) derivados, resultados do processamento ou combinação de dados brutos ou de outros dados; (c) canônicos ou referenciais, coleções de dados consolidados e arquivados geralmente em grandes centros de dados (SAYÃO; SALES, 2015, p. 8-9).

Os dados de pesquisa definem-se conforme sua natureza digital, o que inclui diferentes tipos de dados, níveis de agregação e métodos de investigação científica. O Interagency Working Group on Digital Data do National Science and Technology Council nos USA define dados de pesquisa em formato digital como:

> [...] dados para uso e/ou reutilização em pesquisas científicas, técnica e para aplicações educacionais. [...] referem-se a toda a gama de tipos de dados e formatos relevantes para pesquisa científica, de engenharia, de educação e de [humanidades], em contextos locais, regionais, nacionais e globais com a amplitude correspondente às aplicações e a usos científicos em potencial. [...] dados de pesquisa digitais são quaisquer dados digitais, bem como os métodos e as técnicas utilizados na

> criação e na análise desses dados, [usados para] verificar resultados ou ampliar conclusões científicas, incluindo dados digitais associados a informações não digitais, como metadados associados a amostras físicas (NATIONAL SCIENCE BOARD, 2011, p. 1-33, tradução nossa)[40].

Os dados de pesquisa digitais proliferaram devido aos avanços impulsionados pelas tecnologias da informação e da computação, sendo o uso crescente de tecnologias uma das principais fontes geradoras de dados. Atualmente, os dados de pesquisa são o resultado dos mais diversos processos de investigação mediada por dispositivos técnicos, o que transforma todas as etapas da pesquisa, desde a geração de dados, que pode se dar por meio de um experimento, como um estudo empírico sobre a observação de fenômenos culturais, até a publicação de resultados de pesquisa em um repositório de dados de pesquisa (PAMPEL, 2013).

Assim, percebe-se que os dados de pesquisa irão diferir de acordo com os métodos científicos e que eles dependem de características específicas de cada disciplina científica. Nesse sentido, o próximo tópico tentará delimitar a definição de dados de pesquisa segundo sua utilização pelas diferentes disciplinas científicas.

### 2.3.2 Dados de pesquisa e as diferentes disciplinas

A definição de dados de pesquisa transforma-se em relação às diferentes disciplinas do conhecimento científico. De disciplina a disciplina, de pesquisador a pesquisador, a expressão dados de pesquisa pode significar coisas diferentes e depende do contexto científico em que será utilizada. Os dados de pesquisa são gerados com diferentes propósitos, o que amplia a sua diversidade. Diferentes disciplinas têm e usam linguagem disciplinar específica em torno dos dados de pesquisa (NATIONAL RESEARCH COUNCIL UK, 1999; PIORUN, 2013;

---

[40] *"[...] for use or repurposing for scientific or technical research and educational applications [...] It refers to the full range of data types and formats relevant to all aspects of science and engineering research and education in local, regional, national, and global contexts with the corresponding breadth of potential scientific applications and uses [...]. Digital research data is any digital data, as well as the methods and techniques used in the creation and analysis of that data, that a researcher needs to verify results or extend scientific conclusions, including digital data associated with non-digital information, such as the metadata associated with physical samples"* (NATIONAL SCIENCE BOARD, 2011, p. 1-33).

SAYÃO; SALES, 2015; KELLAM, THOMPSON, 2016; KOLTAY, 2017; HENDERSON, 2017).

Por exemplo, um cientista social pode considerar dados os seus diários, fotografias, resultados de questionários de pesquisas, notas de pesquisa de campo e/ou planilhas. Dados de pesquisa em Ciência Social são criados conforme os objetivos da investigação. Não são espontâneos, são produzidos e muitas vezes se apresentam como recursos digitais na forma de informação estatística, como indicadores econômicos, políticos e sociais. Por exemplo, pode-se citar um questionário de pesquisa, que pode conter informações coletadas a respeito de opiniões, comportamentos e/ou características sociais de um grupo de indivíduos que compõem uma unidade de observação social. As unidades de observação são o *backbone* dos dados de pesquisa em Ciências Sociais e são compostas por fatores de agregação, como em representações temporais, unidades de observação, coordenadas geoespaciais, entre outras categorias de conteúdo social (GERACI; HUMPHREY; JACOBS, 2012).

Um cientista da computação pode utilizar o termo para referir-se ao fluxo de zeros e uns, que fluem em uma rede de computadores. Dados são vistos como valores armazenados em memória artificial, são objetos que podem ser processados e manipulados em um sistema operacional:

> Tudo em computação pode ser codificado em código binário, o que inclui sistemas operacionais e aplicativos de *software*, que são arquivos que contêm instruções para direcionar as operações da CPU, bem como as informações a serem manipuladas pela CPU. Os programas de computador são escritos em linguagens de programação de alto nível, ou seja, [...] são convertidos em instruções binárias na CPU por meio de um compilador que converte o código em uma massa de instruções na CPU [...] os cientistas da computação consideram a saída de um compilador, que normalmente é armazenado em um arquivo digital, como dados binários. Da mesma forma, a saída binária do código interpretado também é chamada de dados

> (GERACI; HUMPHREY; JACOBS, 2012, p. 19, tradução nossa).[41]

Scholtz (2016) apresenta os dados geoespaciais como uma subcategoria de dados de pesquisa que possuem peculiaridades tecnológicas e metodologias específicas no que tange a sua criação, utilização, preservação e compartilhamento. Esses dados contêm informação a respeito de localizações geográficas e coordenadas geométricas que descrevem e modelam as dimensões e as características do planeta terra. Segundo o *Guia de Pesquisa* da Universidade de York, no Canadá:

> Dados geoespaciais, possuem informações explícitas de posicionamento geográfico, o que inclui dados sobre as redes rodoviárias em um Sistema de Informação Geográfica ou uma imagem de satélite georreferenciada. Os dados geoespaciais podem incluir dados de atributos que descrevem os recursos encontrados em um conjunto de dados. Existem dois tipos de dados geoespaciais – *vetor* e *raster*. Os vetores usam objetos geométricos simples, como pontos, linhas e áreas (polígonos), para representar recursos espaciais. O *raster data* utiliza um *grid* para representar sua informação geográfica. [...] são representados em células individuais, em linhas e em sequências de células vizinhas por áreas de coleção agrupadas. Outros tipos de dados *raster* incluem: imagens de satélite – dados de satélite oriundos de deteção remota; modelos de elevação digital – uma série de dados de elevação uniformemente espaçados; arquivos gráficos – mapas digitalizados, fotografias e imagens em

[41] *"Everything in a computer is encoded in binary, including the operating system and application software, which are files containing instructions to drive the operations of the CPU as well as the information to be manipulated by the CPU. Computer programs written in a higher-level language, that is, a level other than the CPU's binary instruction set, are converted to binary instructions either through a compiler, which translates code en masse into the CPU's instruction set, or through an interpreter, which does the translation on the fly. Computer scientists regard the output of a compiler, which is typically stored in a file, as binary data. Similarly, the binary output of interpreted code is also called data"* (GERACI; HUMPHREY; JACOBS, 2012, p. 19).

> formato TIFF, GIF ou JPEG (YORK UNIVERSITY, 2017, on-line, tradução nossa)[42].

Por fim, dados científicos são dados de pesquisa utilizados e/ou produzidos por disciplinas das Ciências Exatas, como a Astronomia e a Biologia; caracterizam-se pelo seu grande volume e necessidade de alta infraestrutura tecnológica para armazenamento, preservação e compartilhamento.

> O Genoma Humano e as Pesquisas Astronômicas geram alguns dos maiores conjuntos de dados, atingem os petabytes (PB). O Sloan Digital Sky Survey, por exemplo, produz cerca de 5 PB por dia. Já o Large Synoptic Survey Telescope (LSST) grava até 30 trilhões de bytes de dados de imagem diariamente a partir do topo de uma montanha no Chile (GRIGG, 2016, p. 180, tradução nossa)[43].

Os dados científicos são descentralizados e gerados em diversos laboratórios e/ou centros de pesquisa por muitos pesquisadores em diferentes países do mundo. Os dados científicos são complexos no que tange a sua análise e manipulação, exigem computadores de alto desempenho e soluções de infraestrutura. Por exemplo, os conjuntos de dados sobre o Genoma Humano e sobre a Pesquisa Astronômica requerem uma enorme quantidade de espaço para armazenamento e

---

[42] *"Geospatial data, GIS data or geodata has explicit geographic positioning information included within it, such as a road network from a GIS, or a geo-referenced satellite image. Geospatial data may include attribute data that describes the features found in the dataset. There are two types of geospatial data* ***- vector and raster. Vector*** *data uses the simple geometric objects of points, lines, and areas (polygons) to represent spatial features.* ***Raster*** *data uses a grid to represent its geographic information. Points are represented by single cells, lines by sequences of neighboring cells, and areas by collection of grouping cells. Other types of raster data include: Satellite Imagery - remotely sensed satellite data. Digital Elevation Models- an array of uniformly spaced elevation data. Graphic Files - scanned maps, photographs, and images in TIFF, GIF, or JPEG format"* (YORK UNIVERSITY, 2017, online)**.**

[43] *"Genomics and astronomy generate some of the largest datasets, often reaching one or more petabytes (PB). The Sloan Digital Sky Survey, for example, produces about five PB a day, and the proposed Large Synoptic Survey Telescope (LSST) will record 30 trillion bytes of image data daily from the top of a mountain in Chile, a data volume equal to two daily Sloan Digital Sky Surveys"* (GRIGG, 2016, p. 180).

processamento computacional, além de serem dados de custo elevado (GRIGG, 2016).

Por fim, os dados de pesquisa são produtos de pesquisa heterogêneos e contextualizados conforme as disciplinas do conhecimento. Nesse contexto, os dados de pesquisa devem ser válidos, compartilháveis, heterogêneos e contextualizáveis dentro de uma comunidade científica. No que tange à Ciência da Informação, os dados de pesquisa são de origem multidisciplinar e dependem da relação que a disciplina possui com a CI. No que segue, apresenta-se a Biblioteconomia de Dados como forma de delimitar o tema dos dados de pesquisa em uma das principais disciplinas da CI.

### 2.3.3 Biblioteconomia de Dados

A Biblioteconomia orientada a dados, em inglês, pode ser chamada de *e-Science Librarianhip*[44] (XIA; WANG, 2014), *Data Librarianship* (KOLTAY, 2017; RICE, 2016; RICE; SOUTHALL, 2016)[45] e/ou *Databrarianship* (KELLAM; THOMPSON, 2016)[46]. Ela corresponde ao interesse de bibliotecários em compreender atividades ligadas ao gerenciamento e à curadoria de todos os tipos de dados, sendo seu foco o tratamento, a gestão e a curadoria de dados de pesquisas em qualquer disciplina científica. Ela busca prover técnicas e serviços para o suporte a coleções de dados de pesquisa, investigando problemas práticos relacionados ao armazenamento, ao descobrimento e ao acesso a dados de pesquisa. Atualmente, é uma tendência de estudos desenvolvida em bibliotecas acadêmicas, principalmente, nos USA, no UK e no Canadá.

As origens da Biblioteconomia de Dados estão no surgimento das

---

[44] *"Data librarianship in general, which is sometimes also called e-science librarianship, has now become a significant part of the academic library workforce"* (XIA WANG, 2014, p. 364).

[45] *"The origins of data librarianship and the title data librarian go back to social science data management (particularly publicly available data sets and geospatial data), and they then evolved in the bioinformatics field. The concept and the title now cover library-related work with both purchased and locally produced digital data in any subject domain"* (SOEHNER et al., 2010).

[46] *"[...] we use the term databrarianship to describe a field that is characterized by a diversity of interests but united by our commitment to ensuring access to data, be they primary research data created by our institutions' researchers or secondary data used for analysis"* (KELLAM; THOMPSON, 2016, p. 1).

coleções de dados de pesquisa oriundas da utilização de métodos quantitativos de pesquisa em Ciências Sociais, especialmente dados públicos produzidos por instituições governamentais, como os sensos eleitorais e os dados sobre a economia. O surgimento de arquivos e bibliotecas de dados no Reino Unido (Data Archives Services), nos Estados Unidos e no Canadá (Data Library Services), entre as décadas de 1960 e 1970, são fatos fundamentais que dão origem à Biblioteconomia de Dados. A forma como serviços de dados surgiram e se desenvolveram difere em relação aos continentes, Europa e a América do Norte. Na Europa, os serviços de dados foram concebidos de forma centralizada pelos conselhos nacionais de pesquisa, vide o Council of European Social Science Data Archives (CESSDA)[47]. Na América do Norte, os serviços de dados foram oferecidos por pequenos centros de dados e/ou computação de departamentos universitários nos USA e no Canadá (RICE; SOUTHALL, 2016).[48]

As Ciências Sociais foram um dos primeiros campos disciplinares a adotar e a compartilhar dados de pesquisa para o reuso, isso por razões de economia e eficiência científica. O desenvolvimento de coleções de dados em Ciências Sociais foi fortemente influenciado pelas ideias computacionais de Alan Turing (1930), pelos métodos de escalonamento e análises multivariadas utilizados pelo World Bank, International Monetary, na década de 1940; pelo surgimento das linguagens de programação COBOL e FORTRAN; e, respetivamente, no Institute of Social Research em Michigan e no Bureau of Applied Social Research, em Columbia, na década de 1950; pela criação dos *packages* SPSS (*social sciences package*), SAS (*statistical analysis system*), na década de 1960, desenvolvidos pelos serviços de dados locais das universidades de Princeton, Northwestern, British Columbia, North Carolina, Wisconsin e Yale. E, finalmente, a profissão de *data librarian* é apresentada pela IASSIST no Congresso Mundial de Sociologia de Toronto em 1974 (RICE; SOUTHALL, 2016; KOLTAY, 2017; KELLAM; THOMPSON, 2016).

A Biblioteconomia orientada a dados não é um novo ramo da Biblioteconomia[49], fundamenta-se em uma diversidade de habilidades já incorporadas e conhecidas por bibliotecários. Práticas tradicionais são

---

[47] CESSDA. Disponível em: <http://www.unidata.unimib.it/?page_id=2136>. Acesso em: maio 2017.

[48] Cronograma de surgimento de Bibliotecas de dados disponível em: <https://docs.google.com/spreadsheets/d/1qmC_z50UDHh3Jwdlu6wGrtz1xjdfumBaad6P6cUpBdY/edit?pli=1#gid=0>. Acesso em: set. 2017.

retomadas com o objetivo de serem aplicadas a conjuntos de dados. Por exemplo, a catalogação e a organização de materiais; a preservação e a curadoria; serviços de usuários e de referência; consultoria e treinamentos são repensados para serem aplicados junto às novas práticas que envolvem o gerenciamento, a curadoria e o compartilhamento de dados. No entanto, a Biblioteconomia de Dados não serve apenas como apoio à descoberta de novas habilidades em Biblioteconomia, ela envolve-se com a geração e a preservação de dados de pesquisa, preocupa-se com praticamente todas as funções tradicionais do trabalho bibliotecário, como: aquisição e desenvolvimento de coleções, organização catalogação de acervos e implementação de serviços de referência aos usuários (RICE; SOUTHALL, 2016; KELLAM; THOMPSON, 2016).

Uma das principais tarefas da Biblioteconomia de Dados é ser *upstream*[50] (GOLD, 2007a), o que significa que ela não deve apenas se preocupar com informação publicada, mas também com fontes potenciais de dados, sendo o seu objetivo compreender como os diferentes tipos de dados podem ser capazes de gerar informação útil para pesquisa (XIA; WANG, 2014; KOLTAY, 2017).

Revisto o conceito, a origem e algumas das características dos estudos sobre dados e dados de pesquisa, percebe-se que a complexidade do conceito de dados revela a necessidade de uma ciência que estude habilidades, métodos, técnicas e tecnologias relacionadas ao uso intensivo de dados. Nesse contexto, propõe-se a *Data Science* como teoria e prática científica utilizada para compreender as diversas dimensões do conceito de dados.

Acredita-se que ela represente um novo desafio de aprendizagem e que se caracterize como uma nova oportunidade aos profissionais bibliotecários e cientistas da informação. A *Data Science* proporciona novos métodos e práticas que podem contribuir para a investigação desses profissionais. A *Data Science* também é considerada como área de estudos interligada à gestão e à análise de dados.

---

[50] *"In a seminal article on cyberinfrastructure, data and libraries in D-Lib Magazine Anna Gold has characterized these areas that fall outside the usual comfort zone of academic librarians as 'working upstream' in the research process, before the point of publication"* (GOLD, 2007a, p. 12).

## *2.4 DATA SCIENCE*

A raiz interdisciplinar do termo *Data Science* tem origem em campos como Econometria, Física, Bioestatística, Ciência da Computação, Estatística e Engenharia. Ela pode ser um sinônimo ou um termo relacionado a outros, como análise de negócios, pesquisa operacional, inteligência competitiva, análise e modelagem de dados e a descoberta de conhecimento em bases de dados. Embora a *Data Science* muitas vezes seja vista como uma novidade, registra-se nas Ciências da Computação desde a década de 1960. Zhu, Xiong (2015) e Chen (2015), ao descreverem o estado da arte da *Data Science*, ressaltam que em 1966 Peter Nauer[51] sugeriu chamar a própria Ciência da Computação de *Datalogia.* Na década de 1990, o Committee on Data of the International Council for Science (CODATA)[52] estabelece como prioridade a realização de estudos sobre os avanços em *Data Science*, utilizando o termo *Data Science* como uma forma de representar a pesquisa orientada a dados na diversas áreas do conhecimento. Nos anos 2000, surgem os primeiros periódicos, como o Journal of Data Science, da Universidade de Columbia, nos USA. A *Data Science* possui duas funções básicas: uma função é a analítica, a outra é a invenção de novas técnicas que podem traçar *insights* durante o uso de dados, o que não era possível antes dos anos 2000. Por fim, a *Data Science* surge, atualmente, como uma combinação entre o desenvolvimento de análises com dados e a elaboração de novos algoritmos inteligentes e capazes de unir a Estatística e a Ciência da Computação (ZHU; XIONG, 2015; CHEN 2015; BULINGAME; NIELSEN, 2012).

Nesse sentido, a *Data Science* pode ser vista como a teoria e a prática de extrair conhecimento de dados. Sendo uma disciplina científica preocupada com a criação de produtos e/ou serviços a partir de dados, ela busca utilizar os ambientes orientados a dados como internet, repositórios de dados, sensores climáticos, GPs e a mais recente tendência de conectar objetos em rede à *Internt of Things*[53] como *locus*

---

[51] NAUR, P. Disponível em: <https://en.wikipedia.org/wiki/Peter_Naur >. Acesso em: maio 2016.

[52] CODATA. Disponível em: <http://www.codata.org/about-codata>. Acesso em: maio 2016.

[53] A principal ideia do conceito de Internet of Things **(IoT**) vem do pervasivo, que, conforme Resmini e Rososati (2011), é a capacidade ou tendência a propagar-se, infiltrar-se, difundir-se através de vários meios, canais, sistemas, tecnologias etc. Para Atzori, Iera e Morabito (2011), a IoT significa a integração de várias tecnologias a soluções de comunicações com a Internet, isso inclui todo e qualquer tipo de objeto e/ou coisas, como, por exemplo, uma televisão, um aparelho de telefone celular, um barco ou relógio. Esses autores identificam um cenário de tecnologias que podem compor a IoT: tecnologias de identificação e rastreamento, redes de

de estudo sobre o fluxo dos dados (ZHU; XIONG, 2015; CHEN 2015).

No entanto, um dos principais problemas de investigação em *Data Science* é vê-la como a ciência dos dados, e não apenas como uma extensão de técnicas da Estatística e da Ciência da Computação. Segundo Chikio Hayashi, a *Data Science:*

> [...] não é apenas um conceito sintético para unificar a análise de dados e os seus métodos relacionados, mas também compreende os seus resultados. A *Data Science* pretende analisar e compreender os fenômenos reais com dados. Em outras palavras, o objetivo da *Data Science* é revelar as características ou as estruturas ocultas de fenômenos naturais, humanos e sociais a partir de dados, com um ponto de vista aplicado à teoria e ao método tradicional. Esse ponto de vista implica maneiras flexíveis e multidimensionais de se pensar (HAYASHI, 1996, p. 41, tradução nossa)[54].

A *Data Science* possui o objetivo de revelar as características e/ou estruturas dos fenômenos naturais, humanos e sociais a partir da utilização de dados. Ela pode ser considerada como uma nova área de estudos para profissionais que procuram se envolver com questões ligadas ao gerenciamento e à análise de dados com base em Estatística e Ciências da Computação (ZHU; XIONG, 2015; CHEN 2015).

O *status* científico da *Data Science* está ligado a diferentes objetivos, como: (a) o estudo de dados científicos, que vê a *Data Science* como método e tecnologia utilizada para conduzir pesquisas científicas por meio do gerenciamento de dados científicos, refere-se aos dados usados durante pesquisas científicas; (b) o estudo de dados de negócios observa a *Data* Science como uma técnica para criação de produtos e/ou serviços de dados, como uma metodologia para extração de soluções de problemas em ambientes de negócios (*business intelligence)*; (c) como a integração entre a estatística, a tecnologia

---

sensores e atuator *networks*, protocolos de comunicação e inteligência distribuída para objetos *smart*, sendo esse apenas um possível cenário entre outros.

[54] *"[...] is not only a synthetic concept to unify data analysis and their related methods, but also comprises its results. Data Science intends to analize and undesrstand actual phenomena whith data. In other words, the aim of Data Science is to reveal the features or the hidden structures of complicted natural, human and social phenomena whith data from a diferente point of view the estabilished or traditional theory and method. This point of view implies multidimentional, dynanmic an flexible ways of thinking"* (HAYASHI, 1996, p. 41).

computacional e a inteligência artificial, esse objetivo se preocupa com as habilidades necessárias ao cientista de dados; e (d) como um método para solução de problemas científicos e/ou de negócios relacionados à extração de conhecimentos a partir de dados, essa perspectiva integra todos os outros objetivos (ZHU; XIONG, 2015).

Smith (2006, p. 163) descreve a *Data Science* como o "[...] estudo a respeito da captura, da análise, da recuperação, do arquivamento, do intercâmbio e da mineração de para encontrar conhecimento e relações inesperadas em dados". Em *Data Science*, existe a necessidade de se extrair valor real a partir de dados. O cientista de dados deve possuir conhecimento sobre o domínio em análise e ser capaz de transmitir mensagens claras e concisas por meio da utilização de tecnologias de dados. Deve compreender o significado dos processos de colecionar, analisar, interpretar e visualizar dados oriundos de qualquer tipo de fonte (VAN DER AALST, 2014).

Assim, a nova disciplina sobre o estudo e a análise de dados exige um pensar de forma sistemática, combinando uma abordagem criativa para resolver problemas geralmente pragmáticos. Essa abordagem pode ser exemplificada na forma de pensar de um engenheiro combinada à de um designer visual (VOULGARIS, 2014).

As bases disciplinares da *Data Science* combinam uma série de conhecimentos, como a mineração de dados, o aprendizado de máquina, o *design* de sistemas, a computação distribuída, a estatística, a engenharia industrial, o conhecimento de domínio, visualização e o *big data*[55].

---

[55] O *big data* está relacionado a conjuntos de dados tão grandes, complexos, diversificados e desorganizados que não é possível aplicar a eles técnicas computacionais convencionais para armazenamento, processamento, compartilhamento e preservação dos dados. O *big data* se refere a dados tão imensos que não se encaixam em único computador, esses dados são tão desestruturados que não se adequam a linhas e colunas dos bancos de dados tradicionais; possuem fluxo tão contínuo que não podem ser armazenados em um *datawarehouse* estático. O *big data* trata de conjuntos de dados tão imensos – no volume, na velocidade e na variedade – que acabam por ser impossíveis de gerir com as ferramentas computacionais convencionais. Ele envolve vários computadores (*parallel or distributed computing*) trabalhando em conjunto um com o outro para analisar dados. O volume de dados não é a chave para compreender o *big data*, mas sim a análise do volume de dados. A questão direcionada é como analisar o grande volume de dados, como converter o *big data* em *insights*, inovações e produtos de dados. O *big data* não se preocupa somente em coletar grandes volumes de dados, mas ele busca transformar os

Sob esse viés, percebe-se que a *Data Science* emerge como um esforço multidisciplinar que combina uma série de habilidades técnicas ligadas a uma variedade de campos do conhecimento, que permite pensar ou prever protótipos e demonstrações que auxiliem no entendimento da ciência orientada a dados (VOULGARIS, 2014).

Os rumos da pesquisa realizada em *Data Science* consolidam-se em práticas e habilidades técnicas utilizadas para extrair e visualizar dados. Embora a ciência de dados seja algo mais amplo, ela pode ser definida como um dos pilares para Biblioteconomia de Dados.

Assim, identificada a combinação das disciplinas e o processo de desenvolvimento prático que compõem a *Data Science*, torna-se relevante identificar as habilidades necessárias aos cientistas de dados.

### 2.4.1 Habilidades do cientista de dados

A exploração do núcleo de habilidades que compõe a *Data Science* fundamenta-se em conhecimentos multidisciplinares que fazem parte da prática de estatísticos, cientistas da computação, analistas e gestores de dados. O modelo clássico das habilidades necessárias ao cientista de dados ressalta a ideia de que todas as habilidades necessárias devem estar disponíveis em uma pessoa. As cinco habilidades que fundamentam essa visão do cientista de dados são: *hacker,* pesquisador, analista quantitativo, conselheiro de confiança *(trusted adviser)*, especialista em negócios *(business expert)*.

Um cientista de dados deve se preocupar em adquirir as habilidades de um *hacker.* O termo denota um profissional inserido em contextos de computação de alto desempenho e criativa. Embora o termo *hacker* possa soar como algo negativo (*cracker*), o *hacker* como alguém que infringe as leis morais da computação, o termo também enfatiza as capacidades criativas dos conhecedores das linguagens de programação. A habilidade *hacker* consiste em possuir *kown how* para extrair dados de lugares remotos onde reside a informação e transformá-los para análise e interpretação. Um *hacker* é um programador

---

dados em valor real Assim, pensar em *big data* envolve presumir um grande volume de dados, desestruturados, de fluxo contínuo e em múltiplos formatos. As fontes para geração de *big data* são diversas (vídeos, sensores, genoma, satélites, *internet of things* etc.). As dimensões do *big data* compreendem fenômenos relacionados a propriedades específicas dos dados, como volume, velocidade e variedade (FEDERER, 2016; DAVENPORT, 2014; BURLINGAME; NIELSEN, 2014; VAN DER AALST, 2014)

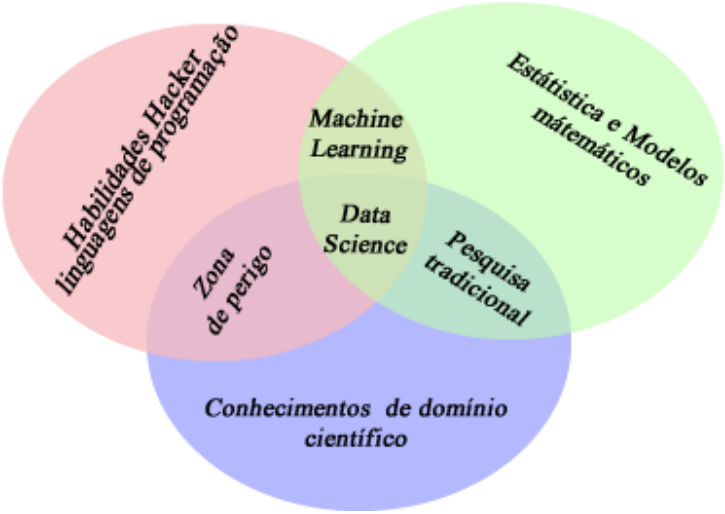
habilidoso. Para tal, deve conhecer uma diversidade de linguagens de programação (DAVENPORT, 2014).

