Anabela Pires Duarte

OS DADOS DE INVESTIGAÇÃO E O PERFIL DOS PROFISSIONAIS DA INFORMAÇÃO

UMA REVISÃO DA LITERATURA

Dissertação de Mestrado em Ciência da Informação, orientada pela Professora Doutora Maria Manuel Borges, apresentada ao Departamento de Filosofia, Comunicação e Informação, da Faculdade de Letras da Universidade de Coimbra.

julho de 2022

FACULDADE DE LETRAS

OS DADOS DE INVESTIGAÇÃO E O PERFIL DOS PROFISSIONAIS DA INFORMAÇÃO

UMA REVISÃO DA LITERATURA

Ficha Técnica

Tipo de trabalho	**Dissertação**
Título	**Os dados de investigação e o perfil dos profissionais da informação**
Subtítulo	Uma revisão da literatura
Autor/a	Anabela Pires Duarte
Orientador/a(s)	**Professora Doutora Maria Manuel Borges**
Júri	**Presidente: Professora Doutora Maria Cristina Vieira de Freitas** **Vogais:** **1. Professor Doutor Jorge Manuel Rias Revez** **2. Professora Doutora Maria Manuel Borges**
Identificação do Curso	**2º Ciclo em Ciência da Informação**
Área científica	**Ciência da Informação**
Data da defesa	**14-julho-2022**
Classificação	**17 valores**

1 2 9 0 FACULDADE DE LETRAS UNIVERSIDADE D COIMBRA

Agradecimentos

À Professora Doutora Maria Manuel Borges, em particular, a minha gratidão pela disponibilidade, incentivo e sabedoria, basilares no meu percurso. Farão sempre parte de mim as aprendizagens académicas e humanas que fiz, sob a sua orientação.

Ao corpo docente do Mestrado em Ciência da Informação, o meu agradecimento pela dedicação na minha formação e o apoio dado aos alunos.

À Dr.ª Sofia Gomes a minha gratidão pelo acolhimento amigo, franco, generoso e pelos preciosos ensinamentos que me deu no estágio.

À Doutora Paula Castilho, por fazer parte da minha vida.

À Paulinha, por ser parte de mim.

Aos meus filhotes, Vasco e Gaspar, os meus maiores companheiros.

SUMÁRIO

LISTA DE FIGURAS vii

INTRODUÇÃO 1

1 O 4º PARADIGMA DA CIÊNCIA, "DATA-INTENSIVE SCIENCE" 9

1.1 DILÚVIO DE DADOS/ BIG DATA 17

1.2. MINERAÇÃO DE DADOS: UMA DAS FERRAMENTAS PARA LIDAR COM O FENÓMENO *BIG DATA* 20

1.3. LONG TAIL 21

2 DADOS DE INVESTIGAÇÃO: DEFINIÇÃO E TAXONOMIA 24

2.1. USO, REUTILIZAÇÃO E PARTILHA: ANÁLISE, CONCEITOS E PROBLEMAS 30

2.2. PRINCÍPIOS FAIR 37

2.3. PLANO DE GESTÃO DE DADOS (PGD) 41

2.4. METADADOS 47

3 INICIATIVAS EM LITERACIA DOS DADOS DE INVESTIGAÇÃO 50

3.1. COMPETÊNCIAS *CORE* PARA OS PROFISSIONAIS DE INFORMAÇÃO 53

3.2 BOAS PRÁTICAS EM OBSERVAÇÃO 58

CONCLUSÃO 66

REFERÊNCIAS BIBLIOGRÁFICAS 72

Resumo

A investigação científica tem repercussões sociais e económicas na sociedade civil e empresarial. É uma constatação que obriga a uma redefinição de todo o paradigma do processo científico, em que os dados de investigação assumem um papel preponderante. Falamos de Ciência Aberta, um movimento de cooperação e partilha no processo científico, que toma sob o seu escopo todas as práticas de acesso aberto, dados abertos, código aberto e investigação replicável aberta.

O objetivo principal deste trabalho é identificar o perfil necessário ao profissional da informação para se afirmar neste contexto. Para o cumprimento deste objetivo identificaram-se, como objetivos específicos, o mapeamento da literatura internacional que incide quer nesta matéria quer na identificação das ações levadas a cabo nas bibliotecas e dos atores que têm desenvolvido ações de literacia de dados.

É um tema emergente, com traços ainda dúbios no que concerne às boas práticas, que exige um delineamento robusto e normativo de estratégias. Definiram-se o que são dados de investigação, apresentando as teorias mais relevantes, numa breve visão transversal às várias áreas académicas. Na senda das práticas no panorama internacional, fez-se uma análise de como são, atualmente, tratados os dados de investigação em Portugal, quem são os seus agentes e que competências devem desenvolver, com particular incidência nos profissionais da área da Ciência de Informação, como suporte à elaboração de Planos de Gestão de Dados ao abrigo das exigências por parte das Agências de financiamento à Ciência, e a receção de normas e preceitos por parte dos investigadores, num mecanismo de interoperabilidade e cooperação entre os vários *stakeholders*.

Palavras-chave: Dados de investigação; E-Science; Bibliotecas académicas; Plano de Gestão de Dados

Abstract

Scientific research has social and economic repercussions on civil and business society. It is a realization that requires a redefinition of the entire paradigm of the scientific process, in which research data assume a preponderant role. We are talking about Open Science, a movement of cooperation and sharing in the scientific process, which takes under its scope all practices of open access, open data, open source and open replicable research.

The main objective of this work is to identify the profile necessary for the information professional to assert himself in this context. To achieve this objective, specific objectives were identified: the mapping of the international literature on this matter, the identification of actions carried out in libraries and the actors that have developed data literacy actions.

It is an emerging theme, with still dubious traits regarding good practices, which requires a robust and normative design of strategies. We will define what research data are, presenting the most relevant theories, in a brief overview of the various academic areas. In the wake of practices on the international scene, we will analyze how research data are currently treated in Portugal, who their agents are and what skills they should develop, with a particular focus on professionals in the field of Information Science, as a support of an elaboration of Data Management Plans under the requirements of Science funding agencies, and the reception of standards and precepts by researchers, in a mechanism of interoperability and cooperation between the various stakeholders.

Keywords: Research data; E-Science; Academic libraries; Data Management Plan

LISTA DE FIGURAS

Figura 1 “Primeiras páginas das edições inaugurais dos dois primeiros periódicos científicos do mundo, ambos inaugurados em 1665, o francês “Jornal des Sçavans” (6.jan) e o britânico “Philosophical Transactions of the Royal Society” (6.mar). Imagem Maurício Tuff *in* As primeiras revistas científicas _____ 9
Figura 2 Slide courtesy Ian Foster *in* (Borgman, 2010) _____ 12
Figura 3 Diapositivo apresentado na exposição feita por Jim Gray à NRC-CSTB1 na Mountain View, CA, em 11 de janeiro, 2007 in (Gray, 2007) _____ 13
Figura 4 Gerada pelos autores em WordClouds.com _____ 17
Figura 5 Long Tail of Data in Institute for Empowering Long Tail Research: A study funded by the NSF in https://sites.google.com/site/ieltrconcept/home _____ 22
Figura 6 Research Life Cycle (University of California, Irvine, Libraries, Digital Scholarship Services, 2019) in (Borgman, 2019) _____ 33
Figura 7 Grau de familiaridade com os Princípios FAIR in (van Selm, 2020) _____ 41
Figura 8 Disciplinas financiadas por agências com requisitos de PGD in (Williams et al., 2017) _____ 45
Figura 9 O papel dos bibliotecários na investigação intensiva de dados in Griffin (2013) ___ 59
Figura 10 A quem os investigadores pedem ajuda para lidar com os dados? (van Selm, 2020) _____ 63
Figura 11 e-Research Life Cycle and data curation in (Lyon, 2007) _____ 63

INTRODUÇÃO

> One of the diseases of this age is the multiplicity of books; they doth so overcharge the world that it is not able to digest the abundance of idle matter that is every day hatched and brought forth into the world. (Barnaby Rich, 1613 *apud* Price, 1986)

Cerca de cinquenta anos antes da primeira publicação científica num periódico, Barnaby Rich (1580-1617) fez um desabafo que, inclusive, deu origem ao chamado "The Barnaby Rich effect", definido como uma alta produção de escritos científicos acompanhados de reclamações sobre a produtividade excessiva de outros autores (Braun & Zsindely, 1985). Diz-nos Price (1986) que a publicação em 1665 de pequenos trabalhos científicos em periódicos foi uma inovação marcante na vida da ciência, com alguma resistência por parte dos cientistas, por considerarem as publicações periódicas menos dignificantes do que monografias. A comunicação da ciência já marcava, à época, uma necessidade não intuída pelos cientistas, imbuídos de um certo conservadorismo e resistência à partilha, julgando proteger assim as suas descobertas.

Price recorda "The Invisible College", expressão que deriva de um grupo de pessoas, em meados do século XVII, que mais tarde se organizaram como a Royal Society of London. Procuravam visibilidade e certificação entre pares, assegurar prioridade aos seus trabalhos de investigação e estarem informados sobre o trabalho feito por outros, colaborando entre si na investigação. Um esboço de alguns pilares na nova forma de fazer e comunicar ciência.

A comunicação revelou ter um papel central na ciência, sendo a publicação considerada a sua face visível (Borges, 2017). Francis Crick (Prémio Nobel pela descoberta da estrutura molecular do ADN) afirmou numa entrevista dada à BBC que a "comunicação é a essência da ciência" (Garvey, 1979).

> A publicação constitui, assim, o objectivo último da ciência, ligando-a, deste modo e estreitamente, ao processo de tornar públicos os resultados da investigação (CRONIN, 2005, p. 11). Karin Knorr-Cetina explicita esse sentido, tornando-o sinónimo do processo de comunicação: a comunicação é dita ser intrínseca à ciência pelo facto de a ciência moderna ser um empreendimento colectivo que depende de os resultados obtidos por cientistas individuais serem retomados por outros cientistas que se fundam

> neles e os desenvolvem. A ciência projeta-se a si mesma no futuro através da comunicação (KNORR-CETINA, 1999, p. 378). (Borges, 2017)

Após a II Guerra Mundial, com o avanço computacional e tecnológico e o chamado dilúvio de dados - "data deluge" (Hey & Trefethen, 2003; Lord & Macdonald, 2003; Borgman *et al.*, 2006; Borgman *et al.*, 2007) acentuam-se os debates e propostas conceptuais[1]. "We are living in the information age" is a popular saying; however, we are actually living in the data age (Han *et al.*, 2012).

A definição do conceito de informação foi um tema central e muito debatido no contexto da Ciência da Informação (Machlup & Mansfield, 1983; Vakkari & Cronin, 1992; Aspray 1985; Fischer 1993; Wellisch 1972; Wersig & Neveling 1975; Cornelius, 2002; Capurro & Hjørland, 2003 *apud* Bates, 2005), mas começou a emergir um fenómeno muito mais impactante: o crescimento exponencial de dados e com ele a procura de soluções teóricas e práticas (Borgman, 2009; Kowalczyk & Shankar, 2011).