O cientista de dados também deve possuir os atributos de um pesquisador científico. Essa aptidão envolve a capacidade de construir experimentos, projetar métodos, reunir, analisar e descrever os resultados de pesquisas científicas que envolvam o uso e o consumo de dados. Geralmente, essa habilidade exige formação acadêmica, o pesquisador ideal deve possui título de doutor. Outra habilidade necessária ao cientista de dados está relacionada ao trabalho realizado por analistas de dados quantitativos, que, além de conhecimentos matemáticos, devem possuir a capacidade de comunicação e de relações humanas, pois são chamados a aconselhar altos executivos sobre decisões internas e externas em empresas onde os dados são o produto principal. O *trusted adviser*, ou conselheiro de confiança, relaciona ter a capacidade de combinar as outras habilidades (*hacker*, pesquisador, analista quantitativo) em uma só. Essa habilidade exige a competência de explicar os processos técnicos da matemática, da computação e da análise de negócios para pessoas não técnicas. Esse tipo de cientista de dados geralmente é chamado para aconselhar líderes de empresas sobre as oportunidades e/ou riscos relacionados aos seus de produtos e/ou serviços baseados em dados. Uma última habilidade requerida ao cientista de dados é o *business expert*; por meio dessa habilidade, o cientista de dados é encarregado de gerar hipóteses e testá-las para fornecer soluções para problemas funcionais em diferentes tipos de negócios (DAVENPORT, 2014).

Outra maneira de delimitar as habilidades do cientista de dados se dá conforme o Diagrama Venn[56] elaborado por Conway (2010). São três os conjuntos que arranjam os conhecimentos necessários aos cientista de dados: habilidades *hacker*, habilidades estatísticas e conhecimento de domínio (Figura 4).

[56] Designam-se por Diagramas de Venn aqueles usados em matemática para simbolizar graficamente propriedades, axiomas e problemas relativos aos conjuntos e sua teoria. Assim, duas curvas que não se tocam e estão uma no espaço interno da outra, simbolizam conjuntos que possuem continência, ao passo que o ponto interno a uma curva representa um elemento pertencente ao conjunto.

Figura 4 - Elementos do trabalho acadêmico.




Fonte: CONWAY (2010).

O diagrama de Venn ressalta a preocupação do cientista de dados com os conhecimentos de um *hacker*. Esse grupo de habilidades não requer que o cientista de dados tenha feito um curso de Ciências da Computação, mas exige que ele possua conhecimentos no mínimo intermediários em linguagens de programação. Um cientista de dados deve entender como integrar múltiplos sistemas e conjuntos de dados. Ele deve ser capaz de conectar conjuntos de dados distintos e identificar novos *insights* em um novo conjunto de dados. O cientista de dados tem que adquirir conhecimentos mínimos para programar, em diferentes linguagens de programação, tais como: *Python*, *R*, *Java*, *Ruby*, *Matlab*, *Pig* ou *SQL*. Ele precisa estar a par de tecnologias como *Apache Hadoop, Hive* e/ou *MapReduce* (CONWAY, 2010).

Além disso, o cientista de dados precisa estar familiarizado com as disciplinas de recuperação, análise e visualização de dados, como: (1) *Natural Language Processing*: focar nas interações entre computadores e seres humanos; (2) *Machine learning:* usar computadores para desenvolver algoritmos; (3) *Conceptual modelling*: ser capaz de compartilhar e articular modelagem; (4) *Statistical analysis*: compreender e trabalhar em torno de possíveis limitações nos modelos; (5) *Predictive modelling*: a maioria dos problemas com o uso de dados está na capacidade de previsão de resultados; (6) *Hypothesis testing*: ser capaz de desenvolver hipóteses e testá-las cuidadosamente por meio de

experimentos práticos. Em suma, a habilidades *hacker* em *Data Science* exigem que o *data scientist* seja capaz de manipular arquivos de texto via linha de comando *Linux* ou *Unix* e pensar suas análises por meio de aplicação de linguagens de programação e algoritmos científicos (CONWAY, 2010).

Outra esfera do diagrama envolve conhecimentos básicos sobre Matemática e Estatística. Essa habilidade do cientista de dados exige que ele conheça métodos e ferramentas básicas aplicadas na quantificação de dados. Não é preciso ser PhD em Estatística, mas um *data scientist* deve saber como interpretar e diferenciar os tipos de análise quantitativa que podem ser aplicados a conjuntos de dados. A *Data Science* envolve a descoberta e a construção de novos conhecimentos, o que requer perguntas que motivem a elaboração de hipóteses quantitativas de predição, que podem ser compreendidas e testadas com métodos e programas de computador fundados em Estatística (CONWAY, 2010).

A terceira esfera, conhecimento de domínio, é aquela na qual o cientista de dados mostra possuir habilidades de *hacker* para programar, extrair e estruturar dados; e provavelmente o cientista de dados já pode relacionar variáveis por meio de Matemática ou Estatística. Essa parte da esfera cientista de dados combina as outras esferas e consolida as principais habilidades exigidas em *Data Science* (CONWAY, 2010).

Além dessas habilidades, é possível classificar o cientista de dados em tipos relacionados às suas habilidades ou campo de conhecimento: *developers*, *researchers*, *creative*, *business people* e *mixed/generic* (VOULGARIS, 2014).

Quadro 2 - Tipos de *data scientists.*

| Tipo | Descrição |
|---|---|
| ***Data developers*** | O objetivo de um *data developer* é adquirir informações úteis a partir de dados. Geralmente, concentram-se em questões técnicas de análise e gestão de dados. Seu trabalho envolve o tratamento de dados para organizá-los em grandes bases de dados, preparando-os para consulta, análise e visualização dos resultados. |
| ***Data researchers*** | São geralmente do mundo acadêmico e demonstram uma sólida experiência em estatísticas ou em qualquer ciência que empregue estatísticas (cientistas sociais). |
| ***Data creative*** | Geralmente, possuem a experiência acadêmica e são excepcionalmente bons com tecnologias (*designers* de *software* para governança e análise de dados), aprendizado de máquina e proficientes em alguma linguagem de programação. |
| ***Data business people*** | Trabalham com negócios e, geralmente, lideram equipes especializadas em *Data Science.* Seu foco é o aumento das receitas de uma empresa. |
| ***Mixed/generic Data scientists*** | São como o *data business people,* mas sem a ampla experiência ou o foco em negócio. |

Fonte: VOULGARIS (2014).

No Quadro 2, ressalta-se que não existem dois cientistas de dados que possuam o mesmo tipo de problema. Acredita-se que o futuro da *Data Science* se baseie em uma mistura perfeita entre habilidades quantitativas e *software skills* (VAN DER AALST, 2014).

Tendo em vista as habilidades necessárias a um cientista de dados, propõe-se interconectar as habilidades requeridas a esse cientista como campo de habilidades também necessárias aos bibliotecários de dados. Busca-se identificar as habilidades necessárias para transformar dados em informação de valor.

### 2.4.2 Habilidades do bibliotecário de dados

O bibliotecário de dados é:

> [...] um bibliotecário que trabalha todo o ou parte do tempo com dados. [...] na maioria dos casos, esse bibliotecário estará trabalhando em algum tipo de biblioteca institucional, seja ela

> acadêmica, escolar, pública ou de negócios. Existem bibliotecários que trabalham com dados fora das bibliotecas, por exemplo, criando taxonomias corporativas ou trabalhando com pesquisadores fora da biblioteca e é provável que haja mais empregos fora das bibliotecas, já que os empregadores percebem como um bibliotecário bem treinado pode ajudar a encontrar, a organizar e analisar dados e/ou informações (HENDERSON, 2017, p. 16, tradução nossa)[57].

Como os materiais tradicionais de uma biblioteca física, os dados também precisam ser coletados, descritos e organizados para que possam ser utilizados por um usuário. Os usuários de dados geralmente precisam encontrar dados que contenham informações pertinentes às suas perguntas de pesquisa, precisam de acesso aos dados brutos e aos dados secundários, precisam poder citar esses dados como citam outros documentos (HENDERSON, 2017, p. 34-35).

Conforme Stuart (2011, on-line), se

> [...] a biblioteca tradicionalmente facilitou o acesso aos documentos, atualmente, os profissionais da informação também podem facilitar o acesso aos dados, mesmo que os dados não se encaixem necessariamente nos mesmos formatos de documentos que as bibliotecas costumavam oferecer.

Assim, um bibliotecário de dados pode trabalhar junto a pesquisadores e a estudantes auxiliando-os durante a manipulação de seus dados de pesquisas, participando da coleta, da organização, da classificação e da preservação dos dados, fazendo dessa prática uma extensão de suas atividades cotidianas. Um bibliotecário de dados pode usar os princípios fundamentais da Biblioteconomia para: (a) adquirir, manter e gerenciar coleções de arquivos de dados usados para análise

---

[57] "*A data librarian is a librarian who works all or part of the time with data. This book also assumes that most librarians will be working in some sort of institutional library, be it academic, school, public, business, or other special library. There are librarians working with data outside of libraries, for example, creating a corporate taxonomy or working with researchers outside of a library setting, and it is likely that there will be more jobs outside of libraries as employers realize how a well-trained librarian can help find, organize, and analyze information"* (HENDERSON, 2017, p. 34-35).

primária e secundária; (b) oferecer serviços de referência e assistência técnica, garantindo o acesso aos dados a toda uma comunidade de usuários; e (c) fornecer ensino relativo ao acesso e à utilização dos conjuntos de dados como elemento de ligação para áreas afins sujeitas às Ciência da Informação (INTERNATIONAL ASSOCIATION FOR SOCIAL SCIENCE INFORMATION SERVICES AND TECHNOLOGY, 2016).

Os bibliotecários de dados, entre outras habilidades, devem se preocupar com avaliação e retenção de dados, promoção de dados, *marketing* de dados, conscientização de dados, coordenação de práticas com dados em toda unidade e instituição, habilidades de negociação e habilidades de gerenciamento de expectativas de dados. O bibliotecário de dados compartilhará as mesmas habilidades de um gestor de dados, no que tange à preservação e à avaliação dos valores econômicos atribuídos aos dados. Em relação ao uso de tecnologias, o bibliotecário de dados deve ter em mente a habilidade *hacker* de um cientista de dados (XIA; WANG, 2014).

O bibliotecário de dados deve possuir as habilidades necessárias para trabalhar com qualquer tipo de dados, sejam eles observacionais, computacionais ou experimentais. Ele deve compreender como as investigações sobre dados modificam as práticas e as teorias que fundamentam a Biblioteconomia. Além de deter o papel de ajudar pesquisadores a coletar, localizar e recomendar recursos de dados apropriados, como o uso de provedores de serviços de dados de pesquisa, os bibliotecários de dados podem oferecer serviços de consultoria sobre o preparo e a aquisição de dados para análise de questões de pesquisa em disciplinas específicas (DEKKER; LACKIE, 2016).

Assim, o trabalho de um bibliotecário de dados pode estar ligado à manipulação de dados, mas também se conecta com o trabalho humano de interpretação, tomada de decisão e comunicação de dados. McCaffrey e Giesbrecht (2016) listam algumas das habilidades e/ou competências necessárias a um bibliotecário de dados. Os autores sugerem três áreas distintas para a atuação desse profissional: (a) gestão e curadoria de dados; (b) visualização de dados e representação de dados geoespaciais; e (c) serviços avançados de referência a dados de pesquisa. Kennan (2016) apresenta um estudo sobre as habilidades requeridas por aqueles que contratam bibliotecários de dados. A autora lista as habilidades exigidas por organizações de pesquisa e universidades nos USA, Canada, UK e Austrália, conforme Quadro 3.

Quadro 3 - Habilidades requeridas aos bibliotecários de dados.

| Habilidades | Descrição |
|---|---|
| **Interpessoais e características comportamentais** | Capacidade relacionada à comunicação formal, como a escrita de documentação técnica e estudos de caso. Outro ponto ligado a este perfil é a capacidade de se adaptar a novas experiências, de estar constantemente em busca de atualização. |
| **Conhecimento contextual sobre ambientes institucionais** | Essa habilidade envolve conhecimentos a respeito de políticas de financiamento oferecidas por agências que fomentam a investigação científica. Ressalta-se que essa habilidade exige a compreensão de processos éticos do processo científico, métodos de pesquisa disciplinar, formas de comunicação científica, propriedade intelectual, formas de acesso, marcos legais e regulatórios (*copyright e creative commons*) e políticas de direito autoral. |
| **Uso de dados** | Inclui a compreensão de tipos de dados (quantitativos, qualitativos), padrões e esquemas de metadados, (Dublin Core, RDF), assim como questões relacionadas a identificadores únicos (*Digital Object Identifiers*) e preservação de dados digitais. |
| **Conhecimentos sobre tecnologias de manipulação de dados** | Os conhecimentos em tecnologia mais relevantes aos bibliotecários de dados abrangem: linguagens de programação (*Python, SQL, Java, XML entre outras*), *design* e estrutura de bases de dados, APIs de recuperação de dados, o *design* centrado no usuário, ferramentas de processamento de linguagem natural, *internet of things* e *big data*. |

Fonte: Kennan (2016).

Assim, é preciso um esforço multidisciplinar de aprendizagem, por parte dos bibliotecários de dados, que devem combinar as habilidades e os conhecimentos necessários à manipulação e à gestão de dados. O papel fundamental dos bibliotecários de dados é buscar adquirir experiência técnica para aquisição, análise e interpretação de dados.

Por fim, o bibliotecário de dados deve transformar os dados em algo reutilizável, compartilhável e preservável em longo prazo. Ele deve conhecer os esquemas de metadados, os repositórios de dados, os conjuntos de dados que são utilizados por seus usuários pesquisadores,

além de ter uma boa noção sobre as mais diversas metodologias e disciplinas científicas. Mas, principalmente, ele deve possuir habilidades ligadas a atividades de coleta, de descrição, de ensino (*data literacy*), de construção de sistemas, de elaboração de serviços de referência e de criação de políticas de gerenciamento de dados institucionais (MCCAFFREY; GIESBRECHT, 2016; KENNAN, 2016; DEKKER, LACKIE, 2016).

O interesse de bibliotecários em adquirir habilidades técnicas para gerenciamento de dados de pesquisa demostra que o uso de dados em Biblioteconomia de Dados pode estar ligado à elaboração de produtos e/ou serviços de dados vinculados ao gerenciamento de dados de pesquisa em bibliotecas. Assim, a próxima questão a ser discutida revela o papel dos bibliotecários de dados em relação ao uso, à gestão e à curadoria de dados de pesquisa.

### 2.4.3 Bibliotecários de dados e os dados de pesquisa

O bibliotecário de dados deve ter experiência em tarefas corriqueiras desenvolvidas em bibliotecas acadêmicas, como gestão, supervisão e liderança. As mesmas habilidades utilizadas para oferecer produtos e/ou serviços a usuários de biblioteca, no que tange à gestão de livros, artigos, materiais audiovisuais, também podem ser aplicáveis para oferta de serviços de dados de pesquisa. Explicam Tenopir et al. (2017) que essas habilidades abrangem disciplinas informacionais que se preocupam com coleta, armazenamento, organização, recuperação e divulgação de informações, seja qual for o assunto ao qual se relacionem.

Existem algumas instituições internacionais que estão preocupadas com a Biblioteconomia de Dados e os dados de pesquisa. Nos USA, é a Association of College and Research Libraries (ACRL)[58]; na Europa, a Ligue des Bibliothèques Européennes de Recherche (LIBER)[59]. Essas instituições declaram que a biblioteca é o espaço institucional ideal para fornecer acesso aos dados de pesquisa.

> [...] (ARL) declarou que as bibliotecas acadêmicas estão em uma posição favorável para ajudar os pesquisadores a enfrentar os desafios do paradigma de pesquisa do uso intensivo de dados, porque elas podem fornecer cada vez mais

[58] ARL. Disponível em: <http://www.ala.org/acrl/>. Acesso em: set. 2017
[59] LIBER. Disponível em: <http://libereurope.eu/about-liber/>. Acesso em: set. 2017

> serviços relacionados a consulta a de dados e têm experiência e habilidades para promover a comunicação cross-departmental, cross-campus e etc na instituições. [...] É sublinhado que os bibliotecários estão familiarizados com as necessidades de pesquisa de pesquisadores e que podem fornecer apoio para o entendimento dos novos modelos de publicação de dados, [...] a (LIBER) declarou que as bibliotecas devem auxiliar o corpo docente na integração do gerenciamento de dados no currículo. (KOLTAY, 2017, p.4-6, tradução nossa)[60]

Os bibliotecários de dados deveriam trabalhar com pesquisadores desde as primeiras fases da investigação científica, ajudando com o processo de documentação e garantindo que os dados sejam preservados, utilizáveis e reutilizáveis em longo prazo. Os bibliotecários deveriam ofertar serviços para auxiliar pesquisadores a compreender e a documentar suas pesquisas. Serviços que compreendem: avaliação e divulgação de dados de pesquisa; consultoria para elaboração de planos de gerenciamento de dados; esquemas de metadados; incorporação de dados em coleções convencionais; alfabetização de dados; oferta de serviços de referência para ajudar os usuários (pesquisadores) a localizar e a citar dados fornecendo soluções para publicação e preservação de dados digitais (WITT; HORSTMANN, 2016).

A gestão de dados de pesquisa é uma tarefa que deve ser assumida por bibliotecários de dados. Os dados de pesquisa são parte essencial do registro acadêmico científico, e a gestão dos dados de pesquisa é cada vez mais vista como uma tarefa elementar para bibliotecas acadêmicas. A gestão de dados de pesquisa é essencial para assegurar a transparência da pesquisa científica, preservar dados e permitir a reutilização e a reanálise de dados, garantindo os avanços do conhecimento científico (TENOPIR et al., 2017).

A gestão de dados compreende a organização, o armazenamento, o acesso e a preservação de dados. Alguns aspetos gerais dessa atividade

---

[60] [...] (ARL) declared that academic libraries are in a favourable position to help researchers meet the challenges of a data-intensive research paradigm because they are increasingly providing data consultation services and have experience and skills in fostering cross-departmental, cross-campus, etc. It is underlined that librarians are familiar with the research data needs of researchers and have been among the supporters of innovative publishing models [...] (LIBER) declared that libraries should assist faculty with the integration of data management into the curriculum.

são: nomeação de arquivos; acesso aos dados; documentação de dados; criação de metadados e vocabulários controlados; armazenamento de dados; arquivamento e preservação de dados; compartilhamento e reutilização de dados; privacidade em dados; propriedade intelectual em dados; e a publicação de dados (HENDERSON, 2017).

Planos de gestão de dados ilustram como os dados de pesquisa poderão ser visíveis, acessíveis e reutilizáveis pelo público desses dados (TIBBO; JONES, 2017). Os serviços de dados de pesquisa são exigidos pelo paradigma de pesquisa pautado no uso intensivo de dados e devem ser integrados a serviços de biblioteca (TENOPIR et al., 2014). Os serviços de dados de pesquisa consistem na gestão e na curadoria de dados de pesquisa, que não são atividades idênticas, mas não parecem estar claramente separadas uma da outra (KOLTAY, 2017,).

Nesse sentido, sugere-se que existem dois tipos de serviços de dados em bibliotecas: serviços de gestão de dados, que objetivam auxiliar pesquisadores enquanto estes estão trabalhando com seus dados (desenvolvimento de planos de gestão de dados, treinamento para documentar e organizar dados, encontrar e assegurar o armazenamento seguro dos dados de pesquisa);e serviços de curadoria de dados, que compreendem o armazenamento e a preservação de dados com segurança e em longo prazo. Uma vez concluída uma pesquisa, é preciso compartilhar e proporcionar a reutilização dos dados (HENDERSON, 2017, p. 27).

No que tange ao uso de dados de pesquisa pelos pesquisadores, os bibliotecários de dados devem possuir os conhecimentos e as habilidades necessárias para aconselhar pesquisadores em relação: (a) à preservação dos resultados da pesquisa; (b) ao gerenciamento e à curadoria de dados, o que inclui descoberta, acesso, disseminação, preservação e portabilidade; (c) apoiar pesquisadores no cumprimento dos diversos mandatos financiadores, incluindo requisitos de acesso aberto; (d) às potenciais ferramentas de manipulação de dados que podem ser usadas pelas disciplinas científicas; sobre mineração de dados; (e) ao uso de metadados (RICE; SOUTHALL, 2016, p. 317).

Os bibliotecários de dados precisam de habilidades técnicas para extrair, analisar e apresentar dados de pesquisa. Assim, um bibliotecário de dados necessita combinar habilidades e conhecimentos para o gerenciamento e a curadoria de dados. Tais competências compreendem organização, armazenamento, acesso e preservação de dados de pesquisa (HENDERSON, 2017).

Tenopir et al. (2014, p. 85, tradução nossa) argumentam que

> [...] as bibliotecas devem estar envolvidas durante todo o processo de pesquisa e precisam se envolver ativamente com a curadoria, aconselhar e preservar os resultados da pesquisa. Alguns papéis adicionais sugeridos para as bibliotecas são desenvolver a conscientização dos pesquisadores para que adotem regras de arquivamento e preservação para os seus dados e busquem treinar os bibliotecários de dados para utilizarem seus dados. [...] os perfis profissionais sugeridos pelos membros da Assossiation of Research Libraries nos USA e no Canadá incluíram como referências [treinamentos sobre] *e-Science*, curadoria e gestão de dados de pesquisa como [tópicos] relevantes [para criação de] serviços de [dados de pesquisa] ligados às comunicações acadêmicas[61].

De acordo com Cox, Verbaan e Sen (2013), um bibliotecário de dados precisa adquirir essas competências para apoiar o gerenciamento de dados de pesquisa. Além disso, papéis existentes na promoção do acesso aberto também estão ligados à agenda do gerenciamento de dados de pesquisa. Embora o papel e as competências necessárias se alinhem aos papéis atuais dos bibliotecários, existem algumas lacunas no conhecimento dos bibliotecários, como as teorias, as metodologias e o jargão técnico de cada tipo de cientista. É preciso que os bibliotecários aumentem a sua compreensão sobre o que os dados de pesquisa significam para os pesquisadores. Assim, o desenvolvimento de habilidades práticas torna-se relevante (COX; VERBAAN; SEN, 2013).

Após apresentadas a origem e as habilidades necessárias aos cientistas e aos bibliotecários de dados, busca-se identificar os instrumentos e as tecnologias passíveis de serem usadas em *Data Science*.

---

[61] "[...] libraries should be involved throughout the research process and need to be actively engaged in curating, advising, and preserving research outputs. Some additional suggested roles for libraries are to develop researchers' data awareness, to adopt a data archiving and preservation role, and to train data librarians. [...] narrative profiles created by ARL members in the U.S. and Canada included references to e-science/data curation and management as an important service supporting faculty success and scholarly communications" (Tenopir et al., 2014, p. 85).

## 2.5 TECNOLOGIAS E INSTRUMENTOS PARA MANIPULAÇÃO DE DADOS

A *Data Science* oferece um vasto corpus de algoritmos, tipos de análise, linguagens de programação e *softwares* que são utilizados para a manipulação de dados. *Data handling*, nome utilizado em inglês para designar práticas de manipulação de dados, está no centro da preparação de dados para pré-análise, como em processos de limpeza, recodificação, criação de documentação, mesclagem, importação e exportação de dados (migração de formatos dados – de um tipo de *software* para outro sem perder conteúdo). Igualmente, a *data handling* se configura como o núcleo de processamento de dados para pós-análise – serve para preparar conjuntos de dados para o armazenamento final em formato apropriado e para sua efetiva preservação em longo prazo em um repositório de dados (RICE; SOUTHALL, 2016).

Os conhecimentos sobre tecnologias de manipulação de dados são uma habilidade fundamental aos bibliotecários de dados. Adquirir competência nessa área exige que se tenha noção do uso de *softwares* e algoritmos utilizados para coleta, manipulação, análise e visualização de dados.

### 2.5.1 *Softwares* e algoritmos para análise de dados

Compreende-se que o *software* científico seja um programa de computador utilizado e desenvolvido por meio de computação científica, especificamente criado para uso e aplicação por uma comunidade científica. Por exemplo: os sistemas de CAD (*Computer Aided Design*) usados por engenheiros e arquitetos para simular plantas de construção civil (SEMELER; PINTO; VIANNA, 2017).

O *software* científico permite traduzir os métodos de concepção das Ciências (sejam elas Exatas, sejam Humanas e/ou Sociais) para algoritmos computacionais. Esses *softwares* aplicam regras (algoritmos) por meio de linguagens de programação. Praticamente todas as tarefas em *Data Science* podem ser executadas por essas ferramentas (*softwares* e linguagens); no entanto, elas dependem de algoritmos.

Cormen et. al. (2002, p. 3) explicam que um algoritmo é:

> [...] qualquer procedimento computacional bem definido que toma algum valor ou conjunto de valores como entrada e produz algum valor ou conjunto de valores como saída. [...] um algoritmo

> é uma ferramenta para resolver um problema computacional bem especificado. O enunciado do problema especifica em termos gerais o relacionamento entre a entrada definida e a saída desejada. [...] um algoritmo é dito correto se, para cada instância da entrada ele para com a saída correta [...] um algoritmo correto responde ao problema computacional dado (CORMEN et. al., 2002, p. 3).

Um algoritmo pode ser especificado em linguagem comum, como inglês, português, ou por meio de linguagens de programação. Ele também pode ser arquitetado como pseudocódigo. Esse pseudocódigo é um tipo de inglês ou português estruturado usado para descrever algoritmos. Os pseudocódigos não são linguagens de programação. Geralmente, são utilizados para se aprender lógica de programação. Por meio de pseudocódigos é possível pensar procedimentos computacionais se concentrando apenas na lógica do algoritmo, sem se distrair com os detalhes de sintaxe das linguagens de programação (DALBEY, 2001).

Na prática, a *Data Science* utiliza algoritmos para cumprir tarefas como as de classificação, regressão, agrupamento, análise de conteúdo e desenvolvimento de mecanismos de recomendação. Por meio da *Data Science*, são aplicados os mais variados tipos de análise estatística. Essas análises consolidam as tarefas executadas pelos algoritmos.

A seguir, apresentam-se dois quadros: o Quadro 4, sobre as tarefas e os tipos de algoritmos aplicados em *Data Science*, e o Quadro 5, sobre os tipos de análise que aplicam esses algoritmos.

No primeiro quadro, é possível visualizar as tarefas executadas por algoritmos comuns utilizados pelos cientistas de dados: classificação, *clustering*, regressão, simulação, análise de conteúdo e desenvolvimento de mecanismos de recomendação.

Quadro 4 - Habilidades requeridas aos bibliotecários de dados.

| TAREFA | O QUE É? PARA QUE SERVE? | ALGORITMOS |
|---|---|---|
| **Classificação** | Os algoritmos de classificação são comumente usados para classificar pessoas ou entidades em um ou mais grupos e também são utilizados para realizar análises preditivas. A | Os algoritmos de classificação incluem: *logistic regression*, *decision trees*, *boosted decision trees*, *multimodal neural* |

| | | |
|---|---|---|
| | classificação algorítmica preocupa-se com a colocação de determinadas amostras em categorias determinadas. Se as categorias forem predefinidas, então se chama de classificação supervisionada. Caso contrário, chama-se de classificação não supervisionada. Para resolver problemas de classificação mais complexos, que não estejam circunscritos nessas duas classes, é possível usar técnicas multimodais para prever outras classes. | *network*, entre outros. |
| ***Clustering*** | O *clustering* usa o aprendizado não supervisionado para agrupar dados em classes distintas. Uma das principais diferenças entre algoritmos de *clustering* é que o resultado do *cluster* é previamente desconhecido. Antes de utilizar-se um algoritmo de *clustering*, não se sabe a que *cluster* cada ponto de dados pertence. A estatística dispõe de uma série de ferramentas para o *clustering*. São utilizados por meio de métricas de distância que servem para agrupar dados semelhantes em conjuntos similares. Os algoritmos de *clustering* também são chamados de classificação não supervisionada, pois correspondem a um processo do qual ainda não se conhece o padrão. | Um dos algoritmos mais empregados é o de *k-means*. Com esse algoritmo, é possível decidir *a priori* quantos grupos se quer agrupar. Por exemplo, se K=3, o algoritmo produz três *clusters*. |
| **Regressão linear** | A regressão linear é uma das técnicas de previsão mais antigas utilizadas em Estatística, e o seu objetivo é prever um determinado | Os algoritmos de regressão incluem: *regression*, *decision trees*, *neural network* e *boosted decision tree* |

| | | |
|---|---|---|
| | resultado em um conjunto de variáveis observadas. Um modelo de regressão linear simples é uma função linear. | *regression*. |
| **Simulação** | A simulação é amplamente utilizada pela indústria para modelar e otimizar os processos do mundo real. Os engenheiros há muito têm se utilizado de técnicas matemáticas para simular a aerodinâmica de asas de aviões ou carros. Por exemplo, a simulação faz com que as empresas economizem custos de pesquisa e desenvolvimento, uma vez que já não têm que fazer todos os seus testes com modelos físicos reais. Além disso, a simulação oferece a oportunidade de testar cenários e ajustar variáveis aos modelos de computador. A Estatística oferece técnicas para simulação e otimização. | Um exemplo é o método de *Markov chain analysis*, que se aplica para simular as mudanças de estado em um sistema dinâmico. A programação linear pode ser usada para otimizar as rotas de transporte de uma empresa de entregas terrestre ou companhias aéreas. Outro exemplo é a simulação de Monte Carlo, possível de ser adotada para encontrar as melhores condições de otimizar certo resultado de negócio, como o lucro. |
| **Análise de conteúdo** | A análise de conteúdo é usada para minerar o conteúdo em arquivos de texto, imagens, vídeos para capturar *insights*, a exemplo da mineração de textos, que utiliza estatística e linguística para entender o significado em textos. A análise de conteúdo pode ser automatizada por vias da mineração de texto e do uso de técnicas da *Natural Language Processing* (NLP), o que permitirá minerar e identificar o significado de palavras-chave contidas em determinado conteúdo. A *Machine learning,* nesse contexto, oferece ferramentas apropriadas para a análise e o reconhecimento de | *Natural Language Processing* (NLP), *Machine learning*. |

| | | |
|---|---|---|
| | padrões em grandes quantidades de textos. | |
| **Mecanismos de recomendação** | Os mecanismos de recomendação são amplamente utilizados em projetos de serviços centrados no uso de dados. A Amazon, por exemplo, utiliza-se desses algoritmos para recomendar produtos com base nas preferências de seus usuários. Há três grandes abordagens para mecanismos de recomendação. A *collaboration filtering* (CF) faz recomendações com base nas semelhanças entre os usuários ou itens. Com a CF, analisam-se os dados desses itens para encontrar semelhanças entre eles. Os dados são retirados das interações entre os usuários e o sistema durante a utilização de conteúdos digitais. A segunda classe de mecanismos de recomendação faz sugestões ao analisar o conteúdo selecionado pelo usuário. Nesse caso, as técnicas de mineração de texto e NLP são adotadas para analisar e recomendar conteúdo. Os conteúdos similares são agrupados, o que constitui a base do mecanismo de recomendação para os novos usuários. A terceira classe desses mecanismos usa algoritmos de aprendizado de máquina para determinar a afinidade do usuário com o produto. | Os algoritmos para elaboração de mecanismo de recomendação são o *Naïve Bayes*, o *Microsoft Association Rules, collaboration filtering* e os *rules package* em R. |

Fonte: BARGA; FONTAMA; HYONG (2014).

Os algoritmos são ferramentas fundamentais no que tange à escolha de um tipo de análise que pode ser aplicada a um conjunto de dados. Um estudo centrado no uso de dados pode concentra-se em diferentes tipos de análises. Segundo, Barga, Fontama e Hyong (2014), elas são: análise descritiva, análise de diagnóstico, análise preditiva e análise prescritiva.

A análise serve para gerar *insights* a partir de dados brutos, criar informação a partir de dados. A função da análise é tornar os dados compreensíveis. Sob esse viés, apresentam-se, em ordem crescente de complexidade (1-descritiva, 2-diagnóstico, 3-preditiva e 4-prescritiva), os tipos de análise que podem ser utilizados em conjunto com técnicas algorítmicas para cumprir as tarefas de classificação, regressão, agrupamento, análise de conteúdo e desenvolvimento de mecanismos de recomendação, conforme o Quadro 5.

Quadro 5 - Tipos de análises de dados.

| TIPO DE ANÁLISE | DESCRIÇÃO | TÉCNICAS E ALGORITMOS |
|---|---|---|
| **1 - Descritiva** | A análise descritiva busca responder à pergunta: O que aconteceu? Esse tipo de análise é baseado em dados históricos e dados atuais. Por exemplo, um analista de negócios centrado em dados pode usar esse tipo de análise para propor soluções modernas de *Business Intelligence*. A descrição é usada para explicar o que está acontecendo em uma determinada situação. Essa classe de análise envolve a intervenção humana. | Comumente, são usadas estatísticas descritivas como gráficos, historiogramas ou *data clustering.* |
| **2 - Diagnóstico** | A análise de diagnóstico ajuda a entender por que certas coisas aconteceram, identifica os principais fatores. Utiliza-se esse tipo de análise para encontrar respostas sobre por que algo específico aconteceu. Ou o que deu errado. As | Uma análise de diagnóstico utiliza técnicas como *clustering*, *classification*, *decision tree* e *content analysis*. Essas técnicas estão disponíveis em |

| | | |
|---|---|---|
| | análises de diagnóstico são úteis para deduzir e inferir o sucesso ou o fracasso de subcomponentes de qualquer iniciativa centrada em dados. | processos como *data mining* e *machine learning* |
| **3 - Preditiva** | A análise preditiva é utilizada para prever acontecimentos. Serve para tentar antecipar a probabilidade de um resultado incerto. Embora esse tipo de análise esteja baseado em dados históricos e atuais, a análise preditiva vai um passo além das análises descritivas. A análise preditiva envolve a construção de modelos complexos de análise, a fim de antever um evento ou tendência. Por exemplo, em um contexto comercial, essas análises seriam realizadas com o foco voltado aos negócios. | As principais técnicas aplicadas aqui são: *neurais networks*, *decision tree*, *random forests*, *boosted decision tree*, *Monte Carlo simulation* e *regression*. |
| **4 - Prescritiva** | O quarto tipo de *analytic*, análise prescritiva, irá sugerir o melhor curso de ação a ser tomada para otimizar os resultados. A análise prescritiva é útil em cenários como: *channel optimization*, *portfolio optimization* e *traffic optimization to find the best route given current traffic conditions*. Esse tipo de análise tem como objetivo otimizar processos, estruturas e sistemas por meio de ações baseadas em análise preditiva – essencialmente dizendo o que se deve fazer com base em uma estimativa do que vai acontecer. | As técnicas mais utilizadas são: *decision tree*, *linear and non-linear programming*, *Monte Carlo simulation*, *game theory* e *data mining*. |

Fonte: BARGA; FONTAMA; HYONG (2014).

Após apresentada a ideia fundamental sobre *software* científico, modelos de algoritmos e tipos de análise em *Data Science*, abordam-se conceitos sobre repositórios de dados de pesquisa, sistemas desenvolvidos para implementar esses repositórios, os protocolos, as linguagens de programação e os *softwares* necessários para coleta e análise de dados de pesquisa. Tais ferramentas são apresentadas porque podem ser utilizadas para a criação de produtos e/ou serviços de dados de pesquisa em bibliotecas. Também é relevante lembrar que os repositórios de dados de pesquisa são a parte da ciberinfraestrutura de *e-Science* que os bibliotecários de dados devem investigar.

### 2.5.2 *Software* para elaboração de repositórios de dados

Os repositórios de documentos digitais surgem no início dos anos 1990 para disseminar publicações e outros tipos de objetos digitais. Desenvolvidos com a utilização de *software* livre, atualmente estão relacionados com os movimentos de abertura[62] ao conhecimento, como o *Open Source Initiative*[63], o *Open Access* e o *Open Data*[64]. Uma das primeiras iniciativas para criação de repositórios foi o arXiv[65], desenvolvido pela Universidade de Cornell, nos USA, iniciado em 1991 como uma biblioteca digital para *preprints* na área da física (CONSELHO NACIONAL DE ARQUIVOS, 2014; RICE; SOUTHALL, 2016).

O termo repositório é explicado no relatório sobre o estado da arte dos repositórios de dados apresentado pelo comitê ibero-americano, o RRCAP em 2010:

> [...] designa um sistema informático em que existe uma plataforma de armazenamento de objetos representados em ficheiros, capaz de incorporar novos objetos à medida que são produzidos ou submetidos. O repositório oferece serviços que são dirigidos a quem deposita, a quem pesquisa e aos administradores do sistema (RODRIGUES et al., 2010, p. 22).

---

[62] OPEN. Disponível em: <http://opendefinition.org//>. Acesso em: ago. 2017.
[63] OSI. Disponível em: <https://opensource.org/history/>. Acesso em: ago. 2017
[64] OD. Disponível em: <https://okfn.org/opendata//>. Acesso em: ago. 2017
[65] ARXIV. Disponível em: <https://arxiv.org/>. Acesso em: ago. 2017

Lynch (2003) explica que um repositório digital se caracteriza como um conjunto de serviços oferecidos por uma instituição para os membros de sua comunidade. Geralmente, estão focados no gerenciamento e na disseminação de materiais digitais produzidos ou criados por membros de uma comunidade.

Segundo ONLINE COMPUTER LIBRARY CENTER, (2007, p. 3), a função essencial dos repositórios digitais confiáveis[66] é "[...] oferecer à sua comunidade-alvo o acesso confiável e de longo prazo aos recursos digitais por ele gerenciados, no presente e no futuro"[67]. Os repositórios digitais devem aceitar a responsabilidade pela manutenção em longo prazo dos recursos digitais, o que implica possuir um sistema que apoie a interoperabilidade[68] da informação digital, que demostre responsabilidade fiscal e sustentabilidade em meio digital, projetar-se enquanto sistema que assegure a gestão, o acesso e a segurança dos objetos digitais nele depositados. Um repositório deve estabelecer metodologias de avaliação, políticas e práticas de desempenho que possam ser auditadas e mensuradas, como regras de sustentabilidade, segurança e infraestrutura tecnológica. Além de assumirem o compromisso com normas de confiabilidade, como aquelas do Open Archival Information System – OAIS que resultam na norma ISO16363:2012[69] –, que lista os critérios essenciais para os repositórios digitais confiáveis (SAYÃO, 2010; CONARQ, 2014; RICE; SOUTHALL, 2016).

Os repositórios podem ser implementados com uma diversidade de tecnologias de *software*, mas, de modo geral, são elaborados com plataformas de *software* livre já conhecidas e implementadas internacionalmente pela comunidade de bibliotecas, como o *DSpace,* o *E-prints,* o *Fedora,* o *Dataverse* e o *CKAN*. Essas plataformas de *software* livre foram desenvolvidas para proporcionar a coleta, a preservação e a divulgação de publicações e dados de pesquisa, mas podem ser usadas para agregar qualquer tipo de conteúdo em formato digital. O Quadro 6 apresenta uma descrição desses *software*s enquanto repositórios de dados de pesquisa.

---

[66] Grupo de trabalho liderado pelo Research Library Group – RLG e pelo Online Computer Library Center – OCLC.2.

[67] *"[..] a mission to provide reliable, long-term access to managed digital resources to its designated community, now and into the future"*

[68] Intercâmbio e compartilhamento e *linkage* com outros repositórios de dados e outros sistemas de informação (repositórios institucionais, bibliotecas digitais de publicações acadêmicas, editoras científicas) (SAYÃO; SALES 2016).

[69] ISO16363:2012. disponível em: <https://www.iso.org/obp/ui/#iso:std:iso:16363:ed-1:v1:en>. Acesso em: set. 2017.

Quadro 6 - *Softwares* para elaboração de repositórios de dados.

| SOFTWARE | DESCRIÇÃO |
|---|---|
| **DSPACE**[70] | Foi lançado pelo MIT Libraries e pelo Hewlett-Packard Labs em 2002 com o objetivo de fornecer um sistema de repositório para documentos digitais resultantes de investigação ou destinados à educação e distribuído com uma licença de código aberto. A partir de 2007, o MIT e a HP criaram a Dspace Foundation, uma organização sem fins lucrativos para promover a plataforma e suportar os seus utilizadores. Em 2009, esse suporte passou para a DuraSpace Foundation, também uma organização sem fins lucrativos dedicada a tecnologias de código aberto e da nuvem para bibliotecas, universidades, centros de investigação e organizações do património cultural. |
| **E-PRINT**[71] | O e-Prints Repository Software é mantido pela School of Electronics and Computer Science, da Universidade of Southampton, no Reino Unido. A plataforma é distribuída com base em licença de código aberto. Além de oferecer as funcionalidades comuns nos repositórios institucionais, tem associado o EPrints Services, uma equipe de consultoria que pode acompanhar um projeto de instalação de um repositório desde a análise e desenvolvimento personalizado até ao fornecimento de serviços de gestão. |
| **FEDORA**[72] | O Fedora não é uma plataforma para repositórios como o Dspace ou o e-Prints, mas uma arquitetura extensível que pode ser usada para desenvolver *software* para repositórios. Criada pela Universidade de Cornell, é atualmente mantida pela fundação DuraSpace, tal como o DSpace. Tem princípios como o da agregação de conteúdos locais e distribuídos em objetos digitais e a associação destes a serviços [Fedora]. A arquitetura inclui ainda um modelo de relações baseado no RDF (Resource Description Framework) do W3C usado para ligar os objetos aos seus componentes. Está disponível em licença de código aberto e tem sido usado em diversas |

[70] DSPACE. Disponível em: <www.dspace.org>. Acesso em: ago. 2017.
[71] E-PRINTS. Disponível em: <http://www.eprints.org/uk/>. Acesso em: ago. 2017.
[72] FEDORA. Disponível em: <https://getfedora.org/pt_BR/>. Acesso em: ago. 2017.

| | |
|---|---|
| | aplicações para bibliotecas digitais, arquivos, repositórios institucionais e sistemas de objetos de aprendizagem. |
| **DATAVERSE**[73] | O *Dataverse Network* (DVN) é um *software* de código aberto utilizado para o gerenciamento de coleções de dados. O principal objetivo do DVN é resolver os problemas de compartilhamento de dados e replicação de informação científica na *web*. Oferece suporte para serviços de arquivamento, *backups*, recuperação da informação, identificadores persistentes baseados em padrões fixos de dados, conversão e preservação de metadados. O DVN propicia a criação dos chamados *dataverses*. Um *dataverses* pode ser um arquivo *web*, ou repositório utilizado para armazenar e compartilhar dados científicos. O desenvolvimento de *software* DVN começou em 2006 no Instituto de Ciências Sociais Quantitativas (IQSS), na Universidade de Harvard. O conceito que rege a implementação do DVN é o de replicação de dados, ou seja, um *dataverse* deve conter a informação necessária para reproduzir um estudo original de modo a proporcionar uma análise empírica sobre o processo exato de como os dados de pesquisa foram gerados e/ou produzidos (KING, 1995). |
| **CKAN** [74] | CKAN é uma ferramenta para criar repositórios de dados abertos. É utilizado para gerenciar e publicar coleções de dados. É bastante usado por instituições de pesquisa e outras organizações que coletam dados. CKAN é um *software* de código aberto, com uma comunidade ativa de colaboradores que o desenvolvem e mantêm uma crescente biblioteca de extensões CKAN. |

Fonte: DSPACE, e-PRINTS, FEDORA, DATAVERSE, CKAN (RODRIGUES, et al. (2010, p. 22).

Um repositório de dados de pesquisa pode estar disponível em duas classes de provedores: os de dados (*Data Providers*) e os de serviços (*Service Providers*). Os provedores de dados mantêm os repositórios de documentos digitais e implementam protocolos como

[73] DATAVERSE. Disponível em: <https://dataverse.org>. Acesso em: ago. 2017.
[74] CKAN. Disponível em: <https://ckan.org/>. Acesso em: ago. 2017.

*Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH)[75], para disponibilizar seus metadados em rede. Os provedores de serviços coletam dados para construção de serviços de valor agregado aos dados, oferecendo buscas a metadados e/ou outros serviços. Em qualquer um dos casos, o repositório de dados deve ser o mais interoperável possível (OPEN ARCHIVES INITIATIVE, 2002; GARCIA; SUNYE, 2006; RICE,; SOUTHALL, 2016).

Alguns métodos e tecnologias garantem a interoperabilidade. Métodos como *Application Programming Interface* (API) proporcionam o intercâmbio de informações entre os repositórios de dados. As APIs são sistemas que permitem, por meio de um conjunto de rotinas e padrões de programação, que um aplicativo de *software* tenha acesso a outro *software* baseado na *web* e reutilize dados e/ou outras informações. Exemplos desse tipo de interface, conforme explanado por Rice e Southall (2016), são o serviço de identificador persistente (PID) do re3data; o índice de citação de dados da Thomson Reuters, usado para rastrear citações de conjuntos de dados na literatura; e o Open Research e Contributor ID (ORCID), que é um nome digital permanente para identificação de cientistas. O código tem 16 dígitos e é único para cada individuo. O ORCID impede ambiguidades na identificação de autores e colaboradores em publicações. Seu funcionamento é semelhante ao do Digital Object Identifier (DOI), código para objetos como artigos científicos, teses e dissertações (RICE; SOUTHALL, 2016).

Uma API pode ser fornecida apenas para parceiros de um *site* ou pode ser disponibilizada publicamente. Embora o uso de APIs seja algo generalizado, não é uma maneira baseada em padrões de troca de informações como o OAI-PMH, isso porque cada sistema pode utilizar seu próprio conjunto de instruções e esquema de dados. As formas baseadas em padrões incluem (OAI -PMH) e outros padrões de dados abertos vinculados, como Resource Description Framework (RDF). O OAI-PMH é um padrão específico para repositórios que se baseiam no padrão de metadados *Dublin Core*[76] para permitir a colheita de registros (RICE; SOUTHALL, 2016).

A busca de dados de pesquisa na *web*, seja por APIs, seja via protocolos, exige a compreensão de técnicas de recuperação de informação, como o *web scraping e* o *harvesting* de metadados.

---

[75] Disponível em: <https://www.openarchives.org/pmh/>. Acesso em: maio 2016.

[76] DUBLIN CORE. Disponível em: <http://dublincore.org/>. Acesso em: set. 2017.

### 2.5.3 *Software* para recuperação de informação

Uma máquina de busca é baseada em *software* para recuperação de informação na *web*. Baseia-se na arquitetura coletor-indexador. Por coletor, Baeza-Yates e Ribeiro-Neto (2013, p. 421) entendem "[...] os programas ou agentes de *software* que percorrem a *web* enviando páginas novas ou atualizadas a um servidor principal onde elas são indexadas [...] São os chamados *robots*, *spider*, *crawlers*, *knowbots* [...]".

A principal utilização de um coletor é na criação de um índice que cubra tópicos abrangentes (geral) ou específicos (vertical). Os coletores também são usados para arquivar conteúdo, para analisar *sites* automaticamente e para extração de estatísticas agregadas (BAEZA-YATES; RIBEIRO-NETO, 2013).

Geralmente, um coletor vertical é usado para agregar dados de diferentes fontes, usualmente fontes de material similar ou relacionado. Comumente, esses coletores são utilizados para baixar informações de catálogos de compras *on-line* e fornecer uma interface para comparação de preços. Um coletor vertical também pode adotar formatos de dados específicos, ou seja, criar coleções restritas a um formato de dados em particular. Como, por exemplo, um coletor de *feeds*, que é um tipo de agregador de conteúdo baseado na *web* (BAEZA-YATES; RIBEIRO-NETO, 2013).

Uma outra forma de se extraírem dados da *Internet* é por meio de um *harvest,* que é um *software* de arquitetura distribuída, utilizado para coletar dados. São dois os principais componentes que estruturam essa arquitetura de busca de dados: os coletores e os *brokers*. Um coletor recolhe e extrai informação de indexação, a partir de um ou mais servidores *web*. Os momentos de coleta são definidos pelo sistema e são periódicos. Um *broker* provê o mecanismo de indexação e a interface de consulta para os dados coletados. Os *brokers* recuperam informação a partir de um ou mais coletores ou outros *brokers* (BAEZA-YATES; RIBEIRO-NETO, 2013).

Segundo explicam Baeza-Yates e Ribeiro-Neto (2013, p. 411), "[...] um dos objetivos do *harvest* é construir *brokers* específicos sobre tópicos focando nos conteúdos do índice para evitar problemas de vocabulário e escalonamento dos índices genéricos". A arquitetura *harvest* provê replicadores. Um replicador pode ser usado para replicar servidores e/ou suas coleções de dados.

Uma maneira comum de se extraírem dados da *Internet* por meio de máquinas de busca é o *web scraping*, método que se define como

processo de extração e combinação de conteúdos *web* de forma sistemática e automatizada. Em tal processo, um agente de *software*, *scraping*, simula o comportamento de navegação humano em servidores *web* copiando e reorganizando dados desorganizados em dados organizados (GLEZ-PENA et al., 2013).

Existem diferentes técnicas de busca para *web scraping*: a manipulação de HTTP[77], a mineração de dados, as ferramentas de *scraping*, a cópia manual e os microformatos. A manipulação de HTTP permite a colheita de dados estáticos e dinâmicos de um *site* através de uma solicitação HTTP. A mineração de dados é um processo automático que reconhece as informações de um *site* de acordo com *scripts* predefinidos que contêm dados incorporados. Ferramentas de *scraping* são *softwares* utilizados para extrair informações relacionadas a *sites* ou a funcionalidades e/ou estruturas de dados na *web*. As ferramentas de *scraping* também servem para extrair dados de redes sociais e são úteis para todos os tipos de atividades de *marketing web*. A cópia manual, embora existam ferramentas de *scraping* disponíveis, às vezes é necessária, por exemplo, em alguns casos em que a informação do *site* for bloqueada contra qualquer forma *web scraping*. Por fim, os microformatos são referentes às tecnologias da *web* semântica, como dados abertos vinculados via RDF. Esses são conjuntos de informações geralmente elaboradas para serem intercambiadas via vocabulários e ontologias (WEBSTER, 2015)[78].

O funcionamento de um *web scraping* ocorre da seguinte maneira: (1) acesso ao *site* via protocolo de comunicação *web* (HTTP), durante o processo de solicitação e resposta entre um cliente, normalmente um navegador da *web*, e um servidor da web; (2) o programa de *scraper* analisa o HTML e pode extrair seu conteúdo; (3) armazena o *output* em conteúdo textual, com o objetivo de transformar o conteúdo extraído em uma representação estruturada para posterior análise e armazenamento (GLEZ-PENA et al., 2013).

As técnicas de extração de dados da *web* podem ser aplicadas por meio da utilização de *softwares* que funcionam como ferramentas de *scraping*. Essas ferramentas transformam os dados desestruturados da *web* em representação estruturada, como os tipos de arquivos em *Comma Separated Values* (CSV) e/ou *eXtensible Markup Language* (XML). Elas interagem com *sites* da mesma forma que os navegadores

---

[77] Hypertext Transfer Protocol (HTTP) é o protocolo utilizado para enviar e receber informações na *web*.

[78] Disponível em: <https://myhelpster.com/what-is-scraping-the-basics-for-everyone/>. Acesso em: set. 2017.

*web*, Chrome ou Mozilla. Existem muitas dessas ferramentas disponíveis na *web*, como import.io[79], Fminer[80], Octoparse[81], OutWit Hub[82], Web Scraper Chrome Extension[83] e ParseHub[84].

O Quadro 7 descreve três ferramentas de *web scraping* que possuem versão de *software desktop* para Windows, Mac, Linux e Web (cloud). Não exigem conhecimentos em programação, no entanto, são integradas com linguagens de programação (Python e R).

Quadro 7 - Softwares para *web scraping*.

| SOFTWARE | ESPECIFICAÇÃO | OS | LICENÇA |
|---|---|---|---|
| **OUTWIT HUB** | É um aplicativo de *software* projetado para extrair automaticamente informações, dados e outros recursos *on-line* ou locais. Reconhece e captura de *links*, imagens, documentos, contatos, vocabulário recorrente em frases, *feeds* em RSS, converte dados estruturados e não estruturados em tabelas formatadas que podem ser exportadas para planilhas ou bancos de dados. O programa inclui uma extensão gratuita para o navegador *web* Mozilla. | MAC<br>WIN<br>LINUX<br>WEB | Free<br>Versão Pro é paga. |
| **Web Scraper Chrome Extension** | É uma extensão do navegador Chrome criada para extração de dados de páginas da *web*. Permite a criação de *sitemaps*; o uso desses mapas de *site* permite extrair os dados da navegação que podem ser exportados em CSV. | MAC<br>WIN<br>LINUX<br>WEB | Free |

[79]IMPORT.IO. Disponível em: <https://www.import.io/>. Acesso em: set. 2017.
[80] FMINER. Disponível em: <http://www.fminer.com/>. Acesso em: set. 2017.
[81] OCTOPARSE. Disponível em: <*https://www.octoparse.com/*>. Acesso em: set. 2017.
[82] OUTWIT HUB. Disponível em: <http://www.outwit.com/>. Acesso em: set. 2017.
[83] Disponível em: <https://chrome.google.com/webstore/detail/web-scraper/jnhgnonknehpejjnehehllkliplmbmhn>. Acesso em: set. 2017.
[84] PARSEHUB. Disponível em: <https://www.parsehub.com/>. Acesso em: set. 2017.

| | | | |
|---|---|---|---|
| **ParseHub** | É um *software* de *web scraping* que suporta extração de dados em *sites* que usam AJAX. Utiliza a tecnologia de *learning machine* para ler e analisar documentos na *web*. Possui versão para usar no navegador; nessa versão, é possível gerar até 5 projetos de *web scraping* com o plano gratuito. | MAC<br>WIN<br>LINUX<br>WEB | Free<br>Versão Pro é paga. |

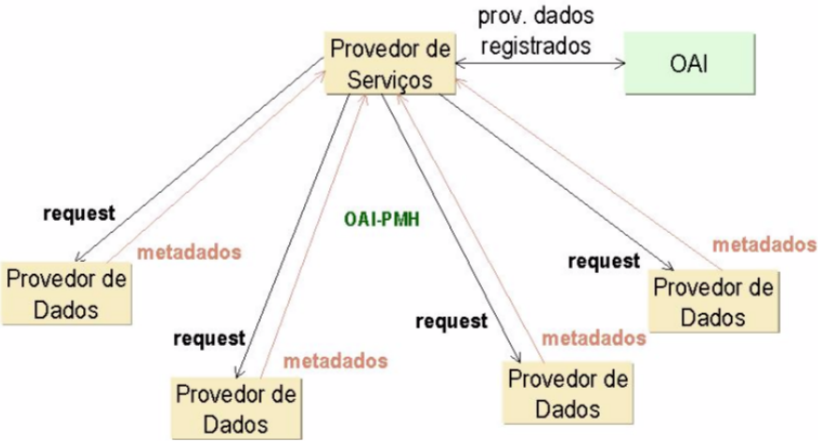
Fonte: OUTWIT HUB; WEB SCRAPER CHROME EXTENSION; PARSEHUB (2017).