Os autores pós-popperianos Kuhn (1970) e Lakatos & Musgrave (1980), ainda que com alguns pontos de divergência concetual, contribuíram para a evolução científica, apresentando um novo conceito: o paradigma na ciência. Nesta nova visão do mundo da ciência, um conjunto de ideias torna-se dominante porque representa uma explicação plausível para os fenómenos observados, ganhando força através do contributo de cada cientista individual (Hansen et al., 2009). Os cientistas deverão munir-se de critérios bem definidos para avaliar de que forma os novos conhecimentos contribuem para o aumento ou diminuição da memória histórica da prática científica. Até aqui, na chamada ciência normal, tradicional, alguns conceitos, teorias na ciência tomavam o lugar quase de dogmas, que perante um colapso revolucionário, davam lugar a novas teorias. Na sua visão o desenvolvimento científico implica a ocorrência de três ou mais componentes: paradigma, ciência normal e colapso revolucionário, cuja evolução concorre para as mudanças de paradigma e para novos olhares sobre a ciência. Na noção de paradigma está implícita a continuidade, o contributo dos cientistas para um paradigma, seja para lhe dar mais robustez ou enfraquecer. Mas não há uma rutura e sim um desenvolvimento quase harmonioso em que toda a comunidade científica tem

[1] Concepções de dados, informação, conhecimento (McDonough, 1963; Hayes, 1969; Borko *et al.*, 1971; Brillouin, 2013).

lugar e voz. A força do novo conceito não está muito afastada do que Kuhn chama de "normal science", ou seja, homens cuja pesquisa é baseada em paradigmas partilhados, estão comprometidos com as mesmas regras e padrões para a prática científica. Esse compromisso e o aparente consenso que ele produz são pré-requisitos para a ciência normal, ou seja, para a génese e continuação de uma determinada tradição de pesquisa (Kuhn, 1970; Kuhn, 1976).

Este conceito fomentou novas evoluções de pensamento em outros autores, onde se inclui Jim Gray, o responsável pelo chamado "quarto paradigma da ciência". Jim Gray considerava a "ciência intensiva em dados" ou *e-Science* como um "quarto paradigma" da ciência - empírica, teórica, computacional e agora baseada em dados (Jim, 2007; Bell *et al.*, 2009; Hey *et al.*, 2009). Com esta nova ciberinfraestrutura, alteraram-se de igual modo os modelos de colaboração na ciência (Hey & Trefethen, 2002; Borgman, 2010).

O dilúvio de dados ganha expressão no fim da II Guerra Mundial num investimento reforçado, dado pelas agências governamentais à ciência, com o propósito de assegurar a segurança nacional por via da evolução tecnológica. Após a guerra o volume da produção de dados mostrou-se cada vez mais denso, acompanhado por um desenvolvimento muito rápido e cada vez mais eficiente e complexo de novas formas de instrumentação e computadores de alto desempenho (Hey & Trefethen , 2003). O impacto do dilúvio de dados nas chamadas *Big Science* e *Litle Science* são objeto de preocupação pelo desequilíbrio que pode causar no acesso aos meios de tratamento de dados (Borgman et al., 2007). Atualmente acresce o problema do ciberespaço e a conceção de dados sem referência no mundo natural, numa produção cada vez mais intensa (Zhu & Xiong, 2015). *Big Data* passou a ser a nomenclatura deste dilúvio de dados, cujas características e dimensões ficaram conhecidas pelos 3V's (Volume, variedade e velocidade) (Chen et al., 2012, Kwon et al., 2014; Laney, 2001), passando mais tarde a 5V's (veracidade e valor) (Hashem et al., 2014; Elragal, 2014; Fadiya et al., 2014; Yang et al., 2014; Panneerselvam et al., 2015; Yaqoob et al., 2016).

Uma das ferramentas que tem representado uma forma muito eficaz de busca de informações em grandes volumes de dados é a mineração de dados, originária da ciência da computação, consolidou-se na ciência de dados e outras áreas a partir da década de 1980 (Gomes et al., 2019).

Neste universo de dados convivem os projetos de investigação de maior relevo, mas em menor número e os projetos de menor dimensão e expressão a que se chama cauda longa (*Long Tail*). Na sua maioria são projetos com financiamento reduzido e equipas de pequena dimensão. Os dados de todo o processo de investigação não são devidamente tratados, correndo o risco de se perderem totalmente. A pequena dimensão destes projetos não retira o valor intrínseco científico dos dados que deles resultam (Heidorn, 2008). A NSF (National Science Foundation) financiou um projeto (Borgman et al., 2016) que procurou soluções para projetos na cauda longa, tanto a nível de infraestrutura tecnológica, como ao nível de alteração de mentalidades e práticas (Borgman, Wallis e Enyedy, 2007; Eddy, 2005; Edwards et al., 2011; Steinhardt e Jackson, 2014).

Percebamos então a origem de todo este movimento de *eScience*, 4º paradigma, pesquisa intensiva de dados, acesso aberto. O que são dados de investigação? As definições percorrem a sua etimologia, o seu significado em obras de referência conceituadas, e as propostas, algumas divergentes, de importantes autores na Ciência da Informação (Wellisch, H., 1972; East, H., 1983; Buckland, 1991). O estado dinâmico do conceito de dados é visível, reveste-se de grande plasticidade face a novas formulações, como a de Shannon (Bates & Maack, 2009) e novos entendimentos (Bechhofer et al. 2010; Drucker, 2011; Burton & Jackson 2012; Gitelman 2013; Ribes & Jackson 2013; Vertesi & Dourish 2011; Edwards, 2013; Zhu & Xiong, 2015). Há correntes com diferentes propostas para uma conceção de dados (Zimmerman, 2008; Griffin, 2013; Príncipe et al., 2021) e optámos por adotar a proposta de Borgman (2010) que os classifica em 4 categorias: observacionais, computacionais, experimentais e registos.

A exigência pelas agências de financiamento (públicas e privadas) da partilha de dados, conduziu a uma maior atenção ao seu tratamento e curadoria. No contexto da Ciência Aberta (CA), frequentemente caracterizada como um guarda-chuva que alberga debaixo do seu escopo vários movimentos que têm por objetivo de remover as barreiras para o compartilhamento de qualquer tipo de produção, recursos, métodos ou ferramentas, em qualquer etapa do processo de pesquisa, levantam a questão do uso, reutilização e partilha dos dados de investigação. Passam a ser uma prioridade nas preocupações de quem faz ciência e depende de financiamento (Latham, 2017). Como tal percebeu-se a importância de formular um ciclo de vida dos dados (Goodman et al.,

2014) e o estabelecimento de procedimentos normativos sobre este fluxo (Wallis et al., 2012).

No processo de investigação não há um consenso sobre a opção de apresentar os dados juntamente com as publicações, e nem todos os académicos se regem pelas mesmas diretrizes no que toca à gestão dos dados das suas publicações (Perrier *et al.*'s ,2020.; Aydinoğlu *et al.*, 2017; Diekmann, 2012; Hall, 2013; Zhu, 2019; Ali-Khan *et al.*, 2017; Wynholds *et al.*, 2011; Hickson *et al.*, 2016; Laine's, 2017; citados por Thoegersen, J. L., & Borlund, P., 2021). No entanto, os cientistas já demonstram ter a noção de que, para além da sua investigação, têm de considerar questões como o controlo de dados, direitos de autoria e incentivos para gerir e partilhar os dados, como parte integrante do processo de investigação (Borgman *et al.*, 2007; Borgman, 2012; Wallis *et al.*, 2010). O uso, partilha e reutilização dos dados coletados numa investigação, estende-se numa baliza temporal muito para além do *terminus* de uma investigação (Karasti *et al.*, 2006; Karasti *et al.*, 2010).

Uma boa gestão dos conjuntos de dados deverá seguir os princípios FAIR. O acrónimo significa: Findable, Accessible, Interoperable e Reusable (Localizável, Acessível, Interoperável, Reutilizável) e pretendem ser um conjunto de princípios orientadores e práticas aceites pela comunidade, numa linguagem e procedimentos acessíveis, claros e globais tanto pelos humanos, como por sistemas computadorizados (van Selm, 2020). Os princípios FAIR são um dos requisitos na formulação dos PGD (Planos de Gestão de Dados) exigidos pela maioria das agências de financiamento da ciência. A literatura permite acompanhar os requisitos dos diferentes tipos de agências de financiamento, de forma bastante elucidativa (Wang et al., 2012; Huang & Huang, 2018; Mejia & Kajikawa, 2018; (Holbrook & Frodeman, 2011; Wang & Shapira, 2011).

O PGD já é uma realidade difundida e praticada por muitas instituições de investigação, ainda que haja algumas falhas na aceitação do seu valor em algumas comunidades (Nitecki & Davis, 2017). Um PGD permite organizar os dados, mantendo-os seguros e garantindo o acesso a quem precisa (Borgman et al., 2007). As práticas na aplicação dos princípios FAIR e da criação de PGD's alargaram-se a muitos países desde a iniciativa da NSF, em 2011, exigindo a todas as propostas de financiamento, de qualquer tamanho ou diretoria, um PGD. Os metadados, os "dados sobre os dados" são igualmente essenciais e devem descrever o formato e o conteúdo do conjunto de dados,

as circunstâncias da sua recolha, procedimentos usados para os manipular ou modelar, custódia, a sua qualidade, informações de preservação e descrição específica da disciplina (Shankar, 2003).

Perante este cenário tecnológico e de produção intensiva de dados, e focando a atenção nos profissionais de informação, analisámos várias iniciativas de literacia em dados de investigação, relatados de forma crítica, permitindo daí deduzir orientações a pôr em prática em outras instituições. (Carlson & Johnston, 2015; Yoon e Schultz, 2017; Rice & Haywood, 2011; Príncipe & Silva, 2018). E é neste universo em constante mutação que viramos a nossa atenção para os profissionais de informação e a reestruturação dos seus papéis tradicionais. Aliando as suas competências seculares de curadoria a competências *core*, que estão habitualmente sob o domínio dos cientistas de computação, acreditamos, e os estudos comprovam-no, que estamos perante um elemento ativo e crítico que proporciona um conjunto de ferramentas e conhecimentos que facilitam e agilizam a investigação e o acesso ao financiamento da ciência (Amante, 2014; Akers et al., 2016). As bibliotecas adquirem competências interdisciplinares, fornecendo uma variedade de serviços de suporte de dados, incluindo instrução e treino (Piorun *et al.*, 2012; Surkis *et al.*, 2017; Shorish, 2015; Federer *et al.*, 2016); orientação para elaboração de PGD (Federer, 2016; Read *et al.*, 2017; Medina-Smith *et al.*, 2016); gestão e curadoria de dados (Leslie M. Delserone, 2008; Newton *et al.*, 2010; Lage *et al.*, 2011); e visualização de dados (Primich, 2010; Hunt, sem data). Entendendo os mundos da informação dos investigadores contemporâneos, bibliotecários de referência, gestores que organizam e implementam programas de serviço, compiladores bibliográficos, designers de sistemas de informação e colaboradores podem ser capazes de construir ambientes de informação que sejam mais favoráveis à pesquisa interdisciplinar (Palmer, 1996).