Após um breve panorama sobre APIs, máquinas de busca e alguns *softwares* de *web scraping*, apresenta-se *o harvesting* de metadados como arquitetura baseada em padrões de troca de informações e norteadora para o processo de coleta e aquisição de dados em repositórios de dados de pesquisa interoperáveis que utilizam *softwares* como o *DSpace,* o *e-Prints,* o *Fedora,* o *Dataverse* e o *CKAN*.

### 2.5.4 *Harvesting* de metadados

O *harvesting* de metadados pode ser operacionalizado por meio da utilização de protocolos e técnicas de interoperabilidade. Protocolos como OAI-PMH norteiam os processos de busca, extração, descrição e organização de metadados, segundo o padrão *Dublin Core*. Este último é um dos mais utilizados em bibliotecas digitais, periódicos e repositórios. Segundo Garcia e Sunye (2006), é possível realizar buscas periódicas em provedores de serviços coletando seus metadados via *request* no protocolo OAI-PMH, conforme o esquema de funcionamento apontado na Figura 5.

Figura 5 - *Harvesting* de metadados.




Fonte: Garcia; SUNYE (2006).

O protocolo OAI-PMH permite coletar todas as coleções de um provedor de dados ou uma em específico. O protocolo OAI-PMH executa requisições HTTP *(Hypertext Transfer Protocol*) baseadas em *request-response*.

As respostas para as solicitações (*requests*) são em formato XML. Os verbos de requisição (comandos transmitidos aos repositórios que utilizam protocolo OAI-PMH) são: *ListIdentifier*, *ListMetadataFormats*, *ListRecord, GetRecord*, *ListSets*. O primeiro *request* é utilizado para recuperar as informações sobre o provedor de dados; o segundo, para recuperar os formatos de metadados disponíveis no repositório; o terceiro recupera um único registro do repositório; o quarto coleta todos os metadados do repositório; o quinto, uma versão abreviada do verbo *ListRecords*, retorna apenas os cabeçalhos dos registros; e, por fim, o *ListSets* lista a estrutura do conjunto de coleções de dados de um repositório (OPEN ARCHIVES, 2002; GARCIA; SUNYE, 2006).

O comando *request?verb=ListSets* serve para identificar qual o código (*setName)* das coleções de cada repositório. Depois de identificar o *Set* específico da coleção, utiliza-se o comando *ListRecords* para extrair a coleção. Nesse contexto, pode-se coletar toda uma base de dados ou uma coleção específica.

Os arquivos de *output*, em formato XML, são facilmente convertidos para outros formatos, como *Comma Separated Values*

(CSV), *JavaScript Object Notation* (JSON) ou *Structured Query Language* (*SQL)*. A coleta de dados em vários provedores é central para a geração de serviços de dados, funcionalidades e replicação de dados dispersos em repositórios.

### 2.5.5 *Software* para manipulação de dados (*Data Handling*)

Os dados, após recuperados em uma fonte, como um repositório de dados de pesquisa, geralmente, necessitam ser ajustados para atenderem aos requisitos de manipulação de dados (pré-análise e processamento de dados para pós-análise). A manipulação de conjuntos de dados primários para secundários e/ou analíticos necessita a incorporação de alguns *softwares* para manipulação de dados que realizem a limpeza e o tratamento adequado aos dados. O Quadro 8 apresenta uma lista de *softwares* utilizados manipulação de dados.

Quadro 8 - *Softwares* para *Data handling.*

| SOFTWARE | DESCRIÇÃO | OS | LICENÇA |
|---|---|---|---|
| **TABLEAU PUBLIC**[85] | O Tableau Public é um *software* gratuito que permite gerar visualizações de dados interativos na *web*. As visualizações no Tableau Public (*vizzes*) podem ser incorporadas a *sites* e *blogs*, compartilhadas em redes sociais, por *e-mail*, e disponibilizadas para *download*. As visualizações são criadas via aplicativo Tableau Desktop Public Edition, não são necessários conhecimentos em programação. | MAC<br>WIN | Free<br>Versão Pro é paga. |

[85] TABLEAU PUBLIC. Disponível em: <http://www.tableausoftware.com/public/>. Acesso em: set. 2017.

| | | | |
|---|---|---|---|
| **OPENREFINE**[86] | OpenRefine, originalmente conhecido como GoogleRefine, é um *software* de limpeza de dados que permite a preparação de dados para pré-análise. O OpenRefine contém uma série de algoritmos de agrupamento e *clastering*. É uma ferramenta de *softwar*e utilizada para trabalhar com dados desorganizados. Aplica-se para dar consistência aos dados, não é um serviço *web*, mas um aplicativo *desktop* usado para processar dados. | MAC<br>WIN<br>LINUX | Free |
| **KNIME**[87] | Permite manipular, analisar e modelar dados de forma intuitiva por meio de programação visual. O KNIME pode ser estendido para executar R e Python. Permite manipulação de números para imagens, moléculas para seres humanos, sinais para redes complexas e estatísticas simples para análises de *big data*. | MAC<br>WIN<br>LINUX | Free<br>Versão Pro é paga. |
| **RAPIDMINER**:[88] | Plataforma de código aberto para análise preditiva, possibilita o *mashup* de dados, a criação de modelos preditivos e a operacionalização de análises preditivas dentro | MAC<br>WIN<br>LINUX | Free<br>Versão Pro é paga. |

[86] OPENREFINE. Disponível em: <http://openrefine.org/>. Acesso em: set. 2017.
[87] KNIME. Disponível em: <http://www.knime.org/>. Acesso em: set. 2017
[88] RAPIDMINER. Disponível em: <http://sourceforge.net/projects/rapidminer/>. Acesso em: set. 2017.

| | | | |
|---|---|---|---|
| | de qualquer processo. | | |
| **GOOGLE FUSION TABLES**[89] | O Google Fusion é uma ferramenta para análise, visualização e mapeamento de dados. Reconhece dados de latitude e longitude do Google Maps e outros tipos de dados e/ou serviços oferecidos pelo Google. | web | Free |
| **ORANGE**[90] | É um *software* para mineração de dados baseado em componentes. Inclui uma gama de técnicas para visualização, exploração, pré-processamento e modelagem de dados. Pode ser usado por meio de uma interface gráfica ou por meio de módulo de programação em Python. | MAC<br>WIN<br>LINUX | Free |

Fonte: TABLEAU PUBLIC; OPENREFINE, KNIME; RAPIDMINER; GOOGLE FUSION TABLES; ORANGE (2017).

A manipulação de dados é uma tarefa essencial em *Data Science* e pode ser desenvolvida com as ferramentas apresentadas. No entanto, acredita-se que o ideal é utilizar *softwares* que integrem suas funcionalidades de manipulação com funcionalidades de análise passíveis de serem implementadas via linguagem de programação, o que de certa forma irá facilitar o processo de análise e interoperabilidade dos dados. Assim, é desejável que o *software* adotado seja capaz de aplicar técnicas e/ou algoritmos utilizados em *Data Science* por meio de interface gráfica e ou linguagem de programação.

[89]GOOGLE FUSION TABLES. Disponível em: <https://support.google.com/fusiontables/answer/2571232>. Acesso em: set. 2017.
[90] ORANGE. Disponível em: <https://orange.biolab.si/>. Acesso em: set. 2017.

### 2.5.6 Linguagens de programação para manipulação e análise de dados

A escolha por uma linguagem de programação depende do tipo de *software* científico que será desenvolvido. Por exemplo, a elaboração de um *web scraping* é geralmente desenvolvida em *Python*. Já a elaboração de uma análise estatística pode ser realizada em R. Cada tipo de linguagem de programação permite a utilização de um ou mais algoritmos. No caso de coleta, análise e visualização de dados oriundos de repositórios de dados de pesquisa, é possível combinar as duas linguagens.

O *site* Data Science Central[91] apresentou um artigo de Granville (2017), *Python Overtakes R for Data Science and Machine Learning,* que resume a tendência de uso de linguagens de programação por parte dos cientistas de dados, com base em uma série de métricas de *proxy* utilizando o Google Trend[92]. O artigo conclui que as linguagens mais utilizadas em *Data Science* nos USA são o Python e o R. Assim, torna-se essencial apresentar breves definições sobre essas linguagens de programação.

#### 2.5.6.1 *R-project*

O R é uma linguagem, um ambiente para computação estatística e uma interface para o desenvolvimento de gráficos. É um projeto de *software* livre criado em 1990 por Ross Ihaka e Robert Gentleman, ambos professores da Universidade de Auckland, na Nova Zelândia. O R é semelhante à linguagem S, também uma linguagem de programação voltada para cálculos estatísticos que foi desenvolvida na Bell Laboratories por John Chambers. Os criadores do R fundamentaram-se na linguagem S e produziram o R para dar aulas de estatísticas em Auckland. O R tem seu nome devido às iniciais de seus idealizadores, Ross e Robert, como uma menção à linguagem S, já que R é a próxima letra do alfabeto seguindo S (R-PROJECT[93], 200?; FOX, 2009).

O R é um conjunto integrado de pacotes de *software* para manipulação de dados, cálculos e exibição gráfica. O que inclui ferramentas para:

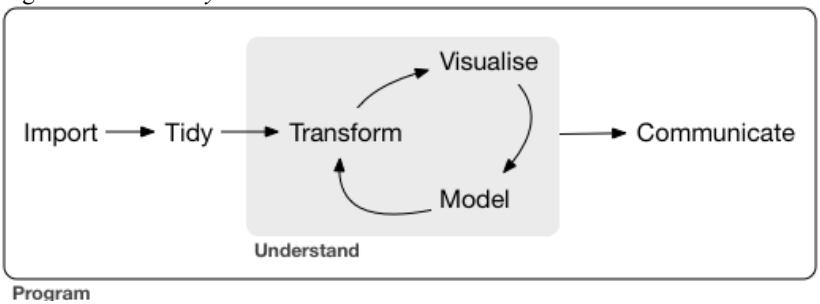
---

[91] DATA SCIENCE CENTRAL. Disponível em: <http://www.datasciencecentral.com/>. Acesso em: ago. 2017.
[92] GOOGLE TREND. Disponível em: <https://trends.google.com/trends/>. Acesso em: ago. 2017.
[93] R-PROJECT. Disponível em: <https://www.r-project.org/>. Acesso em: ago. 2017.

> [...] o efetivo tratamento e armazenamento de dados; para manipular conjuntos de operadores de cálculo de matrizes matemáticas; para coleta de ampla quantidade de dados, uma ferramenta para análise de dados; com interface gráfica para análise e exibição de dados em tela ou em cópia impressa; é uma linguagem de programação bem desenvolvida, simples e eficaz, que inclui condicionais, loops, funções recursivas definidas pelo usuário e na forma de entrada e saída. (R-PROJECT, 200?, online, tradução nossa)[94]

Alguns dos principais componentes do R utilizados em *Data Science* podem ser instalados por meio do pacote *Tidyverse*[95], que significa organizar, arrumar, limpar e deixar dados em ordem. Esse pacote foi criado por Hadley Wickham[96], cientista de dados chefe do projeto RStudio[97], e contém o seguinte fluxo (Figura 6):

Figura 6 - Fluxo *Tidyverse*.




Fonte: GROLEMUND; WICKHAM (2017).

O fluxo do pacote Tidyvers é descrito da seguinte maneira: primeiro, *Tidyvers* importa (import) dados para R, o que geralmente

[94] *"[...] an effective data handling and storage facility; a suite of operators for calculations on arrays, in particular matrices; a large, coherent, integrated collection of intermediate tools for data analysis; graphical facilities for data analysis and display either on-screen or on hardcopy, and; a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities"* (R-PROJECT, 200?, online).

[95] O Tidyverse instala os seguintes pacotes para o R: *broom, dplyr, feather, forcats, ggplot2, haven, httr, hms, jsonlite, lubridate, magrittr, modelr, purrr, readr, readxl, stringr, tibble, rvest, tidyr,* xml2. (GROLEMUND; WICKHAM, 2017).

[96] Disponível em: <http://r4ds.had.co.nz/index.html>. Acesso em: set. 2017.

[97] RSTUDIO. Disponível em: <https://www.rstudio.com/>. Acesso em: set. 2017.

significa extrair informações de um arquivo, banco de dados ou API da *web*. Depois de importar os dados (Tidy), é preciso ajustá-los, o que significa armazená-los de forma consistente, ou seja, cada coluna é uma variável, e cada linha é uma observação. Após ajustar os dados, é preciso transformá-los. A transformação inclui criar observações de interesse, novas variáveis. No que segue, existem dois motores principais de geração de conhecimento: a visualização e a modelagem (GROLEMUND; WICKHAM, 2017).

A visualização é uma atividade fundamentalmente humana. Uma boa visualização irá mostrar se a pergunta feita aos dados está correta ou errada, ou se é preciso coletar dados diferentes. Os modelos são ferramentas complementares para a visualização. Depois de ter feito suas perguntas suficientemente precisas, pode-se usar um modelo para respondê-las. Os modelos são uma ferramenta fundamentalmente matemática ou computacional, cada modelo faz suposições; um modelo não pode questionar seus próprios pressupostos. O último passo é a comunicação; ela é uma parte absolutamente crítica de qualquer projeto de análise de dados. Não importa o quão bem os modelos e a visualização levam a entender os dados, a menos que se possa comunicá-los para outros. Ao redor de todas essas ferramentas está a programação, uma ferramenta de corte transversal que se utiliza em todas as etapas do projeto *Tidyvers* (GROLEMUND; WICKHAM, 2017).

#### 2.5.6.2 *Python*

A linguagem *Python* foi criada por Guido van Rossum em 1991, nos USA. Suporta múltiplos paradigmas de programação: procedimental, orientação a objetos e suporte à programação funcional. Também existem muitos módulos para lidar com dados científicos em *Python*. Os principais pacotes *Python* utilizados para aplicação em projetos de *Data Science* são:

a) *Beautiful Soup*[98]: uma classe utilizada para extrair dados de arquivos HTML e XML;
b) CSV[99]: pacote de dados que implementa classes para ler e gravar dados tabulares em formato CSV. O *Comma*

[98] Disponível em: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>. Acesso em: maio 2016.
[99] Disponível em: <https://docs.python.org/2/library/csv.html>. Acesso em: maio 2016.

*Separated Values* é o formato de importação e exportação mais comum para planilhas e bancos de dados;

*c)* LXML[100]: é um *toolkit* para compilar XML; utiliza classes da linguagem C, como libxml2 e libxlst. Combina a funcionalidade e a completude dessas classes com a simplicidade de uma *API Python* nativa, em sua maioria compatível com os padrões XML;

*d)* *Scipy*[101]: classe científica com diversos módulos, incluindo funções estatísticas. Implementa diversas técnicas úteis para computação científica. Utiliza *NumPy* para lidar com algoritmos numéricos e simbólicos que realizam o processamento matemático;

*e)* *NumPY*[102]: classe básica da linguagem *Python* que permite trabalhar com arranjos, vetores e matrizes de *n* dimensões, de forma comparável e com sintaxe semelhante ao *software* proprietário *Matlab*;

*f)* *Bokeh*[103]: classe da linguagem comumente utilizada para *interactive visualization*. Ela tem como objetivo a geração de representações gráficas para navegadores *web*. *Bokeh* pode ser compilada em diferentes linguagens de programação (*Python,* R e Julia). A classe *Python* permite construir gráficos estatísticos complexos de forma rápida e simples.

*g)* *Pandas*[104]: classe de alta performance para análise e geração de estruturas de dados complexos.

*h)* *Seaborn*[105]: classe para visualização com base na classe *matplotlib*. Fornece uma interface de alto nível para desenhar gráficos estatísticos. Integrada com a *PyData*, incluindo suporte para estruturas *NumPy*, *Pandas* e outras rotinas estatísticas de *scipy* e *statsmodels*.

i) *Matplotlib*[106]: classe de plotagem 2D que produz figuras com qualidade para publicação em uma variedade de formatos impressos e ambientes interativos. *Matplotlib* pode ser usado em *scripts Python* e servidores *web*. Pode

---

[100] Disponível em: <http://lxml.de>. Acesso em: maio 2016.
[101] Disponível em: <https://www.scipy.org>. Acesso em: maio 2016.
[102] Disponível em: <http://www.numpy.org>. Acesso em: maio 2016.
[103] Disponível em: <http://bokeh.pydata.org/en/latest/>. Acesso em: maio 2016.
[104] Disponível em: <http://pandas.pydata.org/>. Acesso em: maio 2016.
[105] Disponível em: <https://web.stanford.edu/~mwaskom/software/seaborn/>. Acesso em: maio 2016.
[106] Disponível em: <http://matplotlib.org>. Acesso em: maio 2016.

gerar gráficos, histogramas, gráficos de barras e gráficos de dispersão.

Acredita-se que esses pacotes da linguagem de programação *Python* são as ferramentas ideais para elaboração de serviços de dados que contemplem os processos de coleta, análise e visualização de dados, fundamentais ao desenvolvimento de um projeto com base no processo de *Data Science*.

Essa linguagem, além de ser amplamente utilizada em *Data Science*, servirá como ferramenta para automatização dos processos de coleta, análise e visualização de dados. Pode ser possível utilizar o *Python* como instrumento de pesquisa aplicada, como *software* científico e ferramenta metodológica de experimentação e elaboração criativa. Assim, uma forma de testar tal aplicação é verificar se os modelos e as técnicas aplicados em *Data Science* podem ser traduzidos para *Python* e/ou *R* e aplicados ao contexto de estudo desta tese, seja diretamente por meio do uso dessas linguagens, seja indiretamente, por meio do uso de *softwares* de interface gráfica que fazem o uso delas.

# 3 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

O objetivo deste capítulo é desenvolver os objetivos específicos (c) realizar uma análise descritiva sobre a composição internacional dos RDPGs; e (d) identificar o corpus de temáticas, países, licenças, *softwares,* instituições e tipos de dados de pesquisa mais utilizados pela área de Geociências. Para isso, apresenta um estudo descritivo sobre o cenário internacional dos RDPGs disponíveis no Registry of Research Data Repositories (re3data) (**ver capítulo 1.1**).

## 3.1 CENÁRIO INTERNACIONAL DOS REPOSITÓRIOS DE DADOS DE PESQUISA EM GEOCIÊNCIAS (RDPGs)

Os repositórios de dados de pesquisa são utilizados para proporcionar o acesso e a preservação dos dados de pesquisa. Neles é possível disponibilizar, compartilhar e fornecer acesso em longo prazo aos dados.

Em específico, busca responder às perguntas lançadas no problema de pesquisa (**ver capítulo 1.3)**, como: quantos repositórios de dados registrados no re3data são sobre Geociências? Quais são as principais instituições? Quais os países e em quais línguas os dados de pesquisa estão disponíveis? Qual a tipologia desses repositórios? Quais tipos de dados são depositados? Quais tipos de acesso e licenças utilizadas? Quais as temáticas e as áreas do conhecimento? E qual o tipo de *software*? Sempre que possível, as respostas apresentarão um RDPGs como exemplo.

Não é objetivo deste estudo analisar todo o *output* dos repositórios de dados de pesquisa registrados no re3data, mas sim propor uma pesquisa sobre o domínio das Geociências e avaliar como os RDPGs elaborados com *Ckan, Dspace, Dataverse, E-Prints* e *Fedora* estão representados no re3data (**ver capítulo 2.5.2**).

A composição internacional dos RDPGs pode ser mapeada por meio do uso do re3data. No total, são (1905) repositórios de dados de pesquisa registrados em (set. 2017), dos quais (604) são catalogados na área das Geociências[107] .

A seguir, apresenta-se a distribuição de RDPGs por países e instituições que possuem coleções de dados de pesquisa em

[107] Lista de Repositórios de dados de Pesquisa em Geociências. Disponível em: < http://www.re3data.org/api/beta/repositories?query=&subjects[]=34%20Geosciences%20(including%20Geography) >. Acesso em: ago 2017.

Geociências, com a ocorrência mínima de (2) RDPGs por instituição (Tabela 1).

Tabela 1 - Países e instituições que possuem RDPGs.

| PAIS | INSTITUIÇÃO | FREQUÊNCIA |
|---|---|---|
| DEU | Helmholtz Centre Potsdam - GFZ German Research Centre for Geosciences | 12 |
| USA | Woods Hole Oceanographic Institution | 11 |
| CAN | Environment and Climate Change Canada | 10 |
| USA | National Science Foundation | 9 |
| USA | Columbia University, Lamont-Doherty Earth Observatory | 8 |
| USA | National Oceanic and Atmospheric Administration | 8 |
| USA | U.S. Geological Survey | 7 |
| USA | National Aeronautics and Space Administration | 6 |
| CAN | Government of Canada | 5 |
| GBR | Natural Environment Research Council | 4 |
| USA | Columbia University, Lamont -Doherty Earth Observatory | 4 |
| AAA [108] | International Council for Science, World Data System | 3 |
| CAN | Canadian Space Agency | 3 |
| CAN | Natural Resources Canada | 3 |
| EEC | European Space Agency | 3 |
| GBR | British Geological Survey | 3 |
| NOR | Norwegian Institute for Air Research | 3 |
| USA | Jet Propulsion Laboratory | 3 |
| USA | University Corporation for Atmospheric Research | 3 |
| AAA | FigShare | 2 |
| AUS | Australian Government, Australian Institute of Marine Science | 2 |

[108] Sigla para internacional.

| | | |
|---|---|---|
| AUS | Australian National Data Service | 2 |
| AUS | Government of South Australia, Department of State Development | 2 |
| CHE | World Meteorological Organization, Global Atmosphere Watch | 2 |
| DEU | Alfred Wegener Institute - Helmholtz Centre for Polar and Marine Research | 2 |
| DEU | Helmholtz Centre for Environmental Research | 2 |
| FRA | University of Lille, ICARE Data and Services Center | 2 |
| GBR | Centre for Ecology & Hydrology | 2 |
| JPN | National Institute of Polar Research, Polar Data Center | 2 |
| NOR | Norwegian Meteorological Institute | 2 |
| RUS | All-Russian Research Institute of Hydrometeorological Information - World Data Center | 2 |
| USA | Columbia University, Earth Institute, Lamont-Doherty Earth Observatory | 2 |
| USA | Cornell University, Albert R. Mann Library | 2 |
| USA | Interdisciplinary Earth Data Alliance | 2 |
| USA | NASA Goddard Space Flight Center | 2 |
| USA | National Aeronautics and Space Administration, Earth Observing System Data and Information System | 2 |
| USA | National Renewable Energy Laboratory | 2 |
| USA | UNAVCO | 2 |
| USA | United States Geological Survey | 2 |
| USA | University of Alaska Fairbanks, Geophysical Institute | 2 |

Fonte: Dados da Pesquisa (2017).

As instituições que mais disponibilizam coleções de dados de pesquisa em Geociências são: o GFZ German Research Centre for Geosciences no Helmholtz Centre Potsdam, na Alemanha, com (12) RDPGs; a Woods Hole Oceanographic Institution nos USA, com (10)

RDPGs; e o Environment and Climate Change Canada, no Canadá, com (9) RDPGs.

O GFZ Research Centre for Geosciences disponibiliza diversas coleções de dados de pesquisa por meio da Information System and Data Center (ISD)[109], um portal de serviços de dados geocientíficos, metadados, documentação e ferramentas de *software* para prática com dados de pesquisa geocientíficos. O ISD fornece acesso a coleções de dados coletados por sensores e satélites. Como exemplo, citam-se os dados da CHAllenging Minisatellite Payload (CHAMP)[110] , que foi uma missão espacial alemã lançada na década de 2000 para coletar dados geocientíficos da atmosférica terrestre. Outra coleção desse mesmo centro é o International Geodynamics and Earth Tide Service (IGETS)[111], que disponibiliza dados sobre as variações temporais do campo de gravidade da Terra. O centro também oferece os dados da Gravity Recovery and Climate Experiment (GRACE)[112], que é um projeto entre a National Aeronautics and Space Administration (NASA) e a GFZ para gerar dados sobre o campo de gravitacional da Terra e medir os reservatórios de água no Planeta em solo, no gelo e nos oceanos.

A segunda instituição que mais disponibiliza RDPGs com coleções de dados de pesquisa em Geociências nos USA é o Woods Hole Oceanographic Institution. O instituto contribui com conjuntos de dados oceanográficos, como o WHOI Ship Data-Grabber System[113], que é uma coleção de dados de pesquisa oceanográficos coletados de navios em curso. Os dados nesse RDPGs são corregistrados com horários, GPS e sistemas de navegação de navios. Além dessa coleção, o instituto disponibiliza o Seafloor Sediments Data Collection[114], uma coletânea de 14.000 amostras estratigráficas e sedimentos geológicos recuperados do fundo do mar.

No Canadá, o Environment and Climate Change disponibiliza compilações de dados do National Pollutant Release Inventory[115], um

---

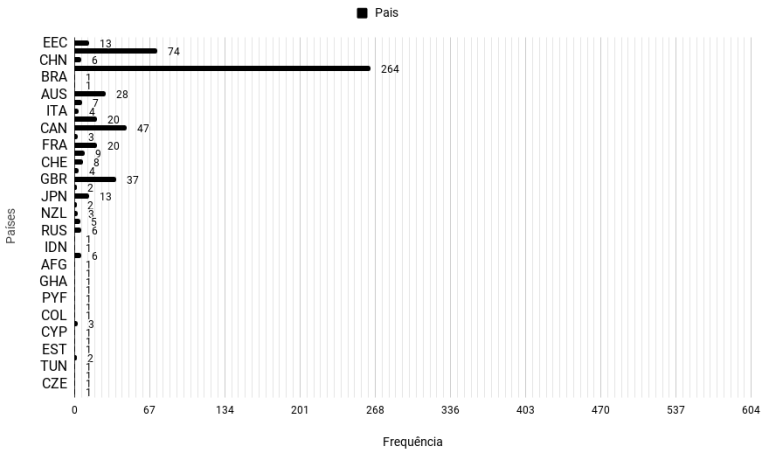
[109] ISD. Disponível em: <http://isdc.gfz-potsdam.de/homepage/>. Acesso em: set. 2017.
[110] CHAMP. Disponível em: <http://www-app2.gfz-potsdam.de/pb1/op/champ/index_CHAMP.html>. Acesso em: set. 2017.
[111] IGETS. Disponível em:<http://dataservices.gfz-potsdam.de/igets/showshort.php?id=escidoc:2387889>. Acesso em: set. 2017.
[112] GRACE. Disponível em: <http://isdc.gfz-potsdam.de/grace-isdc/> . Acesso em: set. 2017.
[113] WHOI Ship Data-Grabber System. Disponível em: <http://4dgeo.whoi.edu/shipdata/index.html />. Acesso em: set. 2017.
[114] SEAFLOOR SEDIMENTS DATA COLLECTION. Disponível em: <http://gis1server.whoi.edu/website/SC2k/viewer.htm >. Acesso em: set. 2017.
[115] NATIONAL POLLUTANT RELEASE INVENTORY. Disponível em: <http://www.ec.gc.ca/inrp-npri/>. Acesso em: set. 2017.