Os estudos disponíveis sobre as necessidades e usos da informação sempre se focaram mais sobre a forma como são feitas as pesquisas, do que com o uso que fazem do seu conteúdo (Paisley, 1968; Case, 2003). No passado a literatura sobre o tema revelava que a preocupação incidia sobre as publicações e os dados criados, em detrimento das atividades de curadoria subsequentes (Latour & Woolgar, 1986; Latour, 1987; Lynch & Woolgar, 1990). Poucas pesquisas nos estudos sociais da ciência sobre origens de dados, foram aplicadas a problemas de gestão e curadoria de dados (Wallis

et al., 2012). Atualmente as competências *core* dos profissionais de informação são amplamente discutidas, com vários organismos ligados à Ciência da Informação a fornecerem ideias, propostas, recomendações (Federer, 2018; Tenopir et al., 2013; Yoon & Schultz, 2017; Nitecki & Davis, 2017; Kenan, 2016; Cox et al., 2012; Semeler et al., 2019).

O objetivo geral deste trabalho é identificar o perfil necessário ao profissional da informação para se afirmar no contexto do 4º paradigma e da Ciência Aberta. Para o cumprimento deste objetivo identificaram-se, como objetivos específicos, o mapeamento da literatura internacional quer sobre esta matéria, quer sobre a identificação das ações levadas a cabo nas bibliotecas e dos atores que têm desenvolvido ações de literacia de dados.

A fundamentação teórica deste estudo, assentou numa revisão narrativa da literatura científica, a trave-mestra que sustentou a prossecução deste estudo. Foi aplicada uma metodologia de natureza qualitativa, com base num estudo exploratório. A pesquisa e recolha de informação realizou-se nas bases de dados subscritas pela Universidade de Coimbra, nomeadamente:

Web of Science, Scopus e Library and Information Science Source da EBSCO – selecionadas pelas apresentação intuitiva e completa. Incluem, entre outros fatores preferenciais, a possibilidade de restrição da pesquisa pela aplicação de filtros e operadores boleanos, acesso a dados bibliométricos, revisão por pares, uma cobertura globalizada e ainda possibilitou acionar alertas de publicação para os temas e/ou autores pretendidos.

Foram, além disso, usados outros recursos como o Repositório Científico de Acesso Aberto de Portugal (RCAAP); *eScholarship Publishing* – plataforma de publicação de acesso aberto subsidiada pela Universidade da Califórnia, gerida pela Biblioteca Digital da Califórnia; e pela relevância e desempenho na área, reconhecidos internacionalmente, foram visualizadas em formato vídeo, on-line, palestras de Christine L. Borgman disponíveis na sua página pessoal:

http://christineborgman.info/presentations/.

Na pesquisa optou-se pela pesquisa avançada com os termos “fourth paradigm” ; “data intensive science” ; “eScience” ; “data deluge” ; “big data” ; “data mining” ;

“long tail” ; “data sharing” ; “FAIR principles” ; “research data” AND “librarian” ; “data management plan” ; “DMP” ; “research data literacy” ; “information literacy” ; “core competencies” AND “information science” ; “core competencies” AND “academic library”; “core competencies” AND “librarians”; “librarians” AND “research data” ; “library roles” ; “roles for science librarians” ; “metadata”. Em todos os termos e expressões foram aplicados os filtros “revistos pelos pares”, “área de ciência da informação”.

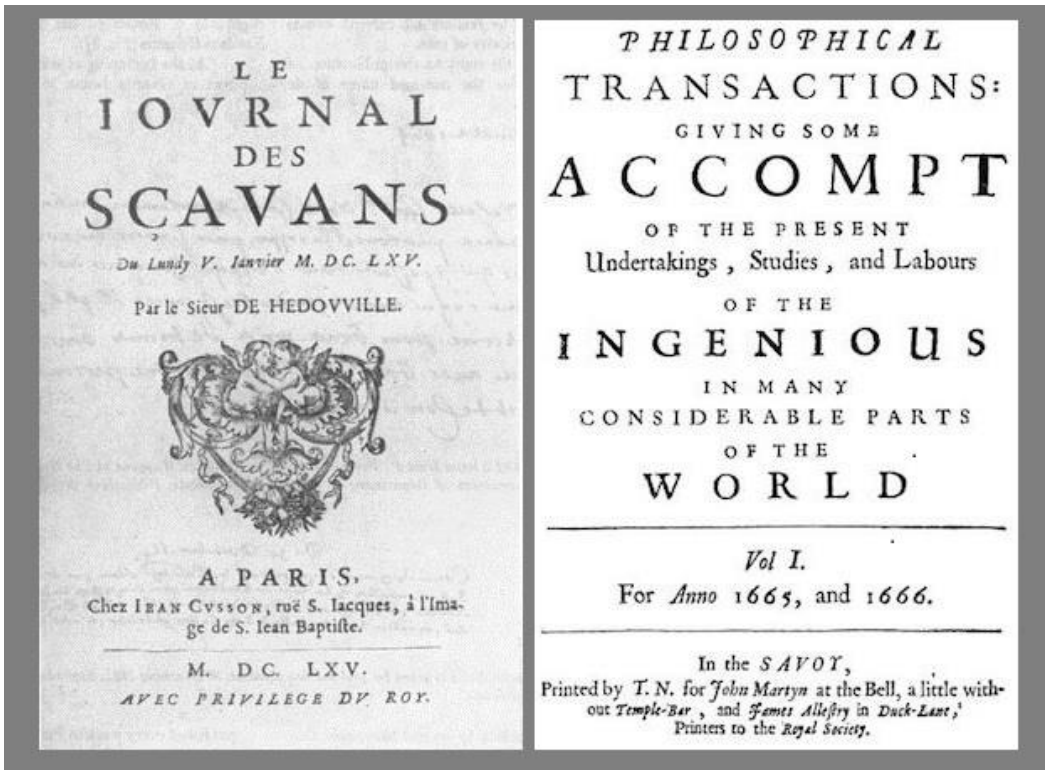
A pesquisa foi feita em inglês, à exceção do RCAAP. A chave de ordenação dos resultados foi o número de citações. A pesquisa foi realizada no período compreendido entre março e junho de 2022.

A estrutura do trabalho divide-se em três capítulos. No 1º capítulo é abordado o 4º paradigma da ciência como consequência do dilúvio de dados, exploramos o conceito de *Big Data* e uma das ferramentas possíveis para lidar com o dilúvio de dados, a mineração de dados, terminando com a problemática da cauda longa. O 2º capítulo foca-se na especificidade dos dados de investigação, fazendo uma revisão dos conceitos apresentados na literatura e explorando as propostas para uma taxonomia para a classificação de tipos de dados de investigação. Após discutir a noção de dados, reflete-se sobre as práticas do uso, reutilização e partilha dos dados, numa análise aos seus benefícios e, ainda, impeditivos. São, ainda, focados os princípios FAIR, associando-os aos planos de gestão de dados que são abordados no ponto seguinte (os seus requisitos, relação com agências de financiamento, impacto direto no processo de investigação científica e os *stakeholders* presentes nesta equação). Termina este capítulo com a discussão sobre a importância dos metadados, já com de um ponto de vista da curadoria. No 3º e último capítulo abordam-se as competências *core* necessárias na reinvenção dos papéis dos profissionais de informação. Exploramos algumas iniciativas em literacia dos dados de investigação, no panorama internacional e nacional, com base na literatura, recolhemos propostas de competências *core*, apresentamos algumas práticas em curso em algumas instituições e tentamos estabelecer um cenário possível do potencial reservado ao futuro desta profissão reinventada.

1 O 4º PARADIGMA DA CIÊNCIA, "DATA-INTENSIVE SCIENCE"

As primeiras revistas científicas foram publicadas em 1665. A primeira, "Journal des Sçavans", em 5 de janeiro de 1665, e dois meses depois a "Philosopical Transactions of the Royal Society" em 6 de março. Foram as precursoras das dezenas de milhares de revistas científicas que existem hoje.

Em algumas áreas da ciência, as revistas científicas são meio de comunicação dos resultados da investigação científica. Acompanham e registam o avanço da ciência, e constituem uma memória dinâmica do avanço do conhecimento científico.



LE
IOVRNAL
DES
SÇAVANS
Du Lundy V. Ianvier M. DC. LXV.
Par le Sieur DE HEDOVVILLE.

A PARIS,
Chez IEAN CVSSON, ruë S. Iacques, à l'Image de S. Iean Baptiſte.
M. DC. LXV.
AVEC PRIVILEGE DV ROY.

PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD

Vol I.
For Anno 1665, and 1666.

In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane, Printers to the Royal Society.

Figura 1 "Primeiras páginas das edições inaugurais dos dois primeiros periódicos científicos do mundo, ambos inaugurados em 1665, o francês "Jornal des Sçavans" (6.jan) e o britânico "Philosophical Transactions of the Royal Society" (6.mar). Imagem Maurício Tuff *in* As primeiras revistas científicas

Os cientistas, à época, já percebiam não terem forma de aceder a tudo e ganhavam consciência da importância de aceder às investigações dos seus pares. Longe do atual

cenário de dilúvio de dados ou *Big Data*, revelava-se já o primeiro impulso no sentido da comunicação da ciência, tocando já em alguns dos seus princípios basilares de partilha. O progresso científico depende da sua disseminação.

Alvin Weinberg, diretor de longa data do "Oak Ridge National Laboratory", consultor do governo e ensaísta, cunhou o termo "*Big Science* " pela primeira vez num artigo intitulado "Impact of Large-Scale Science on the United States", na revista Science, em 1961 (Weinberg, 1961). O termo "*Big Science",* inicialmente, serviu para enquadrar a sua análise da ciência financiada pelo governo e orientada nacionalmente, decorrente da II guerra Mundial e da sua nova economia política da ciência, em que a guerra de alta tecnologia transformava o apoio à pesquisa científica numa prioridade de segurança nacional e prometia transformar cientistas e engenheiros em beneficiários da generosidade da Guerra Fria.

Graças ao contributo de Wiemer houve uma reformulação das estratégias de política científica, na política aberta e no papel dos cientistas. Significativamente, a exploração dessas ideias por Weinberg continuou a evocar discussões e moldou o cânone dos estudos de ciência e tecnologia e os programas de história da tecnologia, ao longo de meio século desde que foram cunhados (Johnston, 2018). A "*Big Science* " é caracterizada por equipamentos caros que devem ser compartilhados entre muitos colaboradores, como aceleradores de partículas ou estações espaciais. A *eScience* e a ciberinfraestrutura são "*Big Science"* nesse sentido, pois são grandes investimentos sociais.

A revolução introduzida pela nova teoria dos autores pós-popperianos Kuhn (1970) e Lakatos & Musgrave (1980) de um novo paradigma na ciência teve grande influência na nova forma de olhar a ciência. À semelhança de obras como "A Física" de Aristóteles, "O Almagesto" de Ptolomeu, os "Principia" e a "Óptica "de Newton, a "Eletricidade" de Franklin, a "Química" de Lavoisier e a "Geologia" de Lyell – essas e muitas outras obras que serviram para definir os problemas e métodos legítimos de um campo de pesquisa para sucessivas gerações de profissionais, Kunh cunha o termo "paradigma" para uma nova visão da ciência. Considerou que os sucessos das obras mencionadas na evolução científica ocorreram por compartilharem duas características essenciais:

> Their achievement was sufficiently unprecedented to attract an enduring group of adherents away from competing modes of scientific activity. Simultaneously, it was sufficiently open-ended to leave all sorts of problems for the redefined group of practitioners to resolve. (Kuhn, 1970)

Jim Gray não ficou alheio a este conceito de paradigma, absorvendo também ele esta noção de continuidade:

> When paradigms change, the world itself changes with them. Led by a new paradigm, scientists adopt new instruments and look in new places. Even more important, during revolutions scientists see new and different things when looking with familiar instruments in places they have looked before. It is rather as if the professional community had been suddenly transported to another planet where familiar objects are seen in a different light and are joined by unfamiliar ones as well. (Kuhn, 1970)

Sobre as anteriores visões da ciência, mas sob novos olhares e perspetivas Gray descreveu o que cunhou como o "4º paradigma da ciência": uma ciência intensiva em dados, cooperativo, em rede e orientada para os dados. No entanto, os sistemas de computação em cluster têm conexão limitada a discos e não possuem software de banco de dados. Jim Gray foi um dos primeiros a antecipar essa necessidade. Em 1995, ele defendeu a construção de clusters de "storage bricks", consistindo em sistemas baratos e equilibrados de unidades centrais de processamento, memória e armazenamento para pesquisa intensiva de dados (Bell *et al.*, 2009)

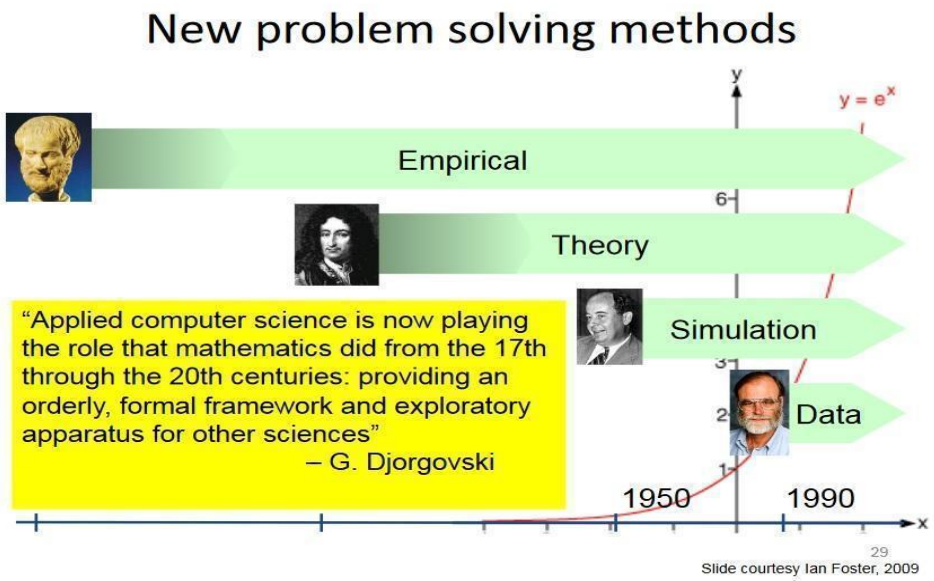


Figura 2 Slide courtesy Ian Foster *in* (Borgman, 2010)

> I wanted to point out that almost everything about science is changing because of the impact of information technology. Experimental, theoretical, and computational science are all being affected by the data deluge, and a fourth, "data-intensive" science paradigm is emerging. The goal is to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other. Lots of new tools are needed to make this happen. (Gray, 2007)

"O mundo da ciência mudou", começa por dizer Jim Gray na sua apresentação, em 2007[2]. O novo modelo sugere que os dados sejam capturados por instrumentos ou gerados por simulações antes de serem processados por software, e as informações ou conhecimentos resultantes sejam armazenados em computadores. Há uma nova expressão da ciência por novos canais.

Gray delineou um apelo em duas partes para o financiamento de ferramentas para captura, curadoria e análise de dados e para uma infraestrutura de comunicação e publicação. Defendeu o estabelecimento de armazenamentos modernos de dados e

[2] Apresentação feita por Jim Gray à NRC-CSTB in Mountain View, CA, em 11 de janeiro, 2007, http://research.microsoft.com/en-us/um/people/gray/JimGrayTalks.htm.

documentos que estão ao mesmo nível das bibliotecas tradicionais. A versão editada da palestra de Jim (Gray, 2007), produzida a partir da transcrição e dos slides de Jim, demonstra que a ciência com uso intensivo de dados consiste em três atividades básicas: captura, curadoria e análise. Os dados vêm em todas as escalas e formas, abrangendo grandes experimentações internacionais; observações interlaboratoriais, de laboratório único e individuais; e potencialmente a vida dos indivíduos (Hansen *et al.*, 2009). Jim Gray definiu *e-Science* como a síntese da tecnologia da informação e ciência, que possibilita enfrentar desafios a escalas anteriormente inimagináveis. Concebia a "ciência intensiva em dados" ou *e-Science* como um "quarto paradigma" da ciência (empírica, teórica, computacional e agora baseada em dados).

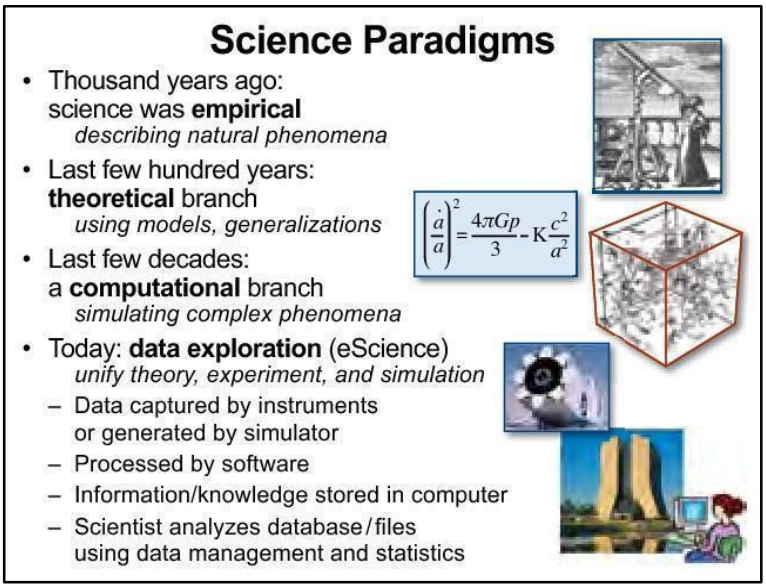


Figura 3 Diapositivo apresentado na exposição feita por Jim Gray à NRC-CSTB1 na Mountain View, CA, em 11 de janeiro, 2007 in (Gray, 2007)

A ciência empírica ou experimental caracteriza o primeiro paradigma, que cedeu lugar ao segundo paradigma, as generalizações teórico-conceituais resultantes do uso de modelos abstratos. O terceiro paradigma é baseado nas simulações assistidas por computadores ou outros equipamentos tecnológicos, enquanto o quarto paradigma é resultante da exploração de dados capturados ou gerados pela simulação (Bell *et al.*, 2009).

O progresso científico implica a existência de uma infraestrutura de informação comum que permita aos cientistas do domínio, explorar os dados disponíveis de forma eficaz e eficiente. Entre os benefícios potenciais da *e-Science* estão: (i) novos métodos de análise de dados e algoritmos mais inteligentes para lidar com a quantidade cada vez maior de dados; (ii) centros de ciência que permitem a computação no lado do servidor de dados, ao mesmo tempo que suportam acesso, intercâmbio e integração de dados; (iii) metadados sofisticados para acesso a dados que suportem independência física e lógica; e (iv) convergência semântica de ferramentas de dados, cruzando fronteiras disciplinares e epistemológicas (Gray *et al.*, 2005). Os cientistas precisam de uma "camada de ferramentas" para apoiar o ciclo de vida da informação desde o projeto inicial da pesquisa por meio de instrumentação, captura de dados, gestão de dados, análise, publicação e curadoria (Borgman *et al.*, 2007).

O lado apelativo de pesquisas intensivas de dados seduz tanto o investigador individual como a comunidade académica num sentido mais lato. A digitalização em larga escala da cultura material e o acesso a livros on-line, periódicos académicos e grandes conjuntos de dados abrem novas oportunidades para bolsas e investigações em qualquer disciplina acadêmica individual e interdisciplinas emergentes, onde a colaboração é a chave para o sucesso. (Griffin, 2013). As simulações de fenômenos muito grandes ou pequenos, rápidos ou lentos ou muito complexos para serem explorados em um laboratório de pesquisa são hoje possíveis graças à eScience produzindo grandes conjuntos de dados. As aplicações de eScience avançam em sintonia com a velocidade computacional (atualmente escala petaflop[3]; em breve escala exaflop e computação em grade[4]), (Griffin, 2013).

[3]In computers, FLOPS are floating-point operations per second. Floating-point is, according to IBM, "a method of encoding real numbers within the limits of finite precision available on computers." *in* What is FLOPS (floating-point operations per second)?

[4]A computação em grade é um grupo de computadores em rede que trabalham em conjunto, como um supercomputador virtual, para executar tarefas grandes, como analisar grandes conjuntos de dados e modelagem do clima. *In* O que é Computação em Grade

A *eScience* tornou-se a norma das ciências, com um progresso mais lento nas ciências não naturais, dadas as diferenças nas culturas epistémicas. Para tal houve um desenvolvimento da infraestrutura técnica, social e política para o conhecimento digital sob as rubricas de ciberinfraestrutura, o termo usado nos EUA, e eScience, o termo mais amplamente usado no Reino Unido e em outros lugares ([U.K. Programa de e-Ciência do Conselho de Pesquisa 2009; Atkins *et al.* 2003 *apud* Borgman, 2010).

O termo *e-Science*, atribuído a John Taylor[5], é definido como: "[...] global collaboration in key areas of science and the next generation of infrastructure that will enable it" (Hey & Trefethen, 2002). Ciberinfraestrutura e *eScience* – ambas cunhadas inicialmente em referência às ciências e tecnologia, e agora usadas de forma mais ampla – referem-se a uma infraestrutura que permite formas de conhecimento que são intensivas em informações e dados, distribuídas, colaborativas e multidisciplinares. *eResearch* tornou-se o termo coletivo para variantes como *eScience*, *eSocial Science* e *eHumanities* (Borgman, 2007).