RDPGs que oferece coleções de dados de pesquisa sobre o lançamento de poluentes (no o ar, na água e na terra). Esses dados servem para identificar as prioridades de prevenção contra a poluição e para apoiar o gerenciamento de substâncias tóxicas e poluentes no ar, na água e na terra. Outro RDPGs canadense é o Canadian Centre for Climate Modelling and Analysis (CCCMA)[116], que oferece dados de pesquisa gerados pela utilização de modelos climáticos desenvolvidos pelo CCCMA. Esses dados são usados para estudar a variabilidade e as mudanças climáticas no Canadá.

Por fim, é relevante destacar que o Brasil possui 1 repositório listado no re3data, o Brazil Exploration and Production Database (BDEP)[117], da Agencia Nacional do Petróleo (ANP). O BDEP disponibiliza dados de pesquisa sobre bacias sedimentares; são mais de 6 petabytes de dados técnicos gerados nas atividades de prospecção petrolífera em todo o território brasileiro desde a década de 2000.

Para responder à outra pergunta, qual a distribuição geográfica desses repositórios e em que língua disponibilizam os dados de pesquisa?, apresentam-se o Gráfico 1 e a Figura 7.

Gráfico 1 – Países.




Fonte: Dados da Pesquisa (2017).

[116] CANADIAN CENTRE FOR CLIMATE MODELLING AND ANALYSIS. Disponível em: < http://www.ec.gc.ca/ccmac-cccma/>. Acesso em: set. 2017.

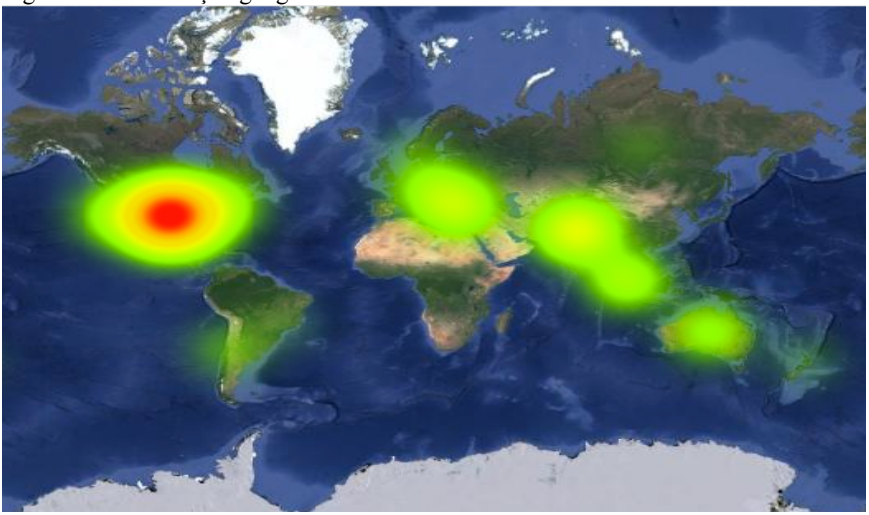
[117] BDEP. Disponível em: <http://www.anp.gov.br/wwwanp/exploracao-e-producao-de-oleo-e-gas/dados-tecnicos>. Acesso em: set. 2017.

No gráfico 1, os USA aparecem com (264) repositórios, sendo o país que mais possui coleções de dados de pesquisa. Elas estão distribuídas entre universidades (Columbia University, Cornell University, University of Alaska Fairbanks) e departamentos governamentais (NASA Goddard Space Flight Center, National Science Foundation, National Aeronautics and Space Administration, Earth Observing System e Woods Hole Oceanographic Institution).

No ranking de (42) países,[118] os que mais possuem RDPGs são: USA (264), DEU (74), CAN (47), Reino Unido, GBR (37), AUS (28), FRA (20), EEC (13), JPN (13), IND (9), CHE (8), NOR (7), CHN (6), NLD (6), RUS (6), SWE (5), AUT (4), ITA (4), BEL (3), DNK (3), NZL (3), GRL (2), IRL (2), UKR (2). Esses países concentram cerca de 90% dos RDPGs.

Na figura 7, percebe-se que os RDPGs estão concentrados na América do Norte e na Europa.

Figura 7 - Distribuição geográfica dos RDPGs.



Fonte: Dados da Pesquisa (2017).

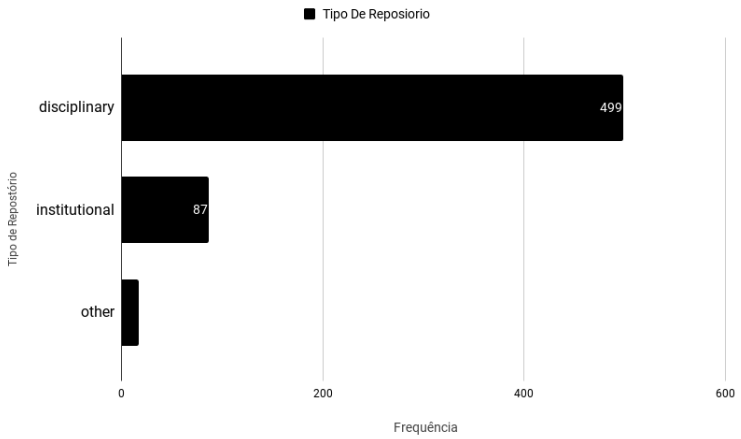
No que tange à língua em que os dados de pesquisa estão disponibilizados, o inglês é a principal (586), sendo apenas 16 desses

[118] COI. Disponevel em: <https://pt.wikipedia.org/wiki/Lista_de_pa%C3%ADses_por_c%C3%B3digo_do_COI>. Acesso em: set. 2017.

RDPGs em outras linguas como o alemão (12) e o francês (2) e apenas (1) em espanhol e (1) russo.

Outro ponto a se observar é a tipologia desses repositórios (Gráfico 2). Os repositórios de dados de pesquisa podem ser: institucionais, disciplinares, multidisciplinares e/ou resultantes de projetos de pesquisa específico (**ver capítulo 1.2**). Os principais tipos de RDPGs são os disciplinares (499) e os institucionais (87); (17) são classificados como de outro tipo.

Gráfico 2 – Tipologia dos repositórios




Fonte: Dados da Pesquisa (2017).

Como exemplo de repositório disciplinar pode-se citar o ShareGeo Open, que é um repositório de dados de pesquisas espaciais que oferece serviços de acesso gratuito a mapas e a dados geoespaciais de provedores de dados no Reino Unido. Foi desenvolvido como parte do projeto EDINA[119], feito para assegurar a continuidade de acesso aos dados científicos pelo setor acadêmico e educacional do Reino Unido. O ShareGeo Open é indexado pelo Thomson Reuters Data Citation Index[120]. A maioria dos formatos de arquivo nesta coleção de dados de

[119] "EDINA is a world-class centre for digital expertise, based at the University of Edinburgh.Working with partners to innovate, deliver and inspire, we transform new ideas into robust solutions" (EDINA, 2017, online). Disponível em: <https://edina.ac.uk/>. Acesso em: set. 2017.

[120] DATA CITATION INDEX. Disponível em : <http://wokinfo.com/products_tools/multidisciplinary/dci/>. Acesso em: set. 2017.

pesquisa pode ser visualizada e analisada usando-se pacotes de dados GIS[121].

O Edinburgh DataShare[122] é um exemplo de repositório de dados de pesquisa do tipo institucional. Ele é produzido pela Universidade de Edimburgo, Escócia, Reino Unido. O DataShare é mantido e hospedado pelo Departamento de Serviços de Informação da Universidade. Os pesquisadores da Universidade de Edimburgo que produzirem dados de pesquisa associados a publicações, como artigos ou relatórios científicos, são convidados a fazer o *upload* de seu conjunto de dados para compartilhamento e preservação no DataShare, no qual um identificador persistente é atribuído a cada conjunto de dados depositado.

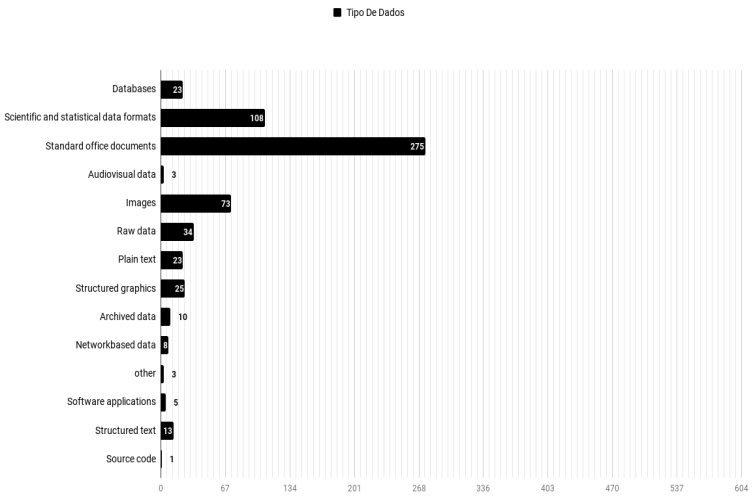
Além do tipo dos repositórios, é relevante apresentar a tipologia dos dados de pesquisa depositados nesses repositórios. Dados de pesquisa são evidências e/ou insumos de pesquisa que são coletados, observados, registrados e/ou criados para fins de análise e que podem produzir resultados de pesquisa para um estudo científico. No Gráfico 3, apresenta-se a tipologia dos dados de pesquisa em Geociências.

---

[121] "A geographic information system (GIS) is a system designed to capture, store, manipulate, analyze, manage, and present spatial or geographic data. The acronym GIS is sometimes used for geographic information science (GIScience) to refer to the academic discipline that studies geographic information system and is a large domain within the broader academic discipline of geoinformatics. What goes beyond a GIS is a spatial data infrastructure, a concept that has no such restrictive boundaries" (WIKPEDEA, 2017). Disponível em : < https://en.wikipedia.org/wiki/Geographic_information_system>. Acesso em: set. 2017.

[122] EDINBURGH DataShare. Disponível em: <http://datashare.is.ed.ac.uk/>. Acesso em: set. 2017.

Gráfico 3 – Tipos de dados de pesquisa




Fonte: Dados da Pesquisa (2017).

Alguns RDPGs disponibilizam dados de tipos especiais, como The Northern California Earthquake Data Center (NCEDC)[123] , que dissemina múltiplos tipos de dados relacionados à predição de terremotos no estado da Califórnia, nos USA. O NCEDC produz dados sísmicos obtidos por meio de sensores sísmicos, geofísicos, eletromagnéticos, de nível de água e sensores de velocidade do vento. Outro exemplo de RDPGs é o Cornell University Geospatial Information Repository (CUGIR)[124], que fornece dados e metadados geoespaciais sobre o Estado de Nova York, nos USA. No entanto, conforme o Gráfico 3, a maioria dos RDPGs disponibiliza dados de pesquisa dos seguintes tipos: documentos de escritório (275), estatísticas e dados científicos (108) e imagens (73).

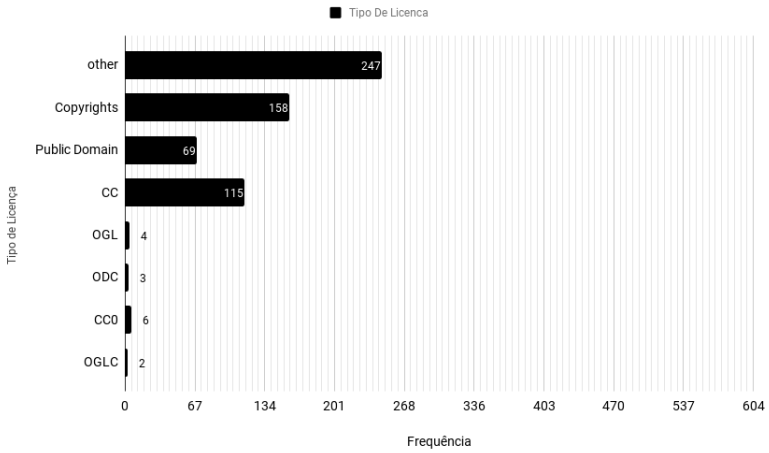
No que segue, outra questão a ser respondida: quais os tipos de acesso e as licenças utilizadas pelos RDPGs? A maioria dos repositórios (499) é de acesso aberto; (101) possuem cláusulas de restrições, (2) são

[123] NCEDC. Disponível em: < http://ncedc.org/_>. Acesso em: set. 2017.

[124] CUGIR. Disponível em: <http://cugir.mannlib.cornell.edu/about.jsp>. Acesso em: set. 2017.

acesso fechado, (2) estão embargados e (2) são fechados. O tipo de licença utilizada pode ser visualizado no Gráfico 4.

Gráfico 4 – Tipo de licença




Fonte: Dados da Pesquisa (2017).

A maioria dos RDPGs (247) utiliza algum tipo de licença especial e relacionada ao uso específico das instituições que oferecem esses dados. Por exemplo, o Community Data Portal[125], que é um repositório disciplinar nos USA apoiado pela University Corporation for Atmospheric Research, embora ofereça seus dados de forma aberta, exige que os usuários se registrem e aceitem os termos de uso[126] dos dados. Ao se cadastrar e aceitar os termos do RDPG, o usuário recebe o direito de usar os dados sem nenhum custo para fins educacionais e não comerciais. Outro exemplo de RDPGs em acesso aberto é o Figshare[127], repositório que permite a publicação de resultados de pesquisa e que adotou Creative Commons Public Domain Dedication (CC0)[128] que libera o conteúdo globalmente sem restrições e nenhum direito

[125] Community Data Portal. Disponível em: <http://cdp.ucar.edu/> Acesso em: ago 2017
[126] UCAR. Disponível em: <https://www2.ucar.edu/terms-of-use> Acesso em: ago 2017
[127] FIGSHARE. Disponível em: <http:// figshare.com> Acesso em: ago 2017> Acesso em: ago 2017
[128]CCO. Disponível em: <https://creativecommons.org/share-your-work/public-domain/cc0/> Acesso em: ago 2017.

reservado. Colocar dados de pesquisa em domínio público sob CC0 é uma maneira de remover qualquer dúvida legal sobre o uso dos dados. Embora o CC0 não exija legalmente que os usuários citem a fonte, isso não afeta as normas éticas para a atribuição nas comunidades científicas e de pesquisa. No entanto, a maioria das licenças e tipos de acesso é especifica de cada RDPGs.

No que segue (Tabela 3), busca-se responder quais são as temáticas e as áreas do conhecimento presentes nos RDPGs. No que tange às temáticas, percebe-se que os RDPGs estão distribuídos em (63) áreas de conhecimento subdivididas em: Ciências Naturais (489), Ciências da Vida (72), Humanidades (33) e Engineering Sciences (10). A Tabela 2 apresenta essa distribuição de áreas do conhecimento, conforme o uso de palavras-chave no re3data e com a ocorrência mínima de (2) termos por área.

Tabela 2 - Distribuição de áreas do conhecimento por palavras - chave.

| ÁREA DE CONHECIMENTO | PALAVRAS | FREQUÊNCIA |
|---|---|---|
| 313 Atmospheric Science and Oceanography | meteorology | 6 |
| 313 Atmospheric Science and Oceanography | climate | 5 |
| 34 Geosciences (including Geography) | atmosphere | 5 |
| 313 Atmospheric Science and Oceanography | oceanography | 4 |
| 313 Atmospheric Science and Oceanography | weather | 4 |
| 1 Humanities and Social Sciences | multidisciplin ary | 3 |
| 3 Natural Sciences | geosciences | 3 |
| 3 Natural Sciences | ocean | 3 |
| 31301 Atmospheric Science | air quality | 3 |
| 31301 Atmospheric Science | meteorology | 3 |
| 31302 Oceanography | physical sciences | 3 |
| 315 Geophysics and Geodesy | earthquake | 3 |
| 315 Geophysics and Geodesy | geophysics | 3 |
| 31502 Geodesy, Photogrammetry, Remote Sensing, Geoinformatics, Cartogaphy | gps | 3 |

| | | |
|---|---|---|
| 34 Geosciences (including Geography) | hydrology | 3 |
| 11 Humanities | multidisciplinary | 2 |
| 21 Biology | biodiversity | 2 |
| 21 Biology | multidisciplinary | 2 |
| 311 Astrophysics and Astronomy | climatology | 2 |
| 313 Atmospheric Science and Oceanography | atmosphere | 2 |
| 313 Atmospheric Science and Oceanography | climatology | 2 |
| 313 Atmospheric Science and Oceanography | cryosphere | 2 |
| 313 Atmospheric Science and Oceanography | geosciences | 2 |
| 313 Atmospheric Science and Oceanography | glaciology | 2 |
| 313 Atmospheric Science and Oceanography | hydrology | 2 |
| 313 Atmospheric Science and Oceanography | ocean | 2 |
| 313 Atmospheric Science and Oceanography | remote sensing | 2 |
| 31301 Atmospheric Science | atmosphere | 2 |
| 31301 Atmospheric Science | ecosystem | 2 |
| 34 Geosciences (including Geography) | agriculture | 2 |
| 34 Geosciences (including Geography) | biomass | 2 |
| 34 Geosciences (including Geography) | climate | 2 |
| 34 Geosciences (including Geography) | climate change | 2 |
| 34 Geosciences (including Geography) | earth science | 2 |
| 34 Geosciences (including Geography) | earth sciences | 2 |
| 34 Geosciences (including Geography) | ecology | 2 |
| 34 Geosciences (including Geography) | emission | 2 |
| 34 Geosciences (including Geography) | geodesy | 2 |
| 34 Geosciences (including Geography) | geography | 2 |
| 34 Geosciences (including Geography) | geology | 2 |

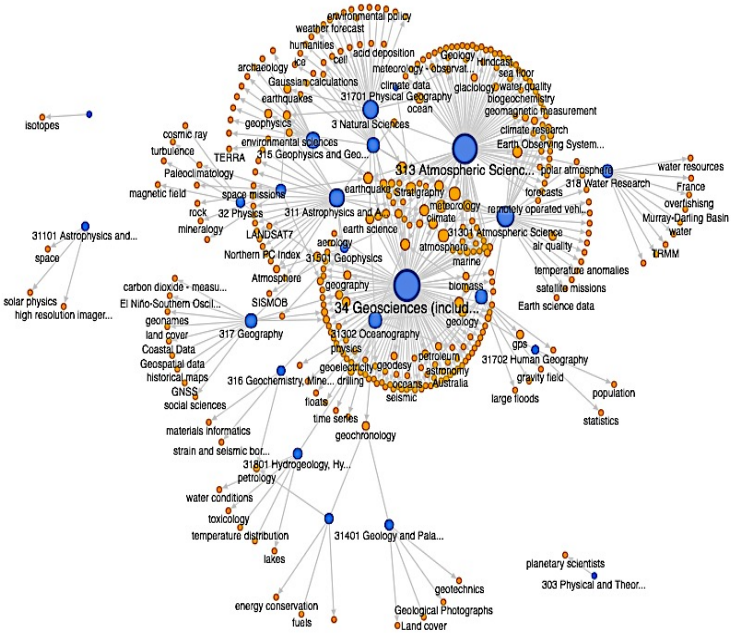
| | | |
|---|---|---|
| 34 Geosciences (including Geography) | marine biology | 2 |
| 34 Geosciences (including Geography) | meteorological sensors | 2 |
| 34 Geosciences (including Geography) | ocean circulation | 2 |
| 34 Geosciences (including Geography) | petroleum | 2 |

Fonte: Dados da Pesquisa (2017).

Os repositórios de dados de pesquisa pertencentes às áreas de Ciências Naturais possuem subáreas específicas como (*Atmospheric Science, Oceanography, Astrophysics, Astronomy, Geophysics, Geodesy, Photogrammetry, Remote Sensing, Geoinformatics, Cartogaphy*) essas subáreas revelaram as temáticas mais relevantes nos dados de pesquisa em Geociências (Tabela 3).

Os temas relacionados à área de Ciências Naturais podem ser visualizados na Figura 8.

Figura 8 - Temas mais frequentes em Ciências Naturais.

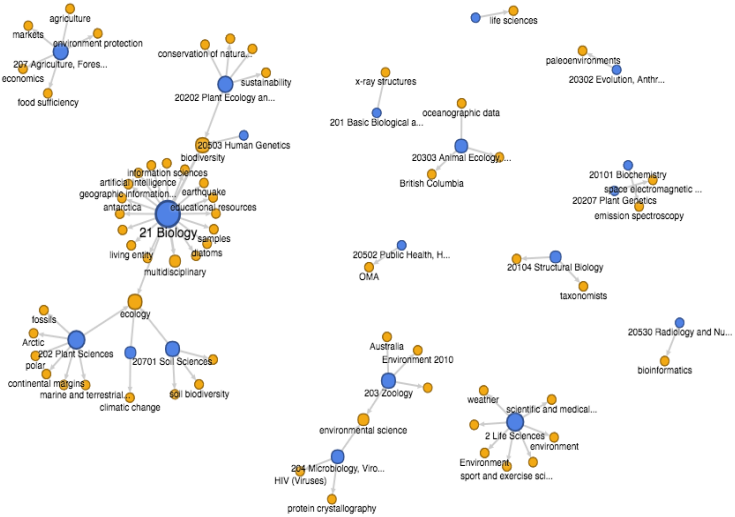



Fonte: Dados da Pesquisa (2017).

Nesse contexto, é possível visualizar (359) temas utilizados para descrever os RDPGs, entre eles os mais recorrentes são meteorologia, climatologia, geodésia e oceanografia. As palavras-chave mais utilizadas em Ciências Naturais revelam as principais temáticas disponíveis nos RDPGs: *telemetry, natural disaster, satélites, site-based (point) ecological data, weather forecast, ice, gps, ocean tides, ocean, seismic data* e *earthquakes*, entre outros temas que podem ser encontrados nos RDPGs. Como exemplo, é possível citar o World Data Centre for Glaciology[129], em Cambridge, no Scott Polar Research Institute. Esse repositório possui coleções de dados de pesquisa sobre os aspectos da neve e do gelo em todo o mundo, cerca de 40.000 registros, desde 1661.

As Ciências da Vida englobam (*Medicine, Zoology, Biology, Agricultural, Economics and Sociology*), entre outras subáreas, conforme a Figura 10.

Figura 9 - Temas mais frequentes em Ciências da Vida.




Fonte: Dados da Pesquisa (2017).

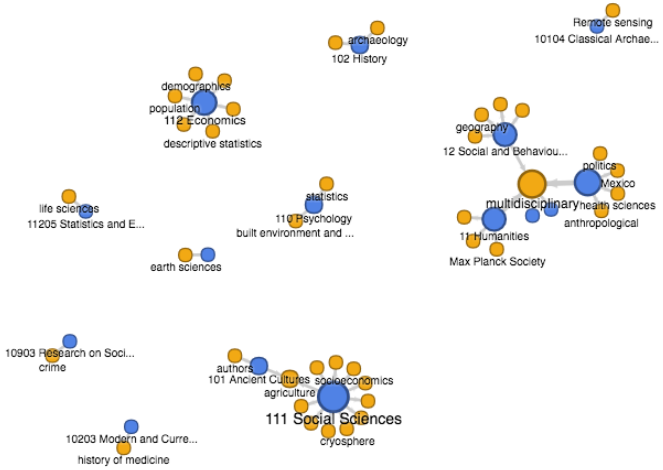
As principais temáticas disponíveis em Ciências da vida são: *biodiversity, ecology, marine and terrestrial animals, taxonomists,*

[129]WORLD DATA CENTRE FOR GLACIOLOGY. Disponível em: <https://www.spri.cam.ac.uk/about/>. Acesso em: set. 2017.

*geospatial data, oceanographic data, Artic, Antarctic, climatic change*, entre outros temas que podem ser encontrados nos RDPGs. Como exemplo, é possível citar o Integrated Ocean Drilling Program (IODP)[130] Bremen Core Repository[131], que disponibiliza dados oceanográficos sobre os oceanos Atlântico e Mediterrâneo, Mar Negro e Mar Báltico. O projeto é desenvolvido pela University of Bremen e representa a participação alemã no IODP. O Bremen Core Repository é de três repositórios do IODP, como o Gulf Coast Repository (GCR)[132], em College Station, Texas, USA, e o Kochi Core Center (KCC)[133], no Japão.

Outra área de conhecimento, as Humanidades, possui (33) RDPGs e abrage disciplinas como (*Ancient Cultures, Pre history, Education, Sciences, Economics Social and Behavioural Sciences, Statistics e Econometrics)*, conforme a Figura 9.

Figura 10 - Temas mais frequentes em Humanidades.




Fonte: Dados da Pesquisa (2017).

---

[130] IODP. Disponível em:< https://www.iodp.org/>. Acesso em setembro 2017.

[131] BREMEN CORE REPOSITORY. Disponível em:< https://www.marum.de/en/Research/IODP-Bremen-Core-Repository.html>. Acesso em setembro 2017.

[132] GCR. Disponível em:< http://iodp.tamu.edu/curation/gcr/index.html >. Acesso em setembro 2017.

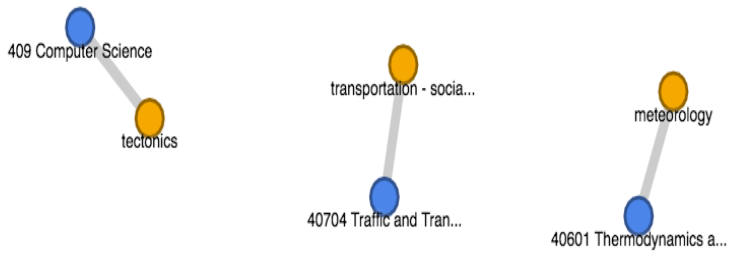
[133] KCC. Disponível em:< http://www.kochi-core.jp/en/iodp-curation/index.html >. Acesso em setembro 2017.

É possível visualizar (54) termos utilizados para descrever os RDPGs nas Humanidades. As temáticas disponíveis são: *multidisciplinary, anthropological, politics, archaeology, marine environment, remote sensing, agriculture, censos, statistics descriptive, statistics, demographics*, entre outros temas que podem ser encontrados nos RDPGs. Como exemplo é possível citar a Database of Places, Language, Culture and Environment (D-PLACE)[134]. Esse RDPGs oferece coleções de dados de pesquisa sobre informações culturais, linguísticas, ambientais e geográficas.

As Engineering Sciences possuem (10) RDPGs que envolvem temáticas como (*Process Engineering*, *Materials Science* e *Computer Science*), conforme Figura 11.

Figura 11 - Temas mais frequentes em Engenharias

.




Fonte: Dados da Pesquisa (2017).

As principais temáticas disponíveis em RDPGs em Engenharia são: *meteorology*, *transportation – social aspects* e *tectonics*. Como exemplo, é possível citar o Monash University Data Repository no Figshare (**ver cápitulo 1.1**), que desde 2014 permite a pesquisadores australianos armazenar, compartilhar e publicar dados de pesquisa Figshare. Todo o conjunto de dados no Figshare recebe um Digital Object Identifier (DOI).