Estabelecido que as tecnologias de dados tradicionais não comportam a dimensão e heterogeneidade dos dados no mundo moderno, a faceta digital ou eletrónica da ciência, ou eScience, apresenta-se como essencial. Claramente, a ciência intensiva de dados, um componente da eScience, deve ir além dos *data warehouses* e sistemas fechados, permitindo o acesso aos dados para os que estão fora das principais equipas do projeto, uma maior integração de fontes e fornecendo interfaces para os cientistas especializados que não são especialistas em administração de dados e computação, e aos demais elementos de suporte ao processo de investigação, onde se destacam os profissionais da informação. Projetos como o Large Hadron Collider (LHC)[6] e o Australian Square Kilometer Array Pathfinder (ASKAP)[7] geram petabytes de dados, que devem ser analisados por centenas de cientistas que trabalham em vários países e falam muitos idiomas diferentes (Fox & Hendler, 2009).

[5] Director General of Research Councils in the UK Office of Science and Technology (OST)

[6] The Large Hadron Collider | CERN

[7] ASKAP Home

Na sequência do trabalho de Gray, foi publicado o livro "O Quarto Paradigma", um volume de pequenos ensaios escritos por académicos e especialistas, organizado em quatro grandes áreas científicas: Terra e Meio Ambiente, Saúde e Bem-Estar, Infraestrutura Científica e Comunicação Académica. Usando exemplos dentro dessas categorias, a eScience é descrita pelo seu impacto em diferentes aspetos da gestão de dados.

Na visão de um cientista de investigação (Regan, 2012), apesar da importância do armazenamento dos dados, o aspeto mais pertinente do ponto de vista do cientista, é a capacidade de redução do volume de dados de forma a ser gerenciável e preciso. Entre algumas soluções propostas está o uso da Inteligência Artificial (AI) e práticas de dados padrão. Mas Reagan levanta a questão: se os computadores começam a superar a capacidade humana de compreender a ciência e os métodos computacionais intensos são necessários para analisar dados, torna-se desanimador para gerações de cientistas que encontram beleza em entender as complexidades da natureza, usando apenas o poder da mente humana. Uma preocupação já referida por outros autores:

> I am afraid that the greatest danger facing information science is losing the sight of users, of human beings. [...] But I am also convinced that the greatest payoff for information science will come if and when it successfully integrates systems and user research and applications. **Society needs such a science and such a profession**. (Saracevic, 1999, grifo nosso)

O elemento humano deve estar presente, como sublinha Borgman *et al.* (2015), ao considerar o conhecimento como "redes robustas de pessoas, artefatos e instituições". Reagan (2012) não diminui, no entanto, o valor do Quarto Paradigma, afirmando que "certainly be considered to provide a sobering and vivid vision of how science must cope with the emergence of big data".

Figura 4 Gerada pelos autores em WordClouds.com

1.1 DILÚVIO DE DADOS/ BIG DATA

> Long predicted by the science community (Hey, A. J. G. & Trefethen, 2003), the popular press is now heralding the wide availability of data for use by anyone, anywhere. Not only have Nature and Science, the premier science journals, published feature sections on "big data" (Community cleverness required, 2008; Data's shameful neglect, 2009; Dealing with data, 2011), so have WIRED magazine (Anderson, 2008), and the Economist (Data, Data Everywhere, 2010). Universities are assessing their rights, roles, and responsibilities for managing and for exploiting data from their researchers (The University's Role in the Dissemination of Research and Scholarship, 2009; Lyon, 2007). (Borgman, 2012)

O dilúvio de dados já se afigurava como um problema em crescimento após a II Guerra Mundial, tendo o avanço tecnológico, mais especificamente a computação, contribuído muito para o emergir das preocupações. Vários autores deram o seu contributo na formulação de respostas ao problema. Havia a necessidade de automatizar o processo de descoberta - dos dados à informação e ao conhecimento - na medida do possível, com automação da gestão de dados, armazenamento e organização de entidades digitais e avanços para o gestão automático de informações. Para tal, seria necessária a anotação automática de dados científicos, com metadados que descrevessem características

interessantes dos dados e do armazenamento e organização das informações resultantes da gestão automatizada do conhecimento dos dados científicos:

> [...]it is evident that *e-Science* data generated from sensors, satellites, high-performance computer simulations, high-throughput devices, scientific images and so on, will soon dwarf all of the scientific data collected in the whole history of scientific exploration (Hey & Trefethen , 2003).

O dilúvio de dados tem impactos diferentes nos diferentes contextos: nas áreas científicas que pertencem à "*Big Science* ", como a física e a astronomia, constroem ferramentas e repositórios para fazer face ao dilúvio, mas na "little science", dependentes de trabalho de campo, carecem de ferramentas e infraestruturas para gerir a crescente quantidade de dados derivados das novas formas de instrumentação: "A few gigabytes of data daily might be a trickle to a high-energy physicist, but waterfall to a habitat ecologist" (Borgman *et al.*, 2007). Atualmente a realidade virtual ou dados no ciberespaço, que fazem parte intrínseca da vida humana e se formam e desenvolvem de forma inconsciente, acrescentam uma maior quantidade de dados à realidade do nosso tempo (Zhu, Zhong, & Xiong, 2009; Zhu & Xiong, 2009). Nesta perspetiva, alguns autores admitem que os dados estão a reinventar-se, não estando limitados a valores de variáveis qualitativas ou quantitativas, ou resultados de medições, ou dados científicos gerados no contexto de observações e experiências científicas. Além de tudo isso, dados também são tudo o que se encontra no ciberespaço. Há cada vez mais dados que não têm referências no mundo natural, como vírus de computador, jogos online e dados inúteis, todos gerados na natureza de dados (Zhu & Xiong, 2015).

Há uma preocupação crescente dos avanços nos programas científicos não serem capazes de acompanhar o aumento das taxas de dados, provenientes dos avanços em dispositivos móveis, sensores digitais, comunicações, computação e armazenamento que agilizaram os meios para recolher dados (Bryant *et al.*, 2008) devido à falta de recursos, à necessidade de pesquisa e desenvolvimento de ferramentas, bem como às plataformas e

infraestruturas necessárias para gerir, analisar e atuar nas crescentes coleções de dados (Bethel *et al.*, 2016).

> There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days, and the pace is increasing. (Google CEO, Eric Schmidt *in* Oracle)

De acordo com a renomada empresa de IT Industrial Development Corporation (IDC; 2011), a quantidade total de dados no mundo aumentou nove vezes em cinco anos (Gantz e Reinsel, 2011). Espera-se que esse número dobre pelo menos a cada dois anos (Chen, Mao & Liu, 2014).

A determinação das origens do termo *Big Data* ondula entre conversas à mesa de almoço na Silicon Graphics Inc. (SGI) em meados da década de 90 e entre a sua difusão, nas iniciativas promocionais da IBM em 2011 e de outras empresas de tecnologia líderes, que investiram na construção do nicho de mercado de análise (Gandomi & Haider, 2015).

Laney propôs **V**olume, **V**ariedade e **V**elocidade (ou os **3V's**) como as três dimensões dos desafios na gestão de dados. Os 3V's surgiram como uma estrutura comum para descrever *Big Data* (Chen *et al.*, 2012, Kwon *et al.*, 2014; Laney, 2001).

Na literatura encontramos várias definições de *Big Data*: é a quantidade de dados além da capacidade da tecnologia de armazenar, gerir e processar de forma eficiente (Manyika et.al, 2011); são ativos de informação de alto volume, alta velocidade e/ou alta variedade que exigem novas formas de processamento para permitir uma tomada de decisão aprimorada, descoberta de insights e otimização de processos (Gartner IT Glossary, s.d.; Gürsakal, 2014); são tecnologias e arquiteturas de nova geração que foram projetadas para extrair valor de conjuntos de dados multivariados de alto volume de forma eficiente, fornecendo captura, descoberta e análise de alta velocidade (Gantz e Reinsel, 2011); é o conjunto de métodos e tecnologias em que novas formas são integradas para desdobrar valores ocultos em conjuntos de dados diversos, complexos e de alto volume (Hashem et.al., 2015); são dados grandes em tamanho/ volume, grandes em variedade (estruturado; semi estruturado; não estruturado) e grande em velocidade de mudança (Elragal, 2014); é o ativo

de informação caracterizado por um Volume, Velocidade e Variedade tão elevados, que requerem tecnologia e métodos analíticos específicos para a sua transformação em Valor (De Mauro *et al.*, 2016). Mas na literatura encontramos o acréscimo de mais dois fatores: **V**eracidade e **V**alor (Hashem *et al.*, 2014; Elragal, 2014; Fadiya *et al.*, 2014; Yang *et al.*, 2014; Panneerselvam *et al.*, 2015; Yaqoob *et al.*, 2016) passando os 3V's a **5V's** (i) **Volume**: É a característica mais importante do Big Data. Representa o tamanho do conjunto de Big Data. (ii) **Variedade**: Vários dados surgem oriundos de vários recursos (internos ou externos). Essas entradas de dados de recursos separados causam variação no conjunto de dados. Os dados externos raramente são estruturais. (iii) **Velocidade**: A taxa de produção de Big Data é notavelmente alta. O grande aumento de dados significa que os dados devem ser analisados mais rapidamente. Quanto mais rápido os dados aumentam, mais rápido a necessidade de dados aumenta; (iv) **Veracidade**: É a exatidão dos dados. Os dados devem ser adquiridos de recursos fiáveis e sua segurança deve ser fornecida. Apenas pessoas autorizadas devem ter permissão de acesso. (v) **Valor**: Um resultado deve ser gerado após todos os procedimentos e o resultado deve enriquecer o processo.

1.2. MINERAÇÃO DE DADOS: UMA DAS FERRAMENTAS PARA LIDAR COM O FENÓMENO *BIG DATA*

A Ciência da Informação, que investiga as propriedades da informação e os métodos e técnicas usados na aquisição, análise, organização, disseminação e uso da informação (Borko, 1968), foi bastante afetada pela mineração de dados. É encarada como uma das soluções viáveis para lidar com o dilúvio de dados, uma busca de informações valiosas em grandes volumes de dados. Tem representado, por isso, uma contribuição significativa para o campo da Ciência da Informação (Chen & Liu, 2004).

> This explosively growing, widely available, and gigantic body of data makes our time truly the data age. Powerful and versatile tools are badly needed to automatically uncover valuable information from the tremendous amounts of data and to transform such data into organized knowledge. This necessity has led to the birth of data mining. The field is young, dynamic, and promising. Data mining has and will continue to make great strides in our journey from the data age toward the coming information age. (Han *et al.*, 2012)

A mineração de dados também denominada “descoberta do conhecimento a partir dos dados” (knowledge discovery from data - KDD) é definida como:

> a extração automatizada ou conveniente de padrões, que representam conhecimento, implicitamente armazenado ou capturado em grandes bancos de dados, armazéns de dados, Web, outros repositórios de informações massivos ou fluxos de dados. Como um campo multidisciplinar, a mineração de dados baseia-se no trabalho de áreas como estatística, aprendizagem da máquina, reconhecimento de padrões, tecnologia de banco de dados, recuperação de informações, ciência de rede, sistemas baseados em conhecimento, inteligência artificial, computação de alto desempenho e visualização de dados (Han *et al.*, 2012, tradução nossa).