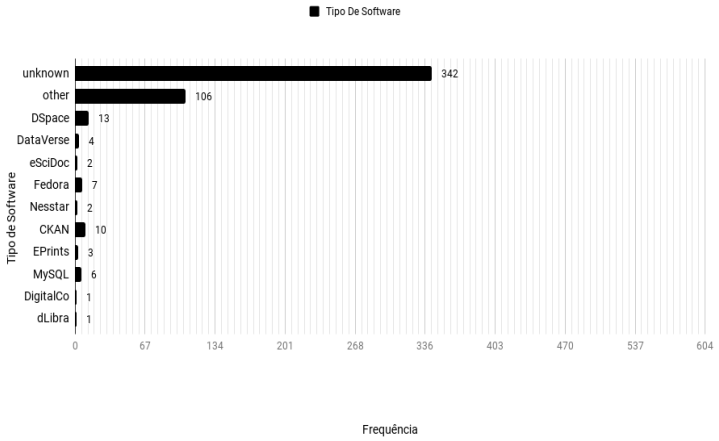
Por fim, no que tange ao uso de tecnologias para elaboração de RDPGs, apresenta-se o Gráfico 5. Nele, busca-se responder à pergunta:

[134] D-PLACE. Disponível em: <https://d-place.org/angular/>. Acesso em: set. 2017.

quais os *softwares* utilizados para elaborar RDPGs? Percebe-se no Gráfico 5 que mais de metade dos *softwares* utilizados (342) para elaborar RDPGs é desconhecida e (106) são classificados como outro tipo de *software*, ou seja, 450 RDPGs são implementados com uma diversidade de tecnologias de *software* específico do repositório, o que, de modo geral, dificulta a identificação do que é utilizado para elaborar esses repositórios de dados de pesquisa. Por exemplo, quais os tipos de APIs e protocolos são utilizados para proporcionar a interoperabilidade dos dados.

Assim, desse conjunto de *softwares* selecionou-se uma amostra contendo RDPGs elaborados com CKAN, Dspace, DataVerse, e-Prints e Fedora (**ver capítulo 2.5.2**).

Gráfico 5 – Tipologia dos *softwares* de repositórios




Fonte: Dados da Pesquisa (2017).

A relevância do uso desses *softwares* é que eles facilitam o acesso, a interoperabilidade e a preservação de dados de pesquisa digitais. Essas plataformas de *software* livre são conhecidas e implementadas internacionalmente pela comunidade de bibliotecas e foram desenvolvidas para proporcionar a coleta, a preservação e a divulgação de publicações e dados de pesquisa. No entanto, podem ser usadas para agregar qualquer tipo de conteúdo em formato digital.

Assim, separou-se da amostra de 604 RDPGs um conjunto de (37) RDPGs elaborados com CKAN, Dspace, DataVerse, e-Prints e Fedora, conforme a Tabela 3.

Tabela 3 - Relação de RDPGs: instituição, tipo de dados, países, acesso e licenças.

| Software | Nome Repositório | Instituição | Tipo de dados | País | Acesso | Licença |
|---|---|---|---|---|---|---|
| CKAN | CCCA Data Centre | Climate Change Centre Austria | Scientific and statistical data formats | AUT | open | Copyrights |
| CKAN | DART Data | University of Leeds | Standard office documents | GBR | open | CC |
| CKAN | GeoPlatform | Federal Geographic Data Committee | Standard office documents | USA | open | Public Domain |
| CKAN | Geothermal Data Repository | U.S. Department of Energy | Structured text | USA | open | CC |
| CKAN | Landcare Research Data Repository | Landcare Research NZ | Scientific and statistical data formats | NZL | open | CC |
| CKAN | London Datastore | Greater London Authority | Standard office documents | GBR | open | OGL |
| CKAN | National Geothermal Data System | Open Energy Information, Geothermal Data Repository | Scientific and statistical data formats | USA | open | CC |
| CKAN | Publicdata.eu | Open Knowledge Foundation | Standard office documents | GBR | open | other |
| CKAN | Water use data | Environment and Climate Change Canada | Scientific and statistical data formats | CAN | open | Copyrights |
| CKAN | eartH2Observe | Plymouth Marine Laboratory | Standard office documents | GBR | open | CC |
| DSpace | Apollo | University of Cambridge, University Library | Standard office documents | GBR | open | CC |

| | | | | | | |
|---|---|---|---|---|---|---|
| DSpace | Banco de Información para la Investigación Aplicada en Ciencias Sociales, Repositorio Institucional | Centro de Investigacion y Docenia Economicas, A.C. | Standard office documents | MEX | open | CC |
| DSpace | DIGITAL.CSIC | Spanish National Research Council | Standard office documents | ESP | open | CC |
| DSpace | DRYAD | Dryad | Standard office documents | AAA | open | other |
| DSpace | DataDOI | University of Tartu | Standard office documents | EST | open | CC |
| DSpace | Earth-prints Repository | | Structured text | ITA | open | CC |
| DSpace | GO-GEO | Higher Education Funding Council for England | Images | GBR | open | other |
| DSpace | Research Data Online | Australian National Data Service | Raw data | AUS | open | Public Domain |
| DSpace | SPECTRa Project | University of Cambridge, Department of Chemistry, Murray-Rust Research Group | Standard office documents | GBR | open | Copyrights |
| DSpace | Scholar's Bank | University of Oregon, UO Libraries | Standard office documents | USA | open | CC |
| DSpace | ShareGeo open | University of Edinburgh, EDINA | Standard office documents | GBR | open | CC |
| DSpace | Wittliff Collections, Texas State University, Digital Collections Repository | Texas State University, University Library | Standard office documents | USA | open | other |
| DSpace | Woods Hole Open Access | Woods Hole Oceanograph | Standard office | USA | open | Copyrights |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Server | ic Institution, Marine Biological Laboratory, Library | documents | | | |
| DataVerse | ICRISAT Dataverse Network | International Crops Research Institute for the Semi-Arid Tropics | Standard office documents | IND | open | CC |
| DataVerse | UCLA Social Science Data Archive Dataverse | University of California at Los Angeles, Library, Social Science Data Archive | Scientific and statistical data formats | USA | open | other |
| DataVerse | UiT Open Research Data Dataverse | The Dataverse Project | Structured text | USA | embargoed | CC0 |
| DataVerse | heiDATA | Heidelberg University, Competence Centre for Research Data | Standard office documents | DEU | open | CC |
| EPrints | LSHTM Data Compass | London School of Hygiene & Tropical Medicine | Scientific and statistical data formats | GBR | open | CC |
| EPrints | Landmap | Joint Information Systems Committee | Scientific and statistical data formats | GBR | restricted | Copyrights |
| EPrints | OceanRep GEOMAR Repository | GEOMAR Helmholtz Centre for Ocean Research Kiel, Library | Images | DEU | open | CC |
| Fedora | | | Images | AUT | open | CC |
| Fedora | 4TU.Centre for Research Data | TU Delft Library | Standard office documents | NLD | open | CC |
| Fedora | ANU Data Commons | Australian National University, | Standard office documents | AUS | open | other |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | Division of Information | | | | |
| Fedora | Center for International Earth Science Information Network | Columbia University, Earth Institute | Standard office documents | USA | open | Copyrights |
| Fedora | NASA Socioeconomic Data and Applications Center | National Aeronautics and Space Administration, Earth Observing System Data and Information System | Standard office documents | USA | open | Copyrights |
| Fedora | TAMBORA | Leibniz Institute for Regional Geography | Standard office documents | DEU | open | other |
| Fedora | UCD Digital Library | Higher Education Authority | Standard office documents | IRL | open | Copyrights |

Fonte: Dados da Pesquisa (2017).

As instituições que utilizam o CKAN estão localizadas em países como GBR, USA, AUT, NZL e CAN. São centros de pesquisa, universidades e entidades governamentais que buscam oferecer dados abertos.

Conforme a Tabela 3, as instituições que utilizam o CKAN. Uma delas é o Climate Change Centre Austria[135](CCCA) Data, que funciona como um centro de dados e informações climáticas, oferecendo dados de observação e medição de cenários quantitativos e qualitativos, bem como os resultados de projetos de pesquisa de climatologia do governo austríaco. Outro RDPGs elaborado com CKAN está na University of Leeds na Inglaterra, o Detection of Archaeological Residues Using Remote-sensing Techniques (DART) DATA[136] , RDPGs que disponibiliza dados de pesquisa arqueológicos capturados por sensoriamento remoto. Outro exemplo, o London Datastore do Greater London Authority (GLA)[137], em Londres, na Inglaterra, disponibiliza cerca de 700 conjuntos de dados sobre informações estatísticas e dados abertos do governo inglês sobre a cidade de Londres.

[135] CCCA Data Centre. Disponível em: <https://data.ccca.ac.at/about>. Acesso em: set. 2017.
[136] DART Data. Disponível em: <http://dartportal.leeds.ac.uk/>. Acesso em: set. 2017
[137] GLA. Disponível em: < https://data.london.gov.uk/dataset/>. Acesso em: set. 2017.

Nos USA, apresenta-se o GeoPlatform[138], da Federal Geographic Data Committee. Esse RDPGs oferece cerca de 160.000 conjuntos de dados geoespaciais, serviços e aplicações para uso público e por agências governamentais norte-americanas. Também nos USA, o Geothermal Data Repository[139], do Energy Departament, oferece uma coleção de dados textuais sobre energia e eficiência energética produzida por pesquisadores do departamento de energia dos USA.

Os (10) repositórios elaborados com CKAN são classificados como do tipo disciplinar (7), institucional (2), sendo (1) de outra categoria. Estão classificados nas áreas de *Geosciences (including Geography)*, *Fine Arts*, *Music*, *Theatre and Media Studies* (1), *Statistics and Econometrics* (1), *Biology* (1), *Atmospheric Science* (1), *Geology and Palaeontology* (1), *Human Geography* (1), *Water Research* (1). As palavras-chave mais usadas para descrição do conteúdo desses repositórios são: *biodiversity*, *climate change*, *climate information*, *energy development*, *exploration*, *hydrology*, *life sciences*, *multidisciplinar* e *water*. Em suma, os principais tipos de dados depositados na plataforma CKAN são *Scientific and statistical data formats* (4), *Standard office documents* (4), *Structured text*. Todos são de acesso aberto mediante licenças CC (4), Copyrights (2), Public Domain (1), OGL (1).

Conforme a Tabela 3, os RDPGs elaborados com DSpace estão localizadas em países como GBR, USA, AUS, ESP, EST, ITL e MEX. As suas instituições são centros de pesquisa, universidades e entidades governamentais e não governamentais.

Na GBR, a University of Cambridge é uma das instituições que utilizam o DSpace para disponibilizar seus dados de pesquisa. O acervo de Cambridge está disponível no Apollo[140], que é o repositório institucional da universidade, preservando e fornecendo acesso ao conteúdo criado por membros de Cambridge. O repositório armazena uma variedade de conteúdo e fornece diferentes níveis de acesso às publicações de pesquisa da universidade.

Nos USA, o Woods Hole Open Access Server da (WHOI) também utiliza o DSpace. Ela disponibiliza a Data Library and Archives (DLA)[141], uma coleção diversificada de registros administrativos, fotografias, jornais pessoais de cientistas, filmes e vídeos, instrumentos

---

[138] GEOPLATFORM. Disponível em: <https://ckan.geoplatform.gov/>. Acesso em: set. 2017.
[139] GEOTHERMAL Data Repository. Disponível em: <https://gdr.openei.org/>. Acesso em: set. 2017.
[140] APOLLO. Disponível em: <https://www.repository.cam.ac.uk/>. Acesso em: set. 2017.
[141] DLA. Disponível em: <http://darchive.mblwhoilibrary.org/>. Acesso em: set. 2017.

históricos, dados de cruzeiros, diários de bordo, diários, planos, histórias orais, além de publicações, mapas, atlas, livros e relatórios técnicos da WHOI. Na EST, a Universidade de Tartu disponibiliza o DataDOI[142], que é um serviço digital que coleta, preserva e distribui dados de pesquisa com identificador digital DOI.

Os (13) repositórios elaborados com Dspace são classificados como do tipo *diciplinary* (3), *institucional* (9), sendo (1) de outra categoria. Estão classificados nas áreas de *Humanities and Social Sciences*, *Geosciences (including Geography)*, *Ancient Cultures*, *Literary Studies*, *Humanities*, *Life Sciences*, *Chemistry*, *Geophysics and Geodesy* e *Geophysics*. As palavras-chave mais usadas para descrição do conteúdo desses repositórios são: *multidisciplinary*, *air, archaeology*, *atmosphere*, *authors*, *planning*, *Gaussian calculations*, *marine biology*, *Mexico*, *politics*, *scientific* e *medical publications.*

Por fim, os principais tipos de dados depositados na plataforma DSpace são: *Standard office documents* (10), *Images* (1), *Raw data* (1), *Structured text* (1). Todos são de acesso aberto mediante licenças CC (7), other (3), Copyrights (2), Public Domain (1).

As instituições que usam o Dataverse, conforme a Tabela 3, estão em países como DEU, IND e USA. Na Alemanha, a Heidelberg University, Competence Centre for Research Data disponibiliza o heiDATA[143], que é um repositório institucional para dados de pesquisa gerido pela Biblioteca Universitária e pelo Centro de Computação da universidade. Na Índia, o International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)[144] é um repositório de dados que coleta, preserva e facilita o acesso aos conjuntos de dados de pesquisa produzidos pelos pesquisadores da ICRISAT. Os dados incluem fenotípica, genotípica, ciências sociais e dados espaciais sobre solo e tempo. Nos USA está o UiT Open Research Data Dataverse[145], um repositório de dados de pesquisa institucional disponibilizado pela UiT The Arctic University of Norway. Outro Dataverse nos USA é o UCLA Social Science Data Archive Dataverse[146], que disponibiliza dados de pesquisa produzidos por professores e estudantes de pós-graduação da University of California, Los Angeles.

---

[142] DATADOI. Disponível em: <http://datadoi.ut.ee/>. Acesso em: set. 2017.
[143] HEIDATA. Disponível em: <https://heidata.uni-heidelberg.de/>. Acesso em: set. 2017.
[144] ICRISAT. Disponível em: <http://dataverse.icrisat.org/>. Acesso em: set. 2017.
[145] UIT OPEN Research Data Dataverse. Disponível em: <https://dataverse.no/dataverse/uit/>. Acesso em: sete. 2017.
[146] UCLA Social Science Data Archive Dataverse. Disponível em: <https://dataverse.harvard.edu/dataverse/ssda_ucla/>. Acesso em: set. 2017.

Os (4) repositórios elaborados com o Dataverse são classificados como do tipo *diciplinary* (1), *institucional* (3). Estão classificados nas áreas de *Humanities*, *Social Sciences Agriculture*, *Forestry*, *Horticulture and Veterinary Medicine*, *Natural Sciences*. As palavras-chave mais usadas para descrição do conteúdo desses repositórios são: *data processing*, *food sufficiency*, *humanities*, *social sciences*. Por fim, os principais tipos de dados depositados na plataforma DataVerse são: *Scientific and statistical data formats, Standard office documents, Structured text*. Conforme os tipos de acesso, eles se configuram como de acesso aberto, mas com tempo de embargo mediante licenças de CC.

Já os RDPGs elaborados com o *software* e-Prints, conforme Tabela 3, estão localizados na DEU e na GBR. O repositório alemão é o OceanRep GEOMAR Repository [147], uma coleção digital de acesso aberto sobre os processos químicos, físicos, biológicos e geológicos nos oceanos. O OceanRep contém o resultado das pesquisas dos pesquisadores e dos estudantes da GEOMAR, o que inclui artigos de periódicos, artigos de conferências, capítulos de livros e teses em texto completo. Os dados de pesquisa estão vinculados às entradas dessas publicações. Na GBR está o LSHTM Data Compass[148], que é um repositório de dados de pesquisa produzidos pela London School of Hygiene & Tropical Medicine e seus colaboradores. Outro RDPGs na GBR é Landmap[149], da Centre for Environmental Analysis (CEDA). O Landmap é repositório que fornece acesso a uma variedade de dados geoespaciais, os quais incluem: fotografias aéreas e imagens de satélites.

Os (3) repositórios elaborados com e-Prints são classificados como do tipo *diciplinary* (2), *institucional* (1). Estão classificados nas áreas de Social and Behavioural Sciences, Agriculture, Forestry, Horticulture and Veterinary Medicine e Natural Sciences. As palavras-chave mais usadas para descrição do conteúdo desses repositórios são *Geosciences*, *health* e *ocean*. Os principais tipos de dados depositados na plataforma e-Prints são: *Scientific and statistical data formats* e *images*. Conforme os tipos de acesso, são de acesso aberto e de acesso restrito mediante licenças CC e Copyrights.

---

[147] OCEANREP GEOMAR Repository. Disponível em: < http://oceanrep.geomar.de/>. Acesso em: set. 2017.
[148] LSHTM Data Compass. Disponível em: <http://datacompass.lshtm.ac.uk/>. Acesso em: set. 2017.
[149] LANDMAP. Disponível em: <http://catalogue.ceda.ac.uk/uuid/7f1280cf215da6f8001eae5c2f019fe8/>. Acesso em: set. 2017.

Por fim, as instituições que disponibilizam seus dados com Fedora, conforme Tabela 3, estão em países como USA, AUS, AUT, DEU, IRL, NLD. Nos USA está o Center for International Earth Science Information Network (CIESIN) [150], da Columbia University, que é um RDPGs interdisciplinar que fornece acesso a uma ampla gama de dados globais, documentação associada e ferramentas de visualização e análise sobre o meio ambiente. NA AUS, o ANU Data Commons[151], produzido pela Australian National University, Division of Information, oferece dados de pesquisa sobre quatro áreas disciplinares: *Earth Sciences*, *Astronomy*, *Phenomics* e *Digital Humanities*. Na Alemanha, está o TAMBORA[152], produzido pela Leibniz Institute for Regional Geography. Os dados de pesquisa disponíveis no TAMBORA são sobre climatologia histórica. Um exemplo na IRL, o UCD Digital Library[153], produzido pela University College Dublin, oferece dados de pesquisa como materiais históricos, fotografias, arte, entrevistas, cartas, entre outros conteúdos.

Os (7) repositórios elaborados com FEDORA são classificados como do tipo *diciplinary* (4), *institucional* (2), sendo (1) de outra categoria. Estão classificados nas áreas de *Geosciences (including Geography)*, (2), *Humanities* (1), *Social Sciences* (1), *Biology* (1), *Atmospheric Science* (1) e *Geography* (1). As plavras-chave mais usadas para descrição do conteúdo desses repositórios são *earth sciences*, *environmental assessment*, *historical maps*, *multidisciplinar*, *software development*, *technology* e *constructione weather*. Em suma, os principais tipos de dados depositados na plataforma Fedora são: *Standard office documents* e *images*. Todos são de acesso aberto mediante licenças CC e Copyrights.

## 4 **CONSIDERAÇÕES**

Ao longo deste estudo, abordaram-se temáticas relacionadas à CI em contextos de *e-Science,* em específico, sobre a *e-Science Librarianhip* ou Biblioteconomia de Dados. A Biblioteconomia de Dados é a disciplina responsável por aplicar os fundamentos da CI, da *e-Science* e da *Data Science* em bibliotecas. As transformações ocorridas

---

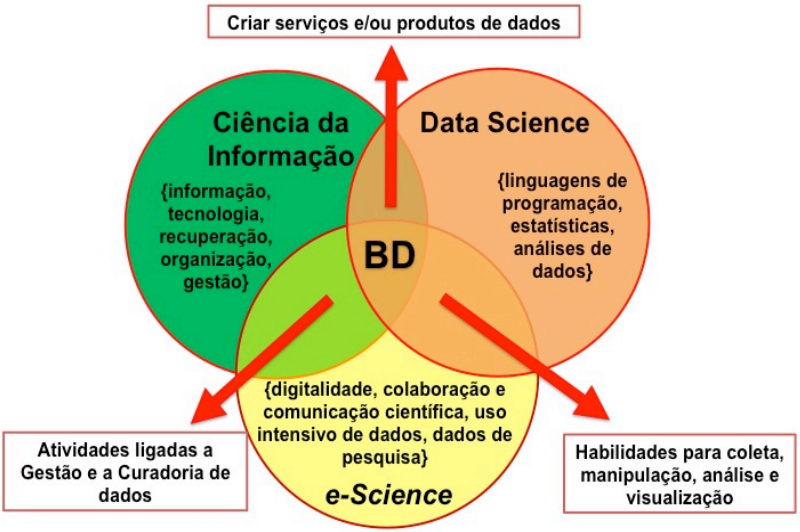
[150] CIESIN. Disponível em: <http://www.ciesin.columbia.edu/data.html/>. Acesso em: set. 2017.
[151] ANU Data Commons. Disponível em: <https://datacommons.anu.edu.au/DataCommons//>. Acesso em: set. 2017.
[152] TAMBORA. Disponível em: <https://www.tambora.org/>. Acesso em: set. 2017.
[153] UCD Digital Library. Disponível em: <http://digital.ucd.ie/> Acesso em: set. 2017.

na CI devido à influência das tecnologias e da ciência com uso intensivo de dados repercutem diretamente nas atividades dos bibliotecários de dados, exigindo que esses profissionais adquiram competências em *Data Science*.

O foco da Biblioteconomia de Dados está relacionado aos estudos sobre dados de pesquisa em repositórios digitais. Os bibliotecários de dados devem utilizar esses ambientes orientados a dados para modificar suas tarefas rotineiras. Nesse sentido, os bibliotecários de dados precisam conhecer as técnicas, os métodos e as tecnologias utilizados em *Data Science* para que possam adquirir as habilidades operacionais necessárias para coleta, manipulação, análise e visualização de dados de pesquisa em repositórios digitais.

Assim, como circunscrever as relações entre CI, e-*Science* e *Data Science* que fundamentam a Biblioteconomia de Dados? Para responder a essa questão, que representa o objetivo geral desta tese, apresenta-se um Diagrama de Venn para sumarizar o que se entendeu por Biblioteconomia de Dados (BD) (Figura 12).

Figura 12 - Diagrama de Venn da Biblioteconomia de Dados.




Fonte: Dados da Pesquisa (2017).

O diagrama de Venn da Biblioteconomia de Dados (Figura 12) representa as relações de união e interação entre CI, *e-Science* e *Data Science.* Nele é possível conceituar a Biblioteconomia de Dados, por

meio seguintes expressões: CI = {informação, tecnologia, recuperação, organização, gestão de repositórios}; *e-Science* = {digitalidade, colaboração e comunicação científica, uso intensivo de dados, dados de pesquisa}; *Data Science* = {linguagens de programação, estatísticas, analises de dados}. Sendo a Biblioteconomia de Dados uma disciplina que se concentra no núcleo do tripé (CI, *e-Science* e *Data Science*), tendo como pilares as atividades de gestão e curadoria de dados em repositórios digitais, pautadas na prática com a coleta, a manipulação, a análise e a visualização de dados, as quais são fundamentais para que os bibliotecários de dados ofertem serviços e produtos a partir de dados em bibliotecas.

Assim, para clarificar as relações desse diagrama, propõe-se uma revisitação de alguns dos principais pontos já discutidos ao longo do texto.

Em um primeiro momento, apresentaram-se os **capítulos 2.1**, **2.2** e **2.3** para identificar-se a Ciência da Informação em contextos de *e-Science*, ressaltando-se o conceito de dados como objeto de estudo para a Biblioteconomia de Dados**.** Em específico, para discutir essas temáticas foram utilizados os seguintes autores: Harold Borko (1968), Tefko Saracevic (2010), David Bawden e Lyn Robinson (2012), Lena Vânia Pinheiro (1998), Rafael Capurro (2007), Michael Buckland (1991), Peter Morville (2014), Luciano Floridi (2010), Scharff e Dusek (2006), David Kaplan (2009), Alberto Cupani (2013), Tony Hey, Stewart Tansley e Kristin Tolle (2009), entre outros.

O conceito de informação e o conceito de tecnologia são fundamentais para se compreender a CI enquanto disciplina científica. A informação está presente, da Física à Biologia, das organizações à vida cotidiana. Um exemplo de definição multifacetada do conceito de informação retém em seu *core* de significados, desde uma simples descrição de como se usam os termos de um dicionário até aplicações tecnológicas complexas e teorias específicas, como a Teoria Matemática da Informação e a Recuperação de Informação.

A tecnologia de modo geral é parte notória do mundo contemporâneo. Significa tanto a satisfação pelos aparelhos que tornam a vida mais cômoda como o entusiasmo ante as possibilidades criadas pelo computador e a internet. Assim, a tecnologia se apresenta como uma realidade polifacetada: não apenas em forma de objetos e conjuntos de objetos, mas também como sistemas, como processos, como modos de proceder, como certa mentalidade (CUPANI, 2013; DUSEK, 2006).

Arrisca-se a propor que algumas correntes de pensamento em CI são ou foram criadas por influência de algum tipo de tecnologia (livro,

computador) e que de maneira interconectada essas tecnologias compõem um rico universo de dispositivos inteligentes que registram o saber mediado por instrumentos e máquinas como computadores e pós-computadores (*smartphones* e *tablets*). Essas tecnologias são inexoravelmente ligadas aos interesses humanos, a suas práticas sociais, às leis entre uma variedade de outros fatores que nos inter-relacionam e interrogam no dia a dia.

Nesse contexto, o lugar epistemológico da informação e da tecnologia relaciona-se ao ponto de vista de quem as estuda; sua aplicação conceitual depende de contexto específico. Tradicionalmente, esses dois conceitos constituem a base dos estudos sobre CI; no entanto, devido às atuais transformações impulsionadas pelas tecnologias, percebe-se que a CI está modificando o foco de seus estudos, de uma visão centrada nas tecnologias e na informação para uma abordagem mais centrada nos estudos sobre dados em formato digital. Os dados, nessa perspectiva, devem ser vistos como os elementos básicos que compõem a informação de natureza digital. Os dados digitais são tão complexos quanto a informação gerada por eles. Virtualmente, todos os tipos de informação em formato digital são compostos por dados. É possível se referir a tudo em meio digital como dados.

Assim, acredita-se que a *e-Science* seja um ramo multidisciplinar do conhecimento científico que se preocupa com o uso intensivo de dados e com a mediação tecnológica em disciplinas científicas, localizada entre os estudos sobre tecnologias, problemas da informação, uso de dados, colaboração e comunicação científica global. Em suma, essas questões estão relacionadas aos contextos impostos pelas infraestruturas (repositórios de dados de pesquisa) da *e-Science* e podem apresentar-se como uma nova classe de estudos em CI. A *e-Science* surge como uma oportunidade de se repensarem as transformações epistemológicas ocorridas nos conceitos de informação e de tecnologia, introduzindo esses conceitos sob a perspectiva dos estudos realizados pelas ciências orientados ao uso intensivo de dados.

Sob esse viés, sugere-se que o uso intensivo de dados seja a base dos estudos de Ciência da Informação em contextos de *e-Science*. No que segue, para consolidar essa abordagem, apresentaram-se os **capítulos 2.3 e 2.4**, contendo os conceitos de: dados de pesquisa, Biblioteconomia de Dados, *Data Science*, habilidades requeridas aos cientistas de dados e habilidades requeridas aos bibliotecários de dados. As relações existentes entre esses conceitos representam o que pode vir a ser considerado estudos sobre Biblioteconomia de Dados contemporânea. Especificamente, quando eles evidenciam a criação de

novos produtos e/ou serviços em bibliotecas relacionados ao uso e consumo de dados de pesquisa.