Apesar de ser utilizada em Ciência da Computação desde 1960, só a partir da década de 1980 o termo se consolidou em outras áreas, sendo uma técnica muito utilizada na ciência de dados. O objetivo final recai sempre na classificação, ou seja, na busca dentro do universo pesquisado de um padrão que permita unir (ou segregar) os entes, objetos ou fenómenos analisados. Pode dizer-se que no final de cada processo de mineração, emerge pelo menos um novo dado, um novo atributo, para cada um dos entes do conjunto (Gomes *et al.*, 2019). Estudos revelam como a mineração de dados se configura como uma técnica essencial na busca de dados “escondidos” no processo de empréstimo nas bibliotecas digitais (Yi *et al.*, 2018).

1.3. LONG TAIL

Long Tail é um termo que deriva da área da Economia, popularizado por Anderson (2004) no contexto do comércio na Internet. Algumas das propriedades e ferramentas da informação na economia da internet, aplicam-se também à atividade científica, na classificação por tamanho dos projetos e escala de produção de dados (Heidorn *et al.*, 2015; Wallis *et al.*, 2013). *Long Tail* refere-se a conteúdos que não estão no *mainstream*, nichos de produtos, que adquirem valor ao estarem acessíveis Este fenómeno é replicado na atividade científica, ou seja, os dados que são muitas vezes desconsiderados, mal indexados

e armazenados tornam-se quase invisíveis. A maioria do trabalho científico é realizado em projetos relativamente pequenos, com parco financiamento e com equipas de pequena dimensão. Os dados em bruto destes trabalhos de investigação, são dados de investigação científica, base de toda a teoria científica. No entanto, pela sua pequena dimensão e visibilidade relativa, não são tratados da mesma forma que os dados provenientes de projetos de investigação de grande exposição (Heidorn, 2008).

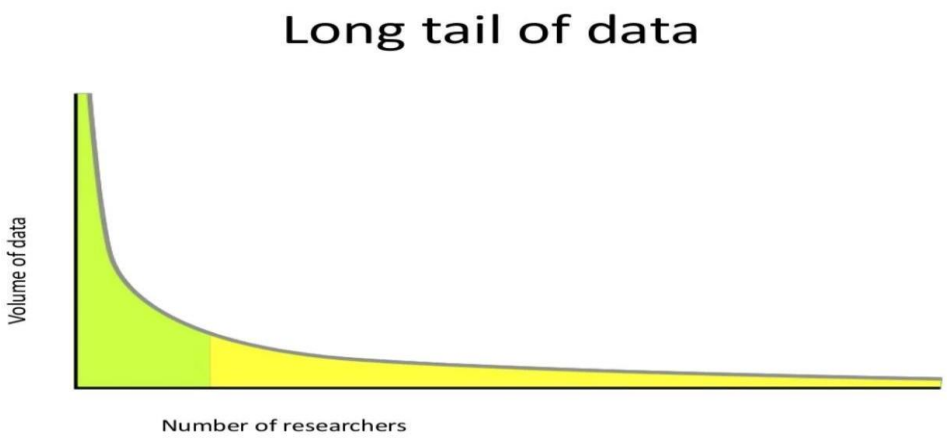


Figura 5 Long Tail of Data in Institute for Empowering Long Tail Research: A study funded by the NSF in https://sites.google.com/site/ieltrconcept/home

O programa de Infraestrutura de software para inovação sustentada, elaborada pela National Science Foundation, concedeu, entre 2012 e 2014, um financiamento[8] ao grupo de investigação``Institute for Empowering *Long Tail* Research" (IELTR)[9], com Foster, Howe, Heidorn e Christine Borgman da Universidade da Califórnia, como investigadores principais, em que os trabalhos foram conduzidos por investigadores em ciência da informação e em ciência da computação (Mitchum, 2012). O estudo apoiou-se na premissa de que projetos científicos maiores e de longo prazo têm muito mais recursos para dedicar à gestão de dados, do que cientistas que trabalham em projetos menores e de curto prazo,

[8] https://nsf.gov/awardsearch/showAward?AWD_ID=1216754

[9] https://sites.google.com/site/ieltrconcept/home

visando fornecer soluções para o problema (Borgman *et al.*, 2016). Usamos este exemplo como demonstração da consciencialização da problemática da cauda longa e a urgência de encontrar soluções. No caso em particular observamos o estudo de 'Data as a Service' (DaaS) que apresenta soluções sugerindo ferramentas como Figshare, Dropbox, Github, Omeka e Zotero. As versões de software são atualizadas automaticamente, a maioria dos aplicativos são independentes de uma plataforma, portanto, parceiros em diferentes sistemas operacionais (por exemplo, Apple, Microsoft, Linux) e podem compartilhar os mesmos aplicativos e recursos de dados. Valiosos recursos para financiamentos reduzidos. Os investimentos na prática científica, por agências de fomento, educadores e investigadores, devem ter uma visão de longo prazo de sustentação do acesso aos dados científicos (Borgman, 2015). «Para citar Jim Gray (National Research Council, 2004): 'que todos os seus problemas sejam técnicos'» (Borgman *et al.*, 2016). Se houver incentivos para os investigadores moverem os dados para computação, para a cloud no período de recolha dos dados *raw*, certamente os investigadores deixarão esses dados disponíveis (Heidorn *et al.*, 2015).

Alguns dos cientistas que trabalham sozinhos ou em pequenas equipas tornam-se especialistas em tecnologia, enquanto outros se recusam a fazer esses investimentos. Muitas vezes, pequenas equipas contam com estudantes de pós-graduação, bolseiros de doutoramento para gerir os seus dados. A experiência muda rapidamente à medida que esses funcionários de curto prazo se formam ou concluem os seus estudos (Borgman, Wallis e Enyedy, 2007; Eddy, 2005; Edwards *et al.*, 2011; Steinhardt e Jackson, 2014). Esta realidade pode e deve ser corrigida, incluindo na equipa de investigação um suporte humano com competências na gestão de dados, em todo o seu ciclo de vida. Os profissionais de informação têm o perfil adequado para o efeito, mantendo um corpo de suporte à investigação de continuidade e fiabilidade para futuras investigações, mantendo o conhecimento adquirido disponível para futuras investigações. Dados e software estão profundamente interligados na prática de pesquisa. Mantê-los, sozinhos ou em conjunto, e dar crédito por contribuições para dados de pesquisa e software, são preocupações crescentes em infraestruturas de conhecimento (Howison e Bullard, 2015; Howison e Herbsleb, 2011, 2013; Uhlir, 2012; Velden et al., 2014 apud Borgman et al., 2016).

2 DADOS DE INVESTIGAÇÃO: DEFINIÇÃO E TAXONOMIA

The most elusive term in data science is 'data'. (Borgman, 2019)

Etimologicamente "data" provém do latim *datu-*, particípio passado do verbo *dāre*, "dar", com significação de "dado", fazendo o plural em data (Porto Editora, 2021). No idioma inglês, o termo é consensual no uso do plural latino - *Data*. Mas o consenso repousa no significante. Já o significado reveste-se de muitas cambiantes. O seu uso e referencial, nas diferentes áreas do conhecimento, foi sempre muito debatido. A questão impõe-se: a que nos referimos quando falamos em dados?

Em algumas obras de referência encontramos definições que apontam "Data" como "facts and statistics collected together for reference or analysis […] quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media" (dicionário de inglês Oxford on-line); "[...] facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors [..] the smallest unit of information to which reference is made[...]"(National Research Council, 1999). Encontramos em vários autores os primeiros esboços da noção de dados, como Otlet e La Fontaine (Wellisch, H., 1972); Bradford contando e correlacionando "dados" de forma inovadora no universo das bibliotecas (East, H., 1983); Borko com uma visão de futuro" An information system is that combination of human and computer-based capital resources which results in the collection, storage, retrieval, communication and use of data" (Borko, 1970 citado por Wellisch, H. 1972). Outro marco importante foi a análise revolucionária de Shannon & Weaver (1949). Antes dele, os engenheiros não tinham meios de calcular a quantidade máxima de informação que poderia ser transmitida por um canal de um determinado tamanho ou configuração. Supunha-se que seria possível continuar a melhorar os canais, para veicular cada vez mais informação. As fórmulas de Shannon permitiam o cálculo da máxima transmissão de informação possível, para uma determinada configuração física (Bates & Maack, 2009).

Para um melhor entendimento da evolução do conceito de dados, é importante compreender a dinâmica da relação entre os conceitos de informação, conhecimento e dados[10]. Com divergências entre os principais autores da área, Buckland (1991) procurou trazer o aspeto da materialidade da Informação (*Information as Thing*), bem como da informação como processo e como fenômeno cognitivo. Capurro e Hjørland (2007) discutiram sobre o conceito de Informação e o seu caráter polissémico e Furner (2016) realizou uma extensa pesquisa, não só sobre a história e etimologia do termo dado, mas trouxe também os vários usos do termo em diferentes áreas do conhecimento (Gomes *et al.*, 2019).

Os dados são carregados de teorias e a sua observação é historicamente contingente (Banning, E. B., 2020). A noção de dados sobre um mesmo assunto difere, consoante o olhar e o objetivo de quem os recolhe, utiliza e reutiliza, depende muito do entendimento de como, onde e quando foram criados (Bechhofer *et al.* 2010; Burton & Jackson 2012; Gitelman 2013; Ribes & Jackson 2013; Vertesi & Dourish 2011; Edwards, 2013). São muitos os fatores que determinam as formas e os tipos de dados, incluindo o estágio da investigação, as características do domínio da pesquisa e o quanto se sabe sobre o problema de pesquisa. Price & Price (1986) faz a distinção entre *Big Science* e *Little Science,* com base em fatores como a maturidade do campo, a consistência dos métodos de pesquisa, o grau de instrumentação compartilhada e os volumes de dados produzidos. Esta distinção tem sido muito importante nos debates sobre dados e muito estudada, (Furner, 2003; Wallis *et al.*, 2012) abrindo novas perspetivas e possibilidades nos procedimentos a ter com os dados gerados.

Atualmente, a urgência na definição dos conceitos em Ciência da Informação, como é o de dados, tem vindo a relativizar-se em prol da urgência de encontrar soluções para o seu aumento exponencial. Em última instância, são uma construção humana, definidos segundo um propósito e uma interpretação, com o objetivo de constituir prova para o que se pretende (Buckland, 1999). Existem em infraestruturas de conhecimento, que governam

[10] Sobre o tema veja Zins, C., 2007; Bates, M. J. ,2005; Gundry 2001; entre outros.

como são criados, geridos, usados e interpretados (Edwards *et al.*, 2013) e servem como prova para os projetos de investigação ou de bolsas de estudo (Borgman, 2019).

Tudo pode ser considerado como dados (Glaser, 2007), apesar do alerta de alguns autores como Alfred Borgmann (1999), de que nem todo o tipo de informação (natural, cultural) é passível de ser reduzida à noção de dados que possam ser transmitidos, processados e/ou produzidos por computadores ou tecnologias afiliadas.