Nesse contexto, para delimitar esses conceitos, alguns autores foram discutidos: Koltay (2015), Zhu e Xiong (2015), Mantra (2012-13), Rice e Southall (2016), National Science Board (2011), Rodrigues et al. (2010), Panpel et al. (2013), Organization for Economic Co-Operation and Development (2004), Tibbo e Jones (2017), Costa e Braga (2016), International Association for Social Science Information Services and Technology (2016), Kellam e Thompson (2016), Henderson (2017), Pryor (2013), Dudziak (2016), Sayão e Sales (2015), National Research Council UK (1999), Geraci, Humphrey e Jacobs (2012), Scholtz (2016), York University (2017), Grigg (2016), Federer (2016), Davenport (2014), Burlingame e Nielsen, (2014), Van Der Aalst (2014), Affelt (2015), Davenport (2014), Hayashi (1996), Chen (2015), Smith (2006), Voulgaris (2014), Barga, Fontama e Hyong (2014), Conway (2010), Mccaffrey e Giesbrecht (2016), Kennan (2016), Partlo (2014), Stuart (2011), Bates (2015), Witt (2016), Dekker e Lackie, (2016), entre outros.

Com a inclusão das tecnologias de dados nas práxis bibliotecárias, percebe-se uma transformação no modo como o bibliotecário vê suas próprias atividades. As atuais mudanças na forma de se conduzir a pesquisa científica em CI oferecem oportunidades para os bibliotecários de dados se envolverem com a ciência de uso intensivo de dados; em específico, eles precisam adquirir habilidades em *Data Science*.

A *Data Science* pode ser considerada como uma nova área de estudos para bibliotecários que procuram se envolver com as questões ligadas à gestão e à análise de dados de pesquisa. Ela investiga questões a respeito de práticas, modelos e metodologias utilizadas para se obter informação útil a partir de dados, que podem ser extraídos de ambientes heterogêneos e complexos. Interessa-se por coleta, análise e visualização de dados, fundamenta-se em disciplinas de áreas exatas, como a Estatística e a Ciência da Computação, mas também envolve as Ciências Sociais e Comportamentais, investigando o comportamento humano durante o uso de tecnologias de dados.

Dessa forma, o bibliotecário de dados precisa aprender mais sobre essa ciência. A tarefa essencial dos bibliotecários de dados será a de transformar suas tarefas e atividades cotidianas, inserindo a *Data Science* no contexto da geração de produtos e/ou serviços de dados em bibliotecas. Acredita-se que a *Data Science* proporcione novos métodos e práticas que podem contribuir para a proposição desses produtos e/ou

serviços de dados. Principalmente, produtos e/ou serviços relacionados a dados de pesquisa.

O conceito de dados de pesquisa aponta uma forte tendência de estudos e práticas profissionais que está se desenvolvendo no cerne das investigações realizadas por bibliotecários de dados e/ou outros cientistas da informação preocupados com os estudos de CI, *e-Science* e *Data Science*. Os dados de pesquisa são uma oportunidade para os bibliotecários de dados oferecerem serviços e/ou produtos de dados a seus usuários. Os dados de pesquisa são um produto para pesquisa, bem como um insumo para a pesquisa. Esse tipo de dados assume papel relevante no que tange aos novos processos de comunicação científica. No futuro, os bibliotecários de dados devem pensar em serviços que liguem pesquisadores a produtos de dados de pesquisa. Os dados de pesquisa devem poder ser coletados sistematicamente, estruturados, documentados e armazenados em repositórios de dados de pesquisa para servir como insumos para futuras pesquisas.

Assim, afirma-se que as atuais tendências de pesquisa em Biblioteconomia de Dados fundamentadas em CI, *e-Science* e *Data Science* estão voltadas para estudos sobre a proveniência, as melhores práticas, o planejamento, a curadoria, entre outras atividades científicas relacionadas à teoria e à *práxis* com dados, principalmente quando esses dados se relacionam com os processos de pesquisa científica em meio digital. Basicamente, o campo da Biblioteconomia de Dados está relacionado com a prática, com o uso e com o consumo de dados.

No que segue, afirma-se que os repositórios de dados de pesquisa são parte da ciberinfraestrutura de *e-Science* e que podem ser vistos como uma das principais áreas de atuação para bibliotecários de dados preocupados em gerar serviços e/ou produtos de dados em bibliotecas.

Nesse sentido, tendo em vista uma abordagem mais técnica, para alcançar objetivo específico (b) – experimentar processos e tecnologias de pesquisa utilizadas por cientistas de dados para manipulação de dados de pesquisa extraídos de RDPG – apresentou-se no **capítulo 2.5** um conjunto de algoritmos, tipos de análise, linguagens de programação e *softwares* utilizados para elaboração, manipulação e análise de repositórios de dados de pesquisa.

Sob esse viés, ressalta-se que os bibliotecários de dados precisam compreender que a complexidade e o ineditismo ligados à práxis do *data scientist* representam um novo desafio de aprendizagem a todos os interessados em repositórios de dados de pesquisa. O uso de algoritmos, *softwares* e linguagens de programação permitirá que os bibliotecários de dados automatizem tarefas e rotinas repetitivas ligadas à

Biblioteconomia e necessárias para o trabalho com dados de pesquisa. É por meio da habilidade *hacker* do cientista de dados que o bibliotecário de dados poderá, efetivamente, trabalhar em ambientes orientados a dados, como os repositórios digitais.

No entanto, um bibliotecário de dados não precisa se tornar um programador, mas deve estar interessado em aprender sobre as linguagens e a lógica de programação dos computadores. Dessa forma, deve conhecer o funcionamento de uma gama de *softwares* científicos usados para transformar, preservar e analisar dados. Um bibliotecário não precisa ser conhecedor de linguagens de programação de banco de dados, mas cada bibliotecário deve compreender os fundamentos das ferramentas de recuperação de informação, como *web scraping* e *harvesting* de metadados. Um bibliotecário não precisa ser um estatístico, necessita entender as bases gerais da estatística básica aplicada em estudos métricos da informação, como a Bibliometria.

Assim, no **capítulo 2.5**, apresentou-se o corpus de algoritmos, tipos de análise, linguagens de programação e *softwares* que são utilizados para a coleta, a manipulação, a análise e a visualização de dados. Acredita-se que essas tecnologias de dados devam ser conhecidas pelos bibliotecários de dados. É necessário que um bibliotecário de dados conheça as principais tarefas e tipos de análise inseridos em *softwares*, como: *OutWitt Hub, Open Refine, Google Fusion*. Esses *softwares* utilizam algoritmos que permitem a realização de diferentes tipos de análise em dados (descrição, diagnóstico, preditiva, prescritiva) os quais permitem aplicar as mais diversas tarefas de estatística em dados como: classificação, *cluster*, regressão, simulação, análise de conteúdo e desenvolvimento de serviços de recomendação.

Além das linguagens de programação e dos *softwares* utilizados em *Data Science*, os bibliotecários de dados devem ter conhecimento sobre *softwares* de elaboração de repositórios de dados digitais. Alguns *softwares*, como *Ckan, Dspace, DataVerse, e-Prints* e *Fedora*, estão no centro dos conhecimentos necessários para que os bibliotecários de dados se familiarizem com processos de gestão, curadoria e acesso aos dados de pesquisa.

Os repositórios de dados de pesquisa são o *locus* de ligação entre os usuários de dados de pesquisa (pesquisadores) e os bibliotecários de dados. Neles os bibliotecários de dados podem oferecer serviços e produtos relacionados ao acesso e à preservação de dados de pesquisa. São plataformas criadas para dar suporte aos dados de pesquisa, preservando o saber resultante do processo de investigação científica.

Nesse sentido, apresentou-se um estudo exploratório e descritivo (**capítulo 3**) sobre o cenário internacional dos RDPGs disponíveis no Registry of Research Data Repositories (re3data) (**ver capítulo 1.1**).

No **capítulo 3**, foi possível alcançar os objetivos específicos (c) realizar uma análise descritiva sobre a composição internacional dos RDPGs; e (d) identificar o corpus de temáticas, países, licenças, *softwares,* instituições e tipos de dados de pesquisa mais utilizados pela área de Geociências.

Com base no **capítulo 3**, é possível afirmar que os Repositórios de Dados de Pesquisa em Geociências (RDPGs) são fontes de dados heterogêneas que possibilitam o acesso e a preservação a uma ampla gama de tipos de dados de pesquisa. Eles armazenam informações/conhecimentos de áreas relacionadas às Ciências Exatas e da Terra como as Ciências Atmosféricas, Geodésicas, Geofísicas, Geológicas e Oceanográficas.

A maioria dos RDPGs está hospedada em países da América do Norte e da Europa. Ao todo, são 604 RDPGs catalogados re3data em (set. 2017), dos quais é possível sumarizar as seguintes características:

a) tipos de dados: documentos de escritório, dados científicos e estatísticos, imagens e dados brutos.

b) os países que possuem mais RDPGS são: USA, DEU, CAN, GBR, AUS e FRA e a língua preponderante nos dados de pesquisa é o inglês;

c) estão inseridos nas seguintes áreas do conhecimento: Ciências atmosféricas e Oceanografia, sendo as principais temáticas: o clima, as mudanças climáticas, a geologia, a hidrologia, o sensoriamento remoto e a oceanografia;

d) em sua maioria, estão em acesso aberto e utilizam algumas das licenças Creative Commons;

e) os *softwares* para desenvolvimento desses repositórios são, em sua maioria, desconhecidos.

Embora a comunidade internacional de bibliotecários incentive a utilização de *softwares* como *Ckan, DSpace, DataVerse, e-Prints* e *Fedora*, eles são pouco utilizados na área das Geociências. Apesar de serem amplamente adotadas por instituições governamentais e comunidades de bibliotecários que defendem o acesso aberto ao conhecimento científico, a elaboração de repositórios de dados de pesquisa com esses *softwares* é pouco difundida. Entre os 604 RDPGs, apenas 37 os utilizam. Alguns exemplos de RDPGS elaborados com esses *softwares* são: *GeoPlatform* (Ckan), *DataShare* (DSpace), o

*UCLA Social Science Data Archive* (*Dataverse*), *OceanRep GEOMAR Repository* (*e-Prints*) e *TAMBORA* (*Fedora*).

Por fim, pode-se concluir o percurso do pensamento proposto nesta tese com algumas recomendações.

Em suma, recomenda-se que os bibliotecários de dados precisam adquirir *know-how* em processos operacionais de coleta, manipulação, análise e visualização de dados. Tal tarefa implica a busca por habilidades inerentes ao cientista de dados e ao cientista da informação. Enquanto cientistas da informação, os bibliotecários de dados devem procurar ressaltar os valores da informação e o processo de geração de conhecimento; enquanto cientista de dados, inquirir os conhecimentos mínimos sobre o uso de linguagens de programação e métodos de pesquisa.

Conforme a International Association for Social Science Information Services And Technology (2016), a formação em Biblioteconomia de Dados está relacionada à formação de nível de pós-graduação em escolas de Biblioteconomia e Ciência da Informação. Já a formação aplicada à *Data Science* deve focar em aprendizados básicos sobre estatística, linguagens de programação, dados de pesquisa repositórios, *softwares* de coleta, manipulação, análise e visualização de dados.

Para conseguir tais habilidades, recomendam-se alguns cursos e/ou formações oferecidos por universidades internacionais, de forma presencial e/ou a distância. Por exemplo, o curso Research Data Management and Sharing, que é oferecido *on-line* de forma gratuita pelo Curating Research Assets and Data Using Lifecycle Education (CRADLE)[154], da Universidade da Carolina do Norte, em Chapel Hill, em parceria com o projeto EDINA[155], da Universidade de Edimburgo. Esse curso oferece uma introdução ao gerenciamento e ao compartilhamento de dados da pesquisa, ciclo de vida dos dados da pesquisa, gestão de dados, melhores práticas para trabalhar com dados (incluindo a organização, a documentação, o armazenamento e a segurança de dados) e repositórios de dados de pesquisa.

Outro curso *on-line* é o Research Data Management Training MANTRA[156]. Oferecido pelo projeto EDINA, da Universidade de

---

[154] CRADLE. Disponível em:<http://cradle.web.unc.edu/>. Acesso em: maio 2017.

[155] EDINA. Disponível em: <https://edina.ac.uk/ >. Acesso em: set. 2017.

[156] MANTRA. Disponível em: <http://mantra.edina.ac.uk/>. Acesso em: set. 2017.

Edimburgo, visa ao gerenciamento de dados de pesquisa. O curso serve tanto para bibliotecários de dados como para outros pesquisadores. Outro curso oferecido pela Universidade Harvard-Smithsonian, no Center for Astrophysics John G. Wolbach, e pela da Harvard Library visa preparar bibliotecários de dados para que adquiram conhecimentos sobre *Data Science*. O *Data Scientist Training for Librarians (*DSTL)[157] tem como objetivo capacitar os bibliotecários junto às atuais demandas e necessidades da gestão de dados, abordando linguagens de programação e métodos de pesquisa para extrair, armazenar, analisar e visualizar dados de pesquisa científica.

Para aprender técnicas estatísticas e linguagens de programação utilizadas em *Data Science*, os bibliotecários de dados podem recorrer a cursos oferecidos por universidades na área de Estatística ou Ciência da Computação. Eles também podem recorrer a formações complementares oferecidas em portais de treinamento *on-line* como Coursera[158], Udemy[159], Data Science Academy[160], EDX[161], entre outros.

Por fim, a Biblioteconomia de Dados é *ad hoc*. A educação formal de bibliotecários de dados não está disponível em todos os países do mundo. Essa característica revela a falta de formação e certificação formal em Biblioteconomia de Dados.

Outra questão está relacionada com a dificuldade em localizar literatura sobre o tema em bases de conhecimento da CI. O que revela que a Biblioteconomia de Dados está em desenvolvimento e carece de atenção por parte dos profissionais da área de CI. Nesse sentido, para adquirir os conhecimentos teóricos mínimos em Biblioteconomia de Dados, recomendam-se algumas obras essenciais.

Para consolidar os conhecimentos sobre CI é fundamental o livro ***Introduction to Information Science***, redigido pelos professores da Universidade de Londres David Bawden e Lyn Robinson (2012). Esse livro reúne diversas abordagens sobre a CI. No que tange a questões ligadas à *e-Science*, o livro ***Fouthr Paradigm: Data-intensive Scientific***

---

[157]DSTL. Disponível em: <http://library.harvard.edu/06112015-1629/registration-open-data-scientist-training-librarians>. Acesso em: maio 2017.

[158] COURSERA. Disponível em:< https://www.coursera.org/>Acesso em: maio 2017.

[159]UDEMY. Disponível em:< https://www.udemy.com/>Acesso em: maio 2017.

[160]DSC. Disponível em:< https://www.datascienceacademy.com.br/>Acesso em: maio 2017.

[161] EDX. Disponível em:< https://www.edx.org/>Acesso em: maio 2017.

***Discovery***, editado em 2009 por Tony Hey, Stewart Tansley e Kristin Tolle. A importância desse texto está no fato de que considera a *e-Science* em todos os seus aspectos, sejam eles relacionados aos avanços científicos e tecnológicos, à comunicação científica ou ao uso intensivo de dados extraídos de sensores tecnológicos.

A temática de dados de pesquisa é revista exaustivamente por Margaret E. Henderson no livro ***Data Management A Practical Guide for Librarians***, de 2017. A autora apresenta os princípios básicos da gestão e das melhores práticas com dados de pesquisa. Outra obra, organizada por Lynda Kellam e Kristi Thompson em 2016, é o livro de artigos ***Databrarianship: the academic data librarian in theory and practice***, que apresenta diversas questões relacionadas à Biblioteconomia de Dados e ao uso de dados de pesquisa pelas distintas disciplinas científicas.

Também sobre Biblioteconomia de Dados, mas com o foco no campo de atuação do bibliotecário de dados, está a obra ***The Data Librarian's Handbook***, escrita em 2016 por Robin Rice e John Southall. Por fim, as questões ligadas à *Data Science* e às habilidades requeridas do cientista de dados aparecem em textos como ***Data scientist: the sexiest job of the 21st century***, publicado na **Harvard Business Review**, em 2012, e escrito por Thomas Davenport e D. J. Patil. Outro texto essencial sobre *Data Science* é **The Data Science Venn Diagram**, elaborado por Drew Conway em 2010. E o último texto essencial a ser lido é "***Big Data at Work: Dispelling the Myths, Uncovering the Opportunities***. Publicado em 2014, trata de todas as questões relacionadas à *Data Science*.

No que tange à realização de trabalhos futuros, sugere-se que os bibliotecários de dados investiguem temas relacionados ao uso de dados de pesquisa em contextos diversos como *big data*, internet das coisas, inteligência artificial, métricas informacionais, comunicação científica, entre outras temáticas. Seria interessante a realização de um estudo sobre as possíveis análises que podem ser aplicadas aos metadados e ao conteúdo depositado em repositórios de dados de pesquisa. Outra forma de abordar os dados de pesquisa em Biblioteconomia de Dados poderia ser a aproximação dos bibliotecários de dados do processo de elaboração de planos de gestão de dados de pesquisa.

Dessa forma, ressalva-se que a Biblioteconomia de Dados deve estar ligada às atuais as práticas com tecnologias digitais, coleta, manipulação, análise e visualização de dados de pesquisa. Segundo recomendações da International Association of Social Science Information Services and Technology (2016), os bibliotecários de dados

devem participar de discussões sobre o compartilhamento de dados, ferramentas de acesso a dados, melhores práticas para a preservação de dados, recursos globais para informação, além de recorrer a questões ligadas a alfabetização estatística, gerenciamento de dados, novas tecnologias, obstáculos ao compartilhamento de dados, formatos de metadados, repositórios digitais, preservação de dados e citação.

A última questão a ser tratada nesta tese diz respeito às ofertas de trabalho para bibliotecários de dados. Até o momento, o autor desta tese, não encontrou oferta de vagas fora dos países da América do Norte e da Europa. Assim, uma sugestão de trabalho futuro poderia ser identificar se, de alguma forma, o perfil profissional dos bibliotecários de dados é requisitado em outras regiões do mundo que não a América do Norte e a Europa.

Nesse sentido, conclui-se esta tese recomendando que os bibliotecários de dados deveriam trabalhar com repositórios de dados de pesquisa, auxiliando pesquisadores desde as primeiras fases da investigação científica, ajudando com o processo de documentação e garantindo que os dados de pesquisa sejam preservados, utilizáveis e reutilizáveis em longo prazo. Um bibliotecário de dados deve usar os valores fundamentais, os princípios éticos, as habilidades e os conhecimentos profissionais da Biblioteconomia para trabalhar com dados de pesquisa, principalmente em repositórios digitais.

Por fim, é preciso um esforço multidisciplinar de aprendizagem, por parte dos bibliotecários de dados, que devem combinar as habilidades e os conhecimentos necessários à Ciência da Informação, à *e-Science* e à *Data Science*.

## REFERÊNCIAS

AMERICAN COUNCIL OF LEARNED SOCIETIES. **Our cultural commonwealth**: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences, 2006. Disponível em: <http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf >. Acesso em: maio 2015.

ANU Data Commons. Disponível em: < https://datacommons.anu.edu.au/DataCommons//>. Acesso em: set. 2017.

APOLLO. Disponível em: <https://www.repository.cam.ac.uk/>. Acesso em: set. 2017.

ARL. Disponível em: <http://www.ala.org/acrl/>. Acesso em: set. 2017.

ATZORI, L.; IERA, A.; MORABITO, G. (2010). The Internet of Things: A survey. **Computer Networks**, vol. 54, n. 15, p. 2787-2805, 2010. Disponível em: <http://www.scirp.org/(S(czeh2tfqyw2orz553k1w0r45))/reference/ReferencesPapers.aspx?ReferenceID=374909>. Acesso em: ago. 2017**.**

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de informação**: conceitos e tecnologia das máquinas de busca. Porto Alegre: Bookman, 2013.

BARGA, R.; FONTAMA, V.; HYONG, T. W. **Predictive Analytics with Microsoft Azure Machine Learning.** 2nd. ed. New York: Springer Science+Business Media New York, 2015.

BAWDEN, D; ROBINSON, L. **Introduction to information science**. London: Facet Publishing, 2012.

BEAUTIFUL Soup Documentation. Disponível em: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>. Acesso em: maio 2016.

BELL, G. Foreword. In: HEY, T.; TANSLEY, S.; TOLLE, K. Fouthr paradigm: Data-intensive Scientific Discovery. **Microsoft**,

2009. Disponível em: <http://research.microsoft.com/en-us/collaboration/fourthparadigm/>. Acesso em: abr. 2015.

BDEP. Disponível em: <http://www.anp.gov.br/wwwanp/exploracao-e-producao-de-oleo-e-gas/dados-tecnicos>. Acesso em: set. 2017.

BOKEH. Disponível em: <http://bokeh.pydata.org/en/latest/>. Acesso em: maio 2016.

BONNEAU, G.-P. et al. Overview and State-of-the-Art of Uncertainty Visualization. In: HANSEN, C. D. et al. (Eds.). **Scientific Visualization**: Uncertainty, Multifield, Biomedical, and Scalable Visualization. London: Springer, Part of Springer Science+Business Media, 2014.

BORKO, H. Information science: what is it? **American Documentation**, v. 19, n. 1, p. 3-5, 1968. Disponível em: <http://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/k—artigo-01.pdf>. Acesso em: 14 ago. 2014.

BRASIL. CONSELHO NACIONAL DE ARQUIVOS (CONARQ). **Resolução nº 39, de 29 de abril de 2014.** Diretrizes para a implementação de repositórios digitais confiáveis de documentos arquivísticos. Rio de Janeiro, 2014. Disponível em: <http://www.conarq.arquivonacional.gov.br/images/publicacoes_textos/diretrizes_rdc_arq.pdf>. Acesso em: ago. 2017**.**

BREMEN CORE REPOSITORY. Disponível em:< https://www.marum.de/en/Research/IODP-Bremen-Core-Repository.html>. Acesso em: set. 2017.

BUCKLAND, M. Information as thing. **Journal of the American Society for Information Science,** v. 42, n. 5, p. 351-36, 1991. Disponível em: <http://people.ischool.berkeley.edu/~buckland/thing.html>. Acesso em: 14 ago. 2014.

CANADIAN CENTRE FOR CLIMATE MODELLING AND ANALYSIS. Disponível em: <http://www.ec.gc.ca/ccmac-cccma/ >. Acesso em: set. 2017.

CAPLAN, P. **Metadata fundamentals for all librarians**. La Vergne: Lightning Source, 2003.

CAPURRO, R. O conceito de informação. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 12, n. 1, p. 148-207, abr. 2007. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/54>. Acesso em: 15 maio 2014.

CCCA Data Centre. Disponível em: <https://data.ccca.ac.at/about>. Acesso em: set. 2017.

CESSDA. Disponível em: <http://www.unidata.unimib.it/?page_id=2136>. Acesso em: maio 2017.

CHAMP. Disponível em: <http://www-app2.gfz-potsdam.de/pb1/op/champ/index_CHAMP.html >. Acesso em: set. 2017.

CHEN, L. **Introduction**: Data Science and BigData Computing. Mathematical Problems in Data Science. Switzerland: Springer International Publishing, 2015. Disponível em: <https://link.springer.com/chapter/10.1007%2F978-3-319-25127-1_1>. Acesso em: 15 maio 2016.

CIESIN. Disponível em: <http://www.ciesin.columbia.edu/data.html/>. Acesso em: set. 2017.

CKAN. Disponível em: <https://ckan.org/>. Acesso em: ago 2017.

CLEVERDON, C. W. ASLIB Cranfield research project on the comparative efficiency of indexing systems. In: ASLIB, 12. **Proceedings**, 421-431, 1960. Disponível em: <http://sigir.org/files/museum/ASLIB%20CRANFIELD%20RESEARCH%20PROJECT-1960/pdfs/frontmatter.pdf>. Acesso em: 15 maio 2015.

CODATA. Disponível em: <http://www.codata.org/about-codata>. Acesso em: maio 2016.

COI. Disponível em: <https://pt.wikipedia.org/wiki/Lista_de_pa%C3%ADses_por_c%C3%B3digo_do_COI>. Acesso em: set. 2017.

CONWAY, Drew. **The Data Science Venn Diagram**. 2010. Disponível em: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>. Acesso em: 15 maio 2015.

CORDEIRO, D. et al.. Da Ciência à e-Ciência: paradigmas da descoberta do conhecimento. **Revista USP**, n. 97, p. 71-80, 2013. Disponível em: <http://www.revistas.usp.br/revusp/article/view/61867/64710>. Acesso em: 15 maio 2015.

CORMEN, T. H; LEISERSON, C; RIVEST, R.; STEIN,C.. **Algoritmos**: teoria e prática. Rio de Janeiro: Campus, 2002.

COSTA, Maira Murrieta; CUNHA, Murilo Bastos da. O bibliotecário no tratamento de dados oriundo da e-Science: considerações iniciais. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 19, n. 3, p. 189-206, jul. 2014. Disponível em: <http://dx.doi.org/10.1590/1981-5344/1900>. Acesso em: maio 2016.

COSTA, Micheli; BRAGA, Tiago. Repositórios de dados de pesquisa no mundo. **Cadernos BAD**, n. 2, p. 80-95, 2016. Disponível em: <https://www.bad.pt/publicacoes/index.php/cadernos/article/view/1585>. Acesso em: 15 set. 2016.

COURSERA. Disponível em: <https://www.coursera.org/>. Acesso em: maio 2017.

COX, A.; VERBAAN, E.; SEN, B. Upskilling liaison librarians for research data management. **Ariadne**, n. 70, 2013. Disponível em: <http://www.ariadne.ac.uk/issue70/cox-et-al>. Acesso em: jun. 2016.

CRADLE. Disponível em: <http://cradle.web.unc.edu/>. Acesso em: maio 2017.

CSV File Reading and Writing. Disponível em: <https://docs.python.org/2/library/csv.html>. Acesso em: maio 2016.

CUGIR. Disponível em: <http://cugir.mannlib.cornell.edu/about.jsp>. Acesso em: set. 2017.

CUPANI, A. **Filosofia da Tecnologia**: um convite. Florianópolis: Ed. da UFSC, 2013.

D-PLACE. Disponível em: < https://d-place.org/angular/>. Acesso em : set. 2017.

DALBEY, J. **Pseudocode Standard**. 2001. Disponível em: <http://users.csc.calpoly.edu/~jdalbey/SWE/pdl_std.html>. Acesso em: set. 2017.

DART Data. Disponível em: <http://dartportal.leeds.ac.uk/ >. Acesso em: set. 2017.

DATA SCIENCE CENTRAL. Disponível em: <http://www.datasciencecentral.com/>. Acesso em: ago. 2017.

DATADOI. Disponível em: <http://datadoi.ut.ee/>. Acesso em: set. 2017.

DATASHARE. Disponível em: <http://datashare.is.ed.ac.uk >. Acesso em: ago 2017.

DATAVERSE. Disponível em: <https://dataverse.org>. Acesso em: ago 2017.

DAVENPORT, T. H. **Big Data at Work**: Dispelling the Myths, Uncovering the Opportunities. Boston: Harvard Business Review Press, 2014.

DAVENPORT, T. H.; PATIL, D. J. Data scientist: the sexiest job of the 21st century. **Harvard Business Review**, v.?, n.?, october, p.70–77, 2012.

DEKKER, H.; LACKIE, P. Technical Data Skills for Reproducible Research. In: KELLAM, L.; THOMPSON, K. **Introduction to Databrarianship**: The Academic Data Librarian in Theory and Practice. Chicago: Association of College and Research Library, 2016.