Se, por um lado, existe a tendência de "cientificar" o conceito, reduzindo dados a uma representação reinterpretável, dentro de padrões formalizados e adequados a uma comunicação, interpretação ou processamento (Reference Model for an Open Archival Information System [OAIS], 2012), por outro, estas descrições de técnica de dados, obscurecem o contexto social onde os dados existem (Borgman *et al.*, 2006).

> Observations that are research findings for one scientist may be background context to another. Data that are adequate evidence for one purpose (e.g., determining whether water quality is safe for surfing) are inadequate for others (e.g., government standards for testing drinking water). Similarly, data that are synthesized for one purpose may be 'raw' for another. (Borgman, 2007; Borgman *et al.*, 2006; Bowker, 2005).

Alguns autores consideram dados como uma representação da natureza, registando o comportamento humano, incluindo trabalho, meios de subsistência e desenvolvimento social que são rapidamente colocados no ciberespaço (Zhu & Xiong, 2015) ou quem considere que todos os dados devem ser entendidos não como data, mas como capta[11] e as convenções criadas para expressar modelos de conhecimento, independentes do observador, precisam de ser radicalmente reformuladas, para expressar a interpretação humanista, enraizada numa relação codependente, entre observador e experiência, para ser expressa de acordo com gráficos construídos, a partir de modelos interpretativos (Drucker, 2011). Numa

[11] Particípio do verbo capiō ("procurar, tirar") | captus (masc.) (fem. capta, neut. captum) | capturado; apreendido, levado, ter sido acolhido, compreendido in https://www.wordsense.eu/captus/#Latin

visão mais inclusiva, podemos considerar dados para ”[...]Além das manifestações digitais da literatura (incluindo texto, som, imagens estáticas, imagens em movimento, modelos, jogos ou simulações)”, mas também a formas de dados e bancos de dados, que geralmente requerem o auxílio de máquinas e softwares computacionais para serem úteis (exemplo de vários tipos de dados de laboratório, incluindo dados espectrográficos, sequenciamento genómico e microscopia eletrónica); dados observacionais (dados de sensoriamento remoto, geoespaciais e socioeconómicos); e outras formas de dados gerados ou compilados, por humanos ou máquinas (Uhlir & Cohen, 2011 *apud* Borgman *et al.*, 2012). De um outro ponto de vista, são o material factual registado, aceite na comunidade científica como necessário para validar os resultados de uma investigação. É importante que o bibliotecário seja capaz de determinar o que constituem os dados “primários” do pesquisador (dados que são gerados ou analisados especificamente para alcançar os resultados do projeto) *versus* os dados “auxiliares” (qualquer dado adicional que é trazido ou gerado para auxiliar para explicar ou entender os dados primários, mas não são usados diretamente para fins de pesquisa).

Primeiro, os investigadores geralmente identificam o seu trabalho com dados como uma série de estágios, pelos quais estes passam. Em segundo lugar, conceptualizar o desenvolvimento e o uso de dados pelo pesquisador como uma série de estágios dentro de um ciclo de vida, naturalmente suporta discussões sobre o processo, métodos e ferramentas usadas para trabalhar com os dados em cada estágio (Carlson, J., 2012). Uma vez que o conceito de dados de investigação é complexo, virtualmente, todo o tipo de informação digital tem potencial para ser dado de investigação se for usado como recurso primário da investigação (Rice and Southall, 2016).

Os dados digitais apresentam-se sob muitas formas. Estas distinções são feitas na base da sua natureza, da sua reprodutibilidade e do nível de processamento a que estiveram sujeitas. Quanto à sua natureza, esta pode incluir números, imagens, vídeo ou *audio streams*, software, algoritmos, equações, animações ou modelos/ animação. Podem ser diferenciados pela sua origem ou, segundo alguns autores, pelo tipo de investigação: observacional, computacional ou experimental: os dados observacionais (como por exemplo a observação direta da temperatura do oceano numa determinada data, a atitude dos eleitores antes de

uma votação, ou fotografias de uma supernova) são registos históricos que não podem ser recolhidos e são portanto arquivados indefinidamente; os dados computacionais são o resultado da execução de um modelo ou uma simulação computacional; no processo experimental há que fazer outra distinção: entre os dados intermédios (os preliminares da investigação) e os dados finais. Isto porque os investigadores só consideram os dados que consideram mais interessantes, sendo os dados intermediários inacessíveis a outros investigadores, (National Science Foundation, 2005).

Há uma outra perspetiva que classifica os tipos de dados quanto ao seu grau de processamento (em bruto ou processados), quanto à sua proveniência (primários ou secundários), quanto à sua dimensão (Big Data e *Long Tail* Data) e quanto ao tipo de investigação a que se associam (dados de observação, experimentais, de simulação, derivados ou compilados e de referência ou canónicos (estáticos ou orgânicos), (Príncipe *et al.*, 2021; Zimmerman, 2008).

Griffin (2013) propõe uma tipologia de dados em que considera como experimentais os dados capturados por instrumentos digitais de alto rendimento e dispositivos de gravação (instrumentos astronómicos sofisticados, aceleradores de partículas, sensores ambientais, equipamentos de diagnóstico médico e muitos outros). Os petabytes de dados que podem resultar daí, geralmente requerem computação significativa para produzir os dados básicos para análise, mas o processo geral é essencialmente automático. Um segundo tipo de dados experimentais é aquele que requer envolvimento humano em algum ponto da captura ou preparação de dados para análise. Exemplos recentes incluem análise de redes sociais.

No âmbito do tema a que nos propomos com o presente trabalho, seguimos a proposta de Christine Borgman (2010) para uma conceção de dados de investigação, que classifica em 4 categorias, de acordo com Long-Lived Data Collections 2005; observacionais, computacionais, experimentais e registos, podendo ser sendo objetivos ou factos e subjetivos ou “alegadas provas” citando Buckland (2006) :

> Entities used as evidence of phenomena for the purposes of research or scholarship (Borgman, 2015)

> Data may exist only in the eye of the beholder: the recognition that an observation, artifact, or record constitutes data is itself a scholarly act. [...] Data is a difficult concept to define, as data may take many forms, both physical and digital. [...] In addition to digital manifestations of literature (including text, sound, still images, moving images, models, games, or simulations), it refers as well to forms of data and databases that generally require the assistance of computational machinery and software in order to be useful, such as various types of laboratory data [...] socioeconomic data; and other forms of data either generated or compiled, by humans or machines. (Borgman, 2011)

O capital humano, a instrumentação e os dados são o capital académico que possibilita o incremento de investimentos em pesquisa, ao reproduzir e verificar novas descobertas, novas perguntas com dados existentes e uma pesquisa de colaboração com criação de dados, a sua partilha e reutilização.

A heterogeneidade dos dados e as questões que levanta demonstram que políticas generalistas não são adequadas. São essenciais políticas que reconheçam e apoiem vários tipos de dados, nas suas diversas condições (National Science Board, 2005).

Vivemos na era do acesso. O acesso ao conhecimento. O acesso aberto aos dados não é necessariamente equivalente aos dados abertos ou à ciência aberta. "Ciência aberta é mais do que a disponibilização em acesso aberto de dados e publicações, é a abertura do processo científico enquanto um todo, reforçando o conceito de responsabilidade social científica" (*Ciência Aberta*, sem data). O conceito de dados abertos pode ser de duas origens diferentes, mas interligadas: Dados abertos governamentais ou dados (que incluem publicações) gerados pelo financiamento governamental de atividades de pesquisa por meio de doações para projetos de pesquisa em universidades ou instituições de pesquisa (Jeffery *et al.*, 2014).

O Acesso Aberto (AA) foi facilitado e promovido pela criação de repositórios a nível nacional (caso do RCAAP, em Portugal, e do NORA, na Noruega, por exemplo), a nível europeu (OpenAire) e regional (La Referencia, América Latina, Caraíbas e Brasil), tornando os repositórios fundamentais na reforma do sistema de comunicação científica (Rodrigues *et al.*, 2004; Chan, 2004; Swan, 2006; Jantz e Wilson, 2008; Gomes, 2013 *apud* Freitas & Corujo, 2021).

2.1. USO, REUTILIZAÇÃO E PARTILHA: ANÁLISE, CONCEITOS E PROBLEMAS

> The ethos of science is communism, in the special sense that the institutional norms of science would make its products part of the public domain, shared by all and owned by none. (Merton, 1968)

Fienberg *et al.*, (1985) sublinha a expressão "Ethos of science" de Merton, (1968) como um ideal a atingir, expresso na disponibilidade das descobertas científicas em toda a comunidade científica. Exemplifica, citando Cavendish (1798), com uma prática prestigiada em tempos passados: entre os melhores investigadores e com uma prática de revistas científicas aberta a uma descrição extensiva, fornecer dados era uma tradição honrada. Hoje, a globalização na ciência é mais expressa, mais explícita, permeada pelos avanços tecnológicos e as novas técnicas de registo que estão a revolucionar a ciência. Percebemos que a Internet e a *World Web Wide* (WWW) ancoram a ciência no ciberespaço, proporcionando um acesso globalizado. Esta globalização, para além de agilizar a comunicação da ciência, incentiva a novas práticas pela simples observação dos resultados dos seus pares, em todos os *stakeholders* do processo de conhecimento. Com as tecnologias de informação e comunicação (TIC) o conceito de "ciência aberta" condiciona positivamente as diretrizes orientadoras para o uso, partilha e reutilização de dados. O conceito de ciência aberta, como sublinham Veiga, Silva & Borges (2021):

> [...]abrange vários termos e práticas como dados abertos, publicação ampliada, dados linkados, revisão por pares aberta, avaliação da ciência aberta (impacto e métricas abertas), recursos abertos (incluindo recursos educacionais abertos), software aberto, acesso aberto ao conhecimento e outros. Esses termos e práticas mostram uma nova lógica não apenas de disponibilização, mas de produção e organização do conhecimento. (Veiga *et al.*, 2021)

A velocidade da produção de dados é superior à capacidade da sua captura, gestão e armazenamento, pré-requisitos para a sua curadoria, ou seja, agregar valor ao seu conteúdo. As opções de dados, metadados, ferramentas analíticas e especializações são constantemente reanalisadas. Muitos, se não a maioria, dos campos científicos, estão a tornar-se mais intensivos em dados com os avanços na instrumentação, como por exemplo

as redes de sensores. À medida que novos instrumentos e formas de dados se tornam disponíveis e as comunidades respondem a novos requisitos para planos de gestão de dados, as práticas científicas estão em fluxo. O fluxo cria oportunidades para estudar a produção, uso e reutilização de dados. Há que questionar como são usados os dados e como são reutilizados (Wallis *et al.*, 2012).