DLA. Disponível em: <http://darchive.mblwhoilibrary.org/>. Acesso em: set. 2017.

DSPACE. Disponível em: <http://www.dspace.org>. Acesso em: maio 2016.

DSC. Disponível em: <https://www.datascienceacademy.com.br/>. Acesso em: maio 2017.

DSTL. Disponível em: <http://library.harvard.edu/06112015-1629/registration-open-data-scientist-training-librarians>. Acesso em: maio 2017.

DUBLIN CORE. Disponível em: <http://dublincore.org/>. Acesso em: set. 2017.

DUDZIAK, E. **Dados de Pesquisa agora devem ser armazenados e citados**. 2016. Disponível em: <http://www.sibi.usp.br/?p=6189>. Acesso em: 10ago. 2016.

DUSEK, V. **Philosophy the Technology:** the technological condition an anthology. Malden: Blackwell Publishing, 2006.

EDX. Disponível em: <https://www.edx.org/>. Acesso em: maio 2017.

E-PRINTS. Disponível em: <http://www.eprints.org/uk/>. Acesso em: ago 2017.

EDINA. Disponível em: <https://edina.ac.uk/ >. Acesso em: set. 2017.

EDINBURGH DataShare. Disponível em : <http://datashare.is.ed.ac.uk/>. Acesso em: set. 2017.

FAY, D. Earth and Environment: introduction. In: HEY, T.; TANSLEY, S.; TOLLE, K. Fouthr Paradigm: Data-intensive Scientific Discovery. **Microsoft**. 2009. Disponível em: <http://research.microsoft.com/en-us/collaboration/fourthparadigm/>. Acesso em: ago. 2017.

FEDORA. Disponível em: <https://getfedora.org/pt_BR/>. Acesso em: ago 2017.

FEDERER, L. Research data management in the age of big data: Roles and opportunities for librarians. **Information Services & Use**, v. 36, n. 1-2, p. 35-43, 2016. Disponível em: <http://content.iospress.com/articles/information-services-and-use/isu797>. Acesso em: 10 ago. 2017.

FIGSHARE. Disponível em: <http:// figshare.com>. Acesso em: ago 2017.

FMINER. Disponível em: <http://www.fminer.com/>. Acesso em: set. 2017.

FLORIDI, L. Open problems in the philosophy of information. **Metaphilosophy**, v. 35, n, 4, p. 554-582, 2004.

FLORIDI, L. Steps forward in the philosophy of information. **Etica & Politica / Ethics & Politics**, v. 14, n. 1, p. 304-310, 2012. Disponível em: <http://www2.units.it/etica/2012_1/FLORIDI.pdf>. Acesso em: 14 maio 2016.

FLORIDI, L. **Very Short Introduction of Information**. Oxford: Oxford University Press, 2010.

FOX, J. Aspects of the Social Organization and Trajectory of the R Project. **The R Journal,** v. 1, 2, 2009. Disponível em: <https://journal.r-project.org/archive/2009/RJ-2009-014/RJ-2009-014.pdf>. Acesso em: ago. 2017**.**

GARCIA, P. A.; SUNYE, M. O Protocolo OAI-PMH para Interoperabilidade em Bibliotecas Digitais. **CONGED**, 2006. Disponível em: <http://conged.deinfo.uepg.br/~iconged/Artigos/artigo_09.pdf >. Acesso em: 15 maio 2016.

GCR. Disponível em: <http://iodp.tamu.edu/curation/gcr/index.html>. Acesso em: set. 2017.

GENBANK. Disponível em: <https://www.ncbi.nlm.nih.gov/genbank/ >. Acesso em: set. 2017.

GEOPLATFORM. Disponível em: <https://ckan.geoplatform.gov/>. Acesso em: set. 2017.

GEOTHERMAL Data Repository. Disponível em: <https://gdr.openei.org/>. Acesso em: set. 2017.

GERACI, D.; HUMPHREY, C.; JACOBS, J. **Data Basics**: An Introductory Text.
Michigan: ICPSR, 2008.

GILCHRIST, A. D. The information paradox. **Scire**: Representación y Organización del Conocimiento, v. 20. n. 2, p. 13-18, 2014. Disponível em: <http://www.ibersid.eu/ojs/index.php/scire/article/viewFile/4203/3772>. Acesso em: 15 maio 2016.

GLA. Disponível em: <https://data.london.gov.uk/dataset/>. Acesso em: set. 2017.

GLEZ-PEÑA, D. et al.. Web scraping technologies in an API world. **Briefings in Bioinformatics**, v. 15, n. 5, p. 788-797, 2013. Disponível em: <http://bib.oxfordjournals.org/content/15/5/788>. Acesso em: 15 maio 2016.

GOLD, A. Cyberinfrastructure, data, and libraries, part 1: A cyberinfrastructure primer for librarians. **D-Lib Magazine**, 13(9/10), 2007a. Disponível em: <http://www.dlib.org/dlib/september07/gold/09gold-pt1.html>. Acesso em: 15 maio 2016.

GOLD, A. Cyberinfrastructure, data, and libraries, part 2: Libraries and the data challenge: Roles and actions for libraries. **D-Lib Magazine**, 13(9/10), 2007b. Disponível em: <http://www.dlib.org/dlib/september07/gold/09gold-pt2.html>. Acesso em: 15 maio 2016.

GONZALES-AGUILAR, A. et al. Visualização de dados, informação e conhecimento. Florianópolis: Ed. UFSC, 2017.

GOOGLE FUSION TABLES. Disponível em: < https://support.google.com/fusiontables/answer/2571232>. Acesso em: setembro 2017

GOOGLE Trend. Disponível em: <https://trends.google.com/trends/>. Acesso em: ago. 2017.

GRACE. Disponível em: <http://isdc.gfz-potsdam.de/grace-isdc/>. Acesso em: set. 2017.

GRAY, J. Jim Gray on eScience: A Transformed Scientific Method. In: HEY, T.; TANSLEY, S.; TOLLE, K. Fouthr Paradigm: Data-intensive Scientific Discovery. **Microsoft**, 2009. Disponível em: <http://research.microsoft.com/en-us/collaboration/fourthparadigm/>. Acesso em: abr. 2015.

GRECO, J. O que é Epistemologia**.** In: GRECO, J.; SOSA, E. **Compêndio de Epistemologia**. São Paulo: Loyola, 2008.

GRIGG, K. Data in the Sciences. In: KELLAM, L.; THOMPSON, K. **Introduction to Databrarianship:** The Academic Data Librarian in Theory and Practice. Chicago: Association of College and Research Library, 2016.

GROLEMUND, G.; WICKHAM, H. **R for Data Science**, 2017. Disponível em: < http://r4ds.had.co.nz/ >. Acesso em: ago. 2017**.**

HAYASHI, C. What is Data Science? Fundamental Concepts and a Heuristic Example. In: CONFERENCE OF THE INTERNATIONAL FEDERATION OF CLASSIFICATION SOCIETIES (IFCS'96), 5th, 1996, Kobe, Japan. **Proceedings**. Kobe, Japan: IFCS'96, 1996.

HEIDATA. Disponível em: <https://heidata.uni-heidelberg.de/>. Acesso em: set. 2017.

HENDERSON, M. **Data Management**: a pratical guide for librarians. Lanham: Rowman & Littlefield Publishers, 2017.

HEY, T.; HEY, J. e-Science and its implications for the library community. **Library Hi Tech**, vol. 24, n. 4, p. 515-528, 2006. Disponível em: <http://eprints.rclis.org/9202/>. Acesso em: abr. 2015.

HEY, T.; TANSLEY, S.; TOLLE, K. Fouthr Paradigm: Data-intensive Scientific Discovery. **Microsoft**. 2009. Disponível em: <http://research.microsoft.com/en-us/collaboration/fourthparadigm/>. Acesso em: 14 abr. 2015.

ICRISAT. Disponível em: <http://dataverse.icrisat.org/>. Acesso em: set. 2017.

IGETS. Disponível em: <http://dataservices.gfz-potsdam.de/igets/showshort.php?id=escidoc:2387889>. Acesso em: set. 2017.

IMPORT.IO. Disponível em: <https://www.import.io/>. Acesso em: set. 2017.

INDEED. Disponível em: <https://www.indeed.com>. Acesso em: maio 2016.

INTERNATIONAL ASSOCIATION FOR SOCIAL SCIENCE INFORMATION SERVICES AND TECHNOLOGY (IASSIST). **Defining data librarian - call for comments**, 2016. Disponível em: <http://www.iassistdata.org/blog/defining-data-librarian-call-comments>. Acesso em: jun. 2016.

IODP. Disponível em: <https://www.iodp.org/>. Acesso em: set. 2017.

ISD. Disponível em: <http://isdc.gfz-potsdam.de/homepage/>. Acesso em: set. 2017.

ISO16363:2012. Disponível em: <https://www.iso.org/obp/ui/#iso:std:iso:16363:ed-1:v1:en>. Acesso em: set. 2017.

JANKOWSKI, N. W. Exploring e-Science: An Introduction. **Journal of Computer-Mediated Communication**, v. 12, n. 2, 549- 562, 2007. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2007.00337.x/pdf/>. Acesso em: 14 abr. 2015.

KCC. Disponível em: <http://www.kochi-core.jp/en/iodp-curation/index.html >. Acesso em: set. 2017.

KELLAM, L.; THOMPSON, K. **Introduction to Databrarianship***:* The Academic Data Librarian in Theory and Practice. Chicago: Association of College and Research Library, 2016.

KENNAN, M. Data management: Knowledge and skills required in research, scientific and technical organizations. In: IFLA WLIC, 2016. Disponível em: <http://library.ifla.org/1466/1/221-kennan-en.pdf>. Acesso em: jun. 2016.

KHUN, T. **A estrutura das revoluções científicas**. São Paulo: Perspectiva, 2000.

KINDLING, M.; PAMPEL, H. Informationsinfrastrukturangebote für digitale Forschungsdaten. **E(hren)Journal**, p. 15-33, 2 fev. 2017. DOI: 10.18452/2341.

KLINE, J. What is Thecnology? In: DUSEK, V. (Ed.). **Philosophy the Technology**: the technological condition an anthology. Malden: Blackwell Publishing, 2006.

KNIME. Disponível em: <http://www.knime.org/>. Acesso em: set. 2017.

KOLTAY, Tibor. Data literacy for researchers and data librarians. **Journal of Librarianship and Information Science**, v. 49, n. 1, 2017. Disponível em: <http://journals.sagepub.com/doi/abs/10.1177/0961000615616450 >. Acesso em: 4 ago. 2017**.**

LANDMAP. Disponível em: <http://catalogue.ceda.ac.uk/uuid/7f1280cf215da6f8001eae5c2f019fe8/ >. Acesso em: set. 2017.

LHC. Disponível em: <https://home.cern/topics/large-hadron-collider>. Acesso em: set. 2017.

LIBER. Disponível em: <http://libereurope.eu/about-liber/>. Acesso em: set. 2017.

LSHTM Data Compass. Disponível em: <http://datacompass.lshtm.ac.uk/>. Acesso em: set. 2017.

LXML - XML and HTML with Python. Disponível em: <http://lxml.de>. Acesso em: maio 2016.

LYNCH, C. A. Institucional Repositories: essential infraestructure for scholarship in
the digital age. **ARL Bimonthly Report 26**, 2003. Disponível em: <http:// www.arl.org/newsltr/226/ir.html>. Acesso em: 25 out. 2004.

MANOVICH, L. What is visualization. **Visual Studies**, 26, n. 1, mar. 2011. Disponível em: <https://www.sfu.ca/cmns/courses/2011/325/Readings/What_is_Visualization.pdf>. Acesso em: maio 2016.

MANTRA. **Do-It-Yourself Research Data Management Training Kit for Librarians**, 2012-13. Disponível em: <http://mantra.edina.ac.uk/libtraining.html>. Acesso em: ago. 2017**.**

MATPLOTLIB. Disponível em: <http://matplotlib.org>. Acesso em: maio 2016.

MCCAFFREY, M.; GIESBRECHT, W. Teaching Data Librarianship to LIS Students. In: KELLAM, L.; THOMPSON, K. **Introduction to Databrarianship***:* The Academic Data Librarian in Theory and Practice*.* Chicago: Association of College and Research Library, 2016.

MONTEIRO, E. **Direitos autorais nos repositórios de dados científicos**: análise sobre os planos de gerenciamento dos dados. 2017. 118 f. Dissertaão (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista – UNESP, Marília (SP), 2017.

MOORE, G.; MOREAU, L. **From metadata to links**. The Sixth Workshop on Open Hypermedia Systems (OHS 6). Berlin: Springer-Verlag, 2001. Disponível em: <http://eprints.soton.ac.uk/id/eprint/254475>. Acesso em: 15 maio 2012.

MORVILLE, P. **Ambient findability**: what we find changes who we become. Farnham: O'Reilly, 2014.

MSTE. **Welcome to Carolyn's Unit on Graphing.** 1997. Disponível em: <http://mste.illinois.edu/courses/ci330ms/youtsey/intro.html>. Acesso em: jun. 2016.

NATIONAL INSTITUTES OF HEALTH. **NIH Data Sharing Policy and Implementation Guidance**, 2003. Disponível em: < https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm>. Acesso em: ago. 2017**.**

NATIONAL POLLUTANT RELEASE INVENTORY. Disponível em: < http://www.ec.gc.ca/inrp-npri/>. Acesso em: set. 2017.

NATIONAL RESEARCH COUNCIL. **A question of balance**: Private rights and the public interest in scientific and technical databases. Washington, DC: The National Academies Press, 1999. Disponível em: <http://www.nap.edu/read/9692/chapter/1>. Acesso em: maio 2016.

NATIONAL RESEARCH COUNCIL (NRC). **A question of balance**: Private rights and the public interest in scientific and technical databases, 1999. Disponível em: <http://www.nap.edu/read/9692/chapter/1>. Acesso em: jun. 2016.

NATIONAL SCIENCE BOARD (NSB). **Cyberinfrastructure vision for 21st century discovery**, 2005. Disponível em: <http://www.nsf.gov/pubs/2007/nsf0728 >. Acesso em: jun. 2016.

NATIONAL SCIENCE BOARD (NSB). **Digital Research Data Sharing and Management**, 2011. Disponível em: < https://www.nsf.gov/nsb/publications/2011/nsb1124.pdf>. Acesso em: ago. 2017**.**

NATURE. Disponível em: <https://www.nature.com/sdata/policies/data-policies >. Acesso em: set. 2017.

NAUR, P. Disponível em: <https://en.wikipedia.org/wiki/Peter_Naur>. Acesso em: maio 2016.

NCEDC. Disponível em: <http://ncedc.org/_>. Acesso em: set. 2017.

BURLINGAME, N; NIELSEN, L. **A Simple Introduction to Data Science**. Wickford: New Street Communications, 2012.

NUMPY. Disponível em: <http://www.numpy.org>. Acesso em: maio 2016.

OCEANREP GEOMAR Repository. Disponível em: <http://oceanrep.geomar.de/>. Acesso em: set. 2017.

OCTOPARSE. Disponível em: <https://www.octoparse.com/>. Acesso em: set. 2017.

OED Online. **datum, n.**, 2015.Oxford University Press. Disponível em: <http://www.oed.com/view/Entry/47434 >. Acesso em: jun. 2016.

ONLINE COMPUTER LIBRARY CENTER – OCLC. **Trustworthy Repositories Audit & Certification**: Criteria and Checklist. 2007. Disponível em: http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying-0. Acesso em: set. 2017.

OPEN ARCHIVES INITIATIVE. **Protocol for Metadata Harvesting** (OAI-PMH). Disponível em: <http://www.openarchives.org/pmh/>. Acesso em: 2 maio 2014.

OPENREFINE. Disponível em: <https://github.com/OpenRefine/OpenRefine>. Acesso em: maio 2016.

ORGANIZATION for Economic Co-operation and Development (OECD). **Declaration on Access to Research Data from Public Funding Organization for Economic Co-operation and Development** (OECD), 2004. Disponível em: <http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157#_ftn0>. Acesso em: 5 out. 2016.

ORANGE. Disponível em: <https://orange.biolab.si/>. Acesso em: set. 2017.

OTLET. P. **Traité de documentation**: le livre sur le livre: théorie et pratique. Bruxelles: Mundaneum, 1934.

OUTWIT HUB. Disponível em: <http://www.outwit.com/>. Acesso em: set. 2017.

PAN-STARRS. Disponível em: <https://panstarrs.stsci.edu/>. Acesso em: set. 2017.

PANDAS. Disponível em: <http://pandas.pydata.org/>. Acesso em: maio 2016.

PANGEA. Disponível em: <http://www.pangaea.de >. Acesso em: ago. 2017.

PAMPEL, H. et al. Making research data repositories visible: the re3data.org
Registry. **PLoS One**, San Francisco, v. 8, n. 11, 2013. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3817176/>. Acesso em: 15 maio
2016.

PARSEHUB. Disponível em: <https://www.parsehub.com/>. Acesso em: set. 2017.

PINHEIRO, L. V. R.; LOUREIRO, J. M. M. Traçados e limites da ciência da informação. **Ciência da Informação**, Brasília, v. 24, n. 1, p. 42-53, 1998. Disponível em: <http://revista.ibict.br/ciinf/article/view/609/611>. Acesso em: 14 ago. 2014.

PIORUN, M. E. **E-Science as a Catalyst for Transformational Change in University Research Libraries**. 2013. 369f. Tese (Doutorado em Filosofia) – Faculty of the Simmons College Graduate School of Library and Information Science. University of Massachusetts Medical School, 2013. Disponível em: <http://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1151&context=lib_articles>. Acesso em: 14 ago. 2016.

PLOS. Disponível em: <http://journals.plos.org/plosone/s/data-availability#loc-depositing-data>. Acesso em: set 2017.

PYTHON. Disponível em: <https://www.python.org/>. Acesso em:

maio 2016.

R-PROJECT. Disponível em: <https://www.r-project.org/>. Acesso em: ago. 2017.

RAPIDMINER. Disponível em: <http://sourceforge.net/projects/rapidminer/>. Acesso em: set. 2017.

RE3DATA. Disponível em: <http://www.re3data.org/about>. Acesso em: ago 2017.

RESMINI, A.; ROSATI, L. **Pervasive Information Architecture.** Designing Cross-Chanel User Experiences. Burlington: Elsevier, 2011.

RICE, R.; SOUTHALL, S. **The data librarian's handbook**. London: Facet Publishing, 2016.

RODRIGUES, E. et al. **Os repositórios de dados científicos**: estado da arte. 2010.
Disponível em:
<http://projeto.rcaap.pt/index.php?option=com_remository&Itemid=2&func=startdown
&id=271&lang=pt>. Acesso em: 5 jun. 2016.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/viewFile/235/22>. Acesso em: 14 ago. 2016.

SARACEVIC, T. Information Science. In: BATES, M. J. ; Maack, M. N. (Eds.) **Encyclopedia of Library and Information Science**. New York: Taylor & Francis, 2010. p. 2570-2586. Disponível em: <https://comminfo.rutgers.edu/~tefko/SaracevicInformationScienceELIS2009.pdf>. Acesso em: 14 ago. 2016.

SAUNDERS, T. Data Science and Data Scientists: What's in a Name? **Information Management**, nov. 11 2013. Disponível em: <http://www.cbigconsulting.com/wp-content/uploads/2014/03/data-scientist.pdf>. Acesso em: 14 ago. 2016.

SAYÃO, M. **Tipos de gráficos**. 200? Disponível em: <www.inf.pucrs.br/~cnunes/ferramentas/Aulas/tipos_de_graficos.pdf>. Acesso em: jun. 2016. Notas de aula.

SAYÃO, L. F.; SALES. L. F. **Guia de gestão de dados de pesquisa para bibliotecários de pesquisadores**. Rio de Janeiro: CNEN, 2015. Disponível em: <http://carpedien.ien.gov.br:8080/bitstream/ien/1624/1/GUIA_DE_DADOS_DE_PES
QUISA.pdf>. Acesso em: 5 out. 2016.

SCHOLTZ, N. Supporting Geospatial Data. In: KELLAM, L.; THOMPSON, K. **Introduction to Databrarianship**: The Academic Data Librarian in Theory and Practice. Chicago: Association of College and Research Library, 2016.

SCIPY. Disponível em: <https://www.scipy.org>. Acesso em: maio 2016.

SDDB. Disponível em: <http://www.scientificdrilling.org >. Acesso em: ago 2017.

SEABORN: statistical data visualization. Disponível em: <https://web.stanford.edu/~mwaskom/software/seaborn/>. Acesso em: maio 2016.

SEAFLOOR SEDIMENTS DATA COLLECTION. Disponível em: < http://gis1server.whoi.edu/website/SC2k/viewer.htm >. Acesso em: set. 2017.

SEMELER, A.; PINTO, A.; VIANNA, W. E-science: An epistemological analysis based on the philosophy of technology. **IFLA Journal**, v. 43, n. 2, p. 198-209, 2017. Disponível em: <http://journals.sagepub.com/doi/10.1177/0340035216678235#articleCitationDownloadContainer>. Acesso em: ago. 2017**.**

SEMIDÃO, R. A. M. **Dados, Informação e Conhecimento enquanto elementos de compreensão do universo conceitual da Ciência da Informação**: contribuições teóricas. Marília, 2014. 198 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Informação, Faculdade de Filosofia e Ciências, Universidade Estadual Paulista –

UNESP, Marília, 2014. Disponível em: < https://repositorio.unesp.br/bitstream/handle/11449/110783/000799485.pdf?sequence=1>. Acesso em: ago. 2017.

SHANNON, C.; WEAVER, W. **The mathematical theory of communication**. Urbana: University of Illinois Press, 1971.

SHCMILLEN, H. **Library and Information Science Education and eScience**: e Current State of ALA Accredited MLS/MLIS Programs in preparing Librarians and Information Professionals for eScience Needs. Denver: Capstone Projects, Paper 1, 2015. Disponível em: <http://digitalcommons.du.edu/lis_capstone/1>. Acesso em: maio 2016.

SILVA, A.; RIBEIRO, F. Information Science and Philosophy of Information: Approaches and Differences. In: DEMIR, H. (Ed.). **Luciano Floridi's Philosophy of Technology**: Critical Reflections., Heidelberg: Springer Science+Business Media Dordrecht, 2012. (Philosophy of Engineering and Technology 8).

SMITH, F. J. Data science as an academic discipline. **Data Science Journal**, n. 5, p. 163-164, 2006. Disponível em: <http://datascience.codata.org/articles/abstract/10.2481/dsj.5.163/>. Acesso em: 14 ago. 2016.

SOCHATS, K. M.; MORSE, E. Visualization. In: WILLIANS, J. G.; CARBO, T. (Eds.). **Information Science**: Still an Emerging Discipline. Pittsburgh, PA: Cathedral Publishing, 1997.

SOEHNER. C.; STEEVES, C.; WARD, J. **E-science and data support services**: A study of ARL member institutions. Washington, DC: Association of Research Libraries, 2010.

TABLEAU PUBLIC. Disponível em: <http://www.tableausoftware.com/public/>. Acesso em: set. 2017.

TAMBORA. Disponível em: <https://www.tambora.org/>. Acesso em: set. 2017.

TENOPIR, C. et al. Research Data Services in European Academic Research Libraries. **LIBER Quarterly**, v. 27, n.1, p. 23–44, 2017. Disponível em:

<https://www.liberquarterly.eu/articles/10.18352/lq.10180/>. Acesso em: jun. 2017.

TENOPIR, C. SANDUSKY, RJ. ALLARD, S.; BIRCH, B. Research data management services in academic research libraries and perceptions of librarians. **Library & Information Science Research**, v. 36, p. 84–90, 2014. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0740818814000255 >. Acesso em: jun. 2016.

TIBBO, H.; JONES, S. **Research data management and sharing**, 2017. Disponível em: <https://www.coursera.org/learn/data-management>. Acesso em: jun. 2017.

UCD Digital Library. Disponível em: <http://digital.ucd.ie/>. Acesso em: set. 2017.

UCLA Social Science Data Archive Dataverse. Disponível em: < https://dataverse.harvard.edu/dataverse/ssda_ucla/>. Acesso em: set. 2017.

UDEMY. Disponível em: <https://www.udemy.com/>. Acesso em: maio 2017.

UIT OPEN Research Data Dataverse. Disponível em: < https://dataverse.no/dataverse/uit/>. Acesso em: set. 2017.

UNIVERSITY OF LEICESTER. **Student Learning Development**. 2009. Disponível em: <http://www2.le.ac.uk/offices/ld/mathematics-help-centre>. Acesso em: jun. 2016.

VALLVERDÚ, J. Computational Epistemology and e-Science: a new way to thinking. **Minds and Machines**, v. 19, n. 4, p. 557-567, 2009. Disponível em: <https://www.academia.edu/493057/Computational_Epistemology_and_e-Science_A_New_Way_of_Thinking?auto=download >. Acesso em: 14 ago. 2016.

VAN DER AALST, W. M. P. Data Scientist: The Engineer of the Future. In: MERTINS, K. et al. (Eds.). **Enterprise Interoperability VI**, 13 Proceedings of the I –ESA Conferences 7, 2014.

VOULGARIS, Z. **Data scientist**: the definitive guide to becoming a data scientist. Basking Ridge: Technics Publications, 2014.

WEB Scraper. Disponível em < https://chrome.google.com/webstore/detail/web-scraper/jnhgnonknehpejjnehehllkliplmbmhn>. Acesso em: setembro 2017.

WEBSTER, S. **What Is Scraping**? The Basics For Everyone. 2015. Disponível em: https://myhelpster.com/what-is-scraping-the-basics-for-everyone/. Acesso em: set. 2017.

WERSIG, G. Information Science: the study of postmodern knowledge usage. **Information Processing & Management**, v. 29, n. 2, p. 229-239, 1993.

WHOI Ship Data-Grabber System. Disponível em: < http://4dgeo.whoi.edu/shipdata/index.html />. Acesso em: set. 2017.

WITT, M.; HORSTMANN, W. International approaches to research data services in libraries. **IFLA Journal**, v. 42, n. 4, p. 1-2, 2016. Disponível em: <https://www.ifla.org/files/assets/hq/publications/ifla-journal/ifla-journal-42-4_2016.pdf>. Acesso em: ago. 2017**.**

INTERAGENCY WORKING GROUP ON DIGITAL DATA. **Harnessing the Power of Digital Data for Science and Society**, 2009. Disponível em: <https://www.nitrd.gov/about/harnessing_power_web.pdf >. Acesso em: 4 ago. 2015.

WORLD DATA CENTRE FOR GLACIOLOGY. Disponível em: <https://www.spri.cam.ac.uk/about/>. Acesso em: set. 2017.

XIA, J.; WANG, M. Competencies and Responsibilities of Social Science Data Librarians: An Analysis of Job Descriptions. **College & Research Libraries**, v. 75, n. 3. p. 362-388, 2014. Disponível em: <http://crl.acrl.org/content/75/3/362.abstract>. Acesso em: 4 ago. 2015**.**

YORK UNIVERSITY. **Geospatial Data**: What is Geospatial Data?, 2017. Disponível em:

<http://researchguides.library.yorku.ca/c.php?g=679467&p=4793119>. Acesso em: ago. 2017**.**

ZHU, Y.; XIONG, Y. Towards Data Science. **Data Science Journal**, 14, p. 8, 2015. Disponível em: <http://doi.org/10.5334/dsj-2015-008>. Acesso em: jun. 2016.