Goodman *et al.*, (2014) estabeleceram um conjunto de 10 regras gerais, linhas orientadoras que permitam ter sempre em vista a importância dos dados, durante todo o seu ciclo no processo de investigação: valorizar os seus dados; partilhá-los *online* com identificador permanente; ter sempre em mente a sua reutilização; publicar o fluxo de trabalho com contexto; associar os dados às suas publicações sempre que possível; publicar o seu código (mesmo os pequenos bits); especificar como pretende ter o seu trabalho reconhecido/ obter crédito; promover e usar repositórios de dados; recompensar os colegas que compartilharem os seus dados corretamente; ser um impulsionador da ciência de dados.

Os laboratórios de pequeno e médio porte, associados à cauda longa, geralmente encontram "atritos científicos" (Edwards *et al.*, 2011) que impedem a gestão, descrição e armazenamento de dados. Esses atritos também retardam o progresso científico e impedem o fluxo de informações dentro e entre as equipas de pesquisa. Fricções comuns incluem falta de padronização, metadados incompletos ou incompatíveis, reivindicações de propriedade intelectual, práticas de partilha de dados e restrições de recursos humanos (Edwards, 2010; Edwards *et al.*, 2011; Mayernik *et al.*, 2011). O atrito de metadados, em particular, é um exemplo de como a falha em documentar conjuntos de dados inibe o uso e a reutilização de dados (Mayernik *et al.*, 2011). Sem metadados, os conjuntos de dados podem transformar-se em planilhas de linhas e colunas não rotuladas ou em sequências de números indecifráveis. Os esquemas de metadados facilitam a descoberta e a partilha de dados em escala global. Estes são conjuntos de elementos – por exemplo, título, criador, data de publicação – projetados para atender às necessidades de uma determinada comunidade. O uso de esquemas de metadados estruturados e extensíveis é uma solução proposta para acesso, descoberta e partilha de dados (Edwards *et al.*, 2011; Getty Research Institute, 2008). No entanto, a criação manual de metadados é uma grande barreira para a gestão eficiente de dados (Mayernyik, 2011). A automatização da anexação de metadados, sempre

que possível, pode melhorar a quantidade e a qualidade do uso de metadados na prática (Borgman *et al.*, 2016, tradução nossa).

A partilha de dados facilita e faz avançar a ciência. Os seus benefícios não se limitam à comunidade científica, acolhem e influenciam toda a sociedade, sem exceção. É na partilha que novas perspetivas se levantam, que se identificam erros, se desencorajam fraudes. Contribui igualmente para o treino de novos investigadores e evita desperdício de recursos. A ciência enquanto tomada de decisão é mais avançada e promove benefícios para o pesquisador (Piwowar *et al.*, 2007). De acordo com um estudo publicado na PlosOne, o compartilhamento dos dados de pesquisa foi associado a um aumento de 69% nas citações, independentemente do fator de impacto do periódico, data de publicação e país de origem do autor (Fienberg *et al.*, 1985; Piwowar *et al.*, 2007; Veiga *et al.*, 2021).É consensual, na comunidade científica, que o número de citações de um artigo indica o seu impacto em pesquisas subsequentes, embora haja debates sobre se o impacto reflete a qualidade e qual o efeito da auto citação e da coautoria na contagem de citações (Avkiran 1997; Garfield 1986; Glanzel e Schubert 2001; Persson *et al.* 2004; Rousseau 1992 *apud* Zhao, 2010).

Os investigadores preocupam-se com a disponibilidade das suas publicações a longo prazo; poucos estão dispostos a fazer investimentos comparáveis na longevidade dos seus dados. As infraestruturas de conhecimento são definidas mais simplesmente como redes robustas de pessoas, artefactos e instituições que geram, compartilham e mantêm conhecimento específico sobre o ser humano e mundos naturais (Borgman, et.al. , 2015).

Para obtenção de melhores dados, ou seja, dados adequados para curadoria, reutilização e partilha, a sua recolha deve ser feita da forma mais limpa e o mais cedo possível, no seu ciclo de vida. Estabelecer normas sobre fontes, estruturas e formatos de dados promoverão o desenvolvimento de infraestrutura de informação.

A integridade dos dados começa nos estágios iniciais do ciclo. A menos que cientistas e outros utilizadores subsequentes de dados de implantações de sensores dinâmicos possam confiar na integridade dos dados, em cada estágio de processamento, esses dados terão um valor mínimo (Borgman *et al.*, 2007).

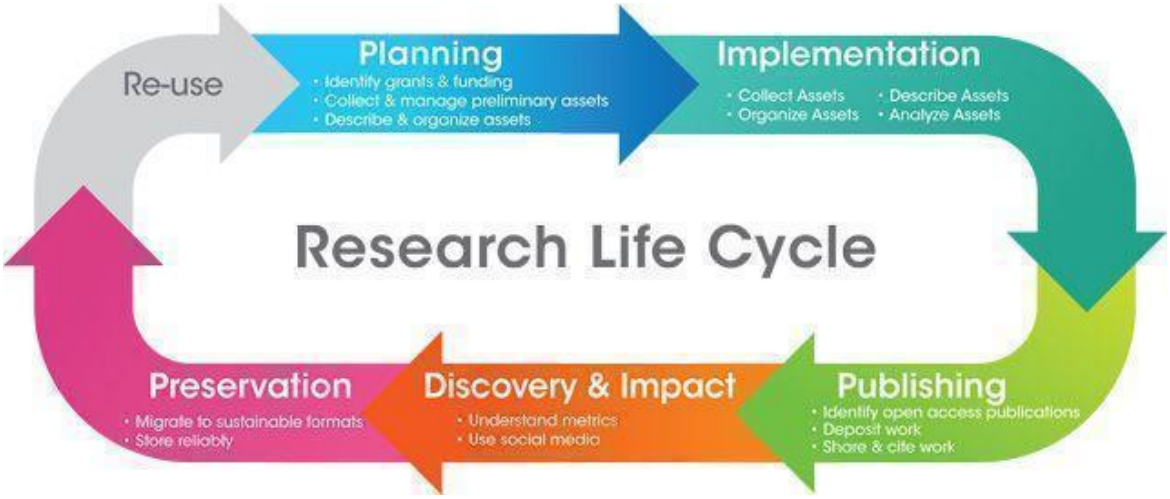


Figura 6 Research Life Cycle (University of California, Irvine, Libraries, Digital Scholarship Services, 2019) in (Borgman, 2019)

O Ciclo de vida da investigação encoraja os investigadores a repensar a forma dos seus trabalhos terem um impacto duradouro através da disseminação e preservação dos seus produtos de investigação. Os profissionais de informação são elementos basilares na infraestrutura do conhecimento, com os seus conhecimentos e práticas especializadas. Há o risco dos investigadores, ao não preservarem devidamente os seus conjuntos de dados, não terem os seus trabalhos citados ou publicados, com a fiabilidade que lhes garanta a disseminação entre pares, ou os seus dados não serem encontrados e acessíveis a outros investigadores, quebrando assim as linhas de investigação. O requisito de um PGD e o depósito dos dados associados às publicações são, entre outros (políticas para a ciência aberta, acesso aberto, etc.) a confirmação de que os dados de investigação são "bens" a preservar seja para a sua reutilização, para a sua reprodutibilidade ou quaisquer outros objetivos que visem o avançar da ciência em cooperação.

> The lives and afterlives of data depend upon many factors, such as their perceived value and the efforts invested in their curation. In data science, we ignore knowledge infrastructures at our peril. Identifying principles for what to keep, why, how, and for how long, is the challenge of our day. (Borgman, C. L., 2019)

É crucial uma base de confiança na origem dos dados para a sua reutilização. Como apontam Carlson e Anderson (2007), há uma cisão entre os dados e quem os utiliza ou recolhe, suscitando a questão se são fiáveis ou entendidos por quem os está a reutilizar. Nos casos de estudo que os autores apresentaram, esta desconexão mostrou que era imperativo que os dados fossem inteligíveis, bem como explícitos os seus contextos de produção, aliados a sistemas de verificação e avaliação apropriados e fiáveis. Percebeu-se também, pelas entrevistas e observação, que as práticas dos investigadores no que toca a dados são altamente específicas e qualitativas, mesmo nas disciplinas quantitativas. Os dados são "cozinhados", ou seja, há uma alteração na forma como são tratados os dados primários. No exemplo apresentado - Skyproject[12], uma imagem pode ser vista por filtros diferentes, alterando o resultado para o investigador, há uma calibração necessária com determinado algoritmo, que terá de se manter igual para todas as imagens. Como este, muitos outros casos de estudo afloram esta problemática (Banning, 2020).

A fiabilidade dos dados também pode ficar comprometida caso ocorra a recolha e o processamento de dados pelos mesmos agentes, ficando muita informação por explicitar a futuros reutilizadores, por haver um conhecimento tácito que não é revelado (Kanfer *et al.*, 2000). Num estudo de caso levado a cabo por Borgman *et al.* (2007), concluem que os cientistas estão mais dispostos a partilhar dados de estudos já publicados, do que dados de investigações que planeiam publicar, facto já constatado por alguns autores (Latour, 1987; Latour & Woolgar, 1986). Os cientistas ainda não sentiam segurança e confiança na partilha dos dados das suas investigações, para além de alguma falta de incentivos. Um movimento

[12] Is a £10M project initiated in 2001 by a consortium of 11 departments. Its distributed team is composed of scientists, software developers, and managers aiming at building the infrastructure for a data grid for U.K. astronomy, which will form the U.K. contribution to a global virtual observatory. It works closely with similar projects worldwide through an international alliance. The infrastructure developed enables the first beta-users to perform queries across distributed datasets through the SkyProject portal in order to access sequences of observations from a range of telescopes. Demonstrator tools accessed via a PC-based ‘‘workbench’’ have included the self-assembly of sky-movies and automated filtering and processing of observation data held on a range of distributed databases. *in* What Are Data? The Many Kinds of Data and Their Implications for Data Re-Use

que tem tendência a melhorar com as medidas certas. Borges (2017) sublinha num estudo de Tenopir *et al.,* (2011):

> [...]algumas das algumas das razões que podem encorajar os investigadores a partilhar ao menos uma parte dos seus dados e a reutilizar outros, em condições definidas, se obtiverem alguma vantagem com isso: a obtenção do crédito através da citação da origem dos dados ou a cópia dos artigos onde os dados foram utilizados encontram-se entre os exemplos apontados.[...]Não temos qualquer dúvida em afirmar que, das funções que cumpre a comunicação da ciência, aquela ligada às questões da propriedade intelectual é a que tem potenciado a transformação porque está intrinsecamente ligada ao reconhecimento do contributo e assim à 'acumulação de crédito'. (Borges, 2017)

Num estudo de pequena dimensão, mas muito detalhado, relativo a uma Universidade de Investigação (Pritchard, Carver & Anand, 2004), concluiu-se que os investigadores com procedimentos mais automatizados na recolha e análise de dados eram os mais propensos a partilhar dados brutos e análises; estes também tendiam a ser os maiores grupos de pesquisa. No caso em que a produção de dados era automatizada, mas outros procedimentos relativos à gestão dos dados eram mais laboriosos, os investigadores tendiam a guardar os dados sem partilhar. Esses comportamentos ocorreram em todas as disciplinas; eles não eram específicos para a ciência.

Como sublinha Borgman, há elementos de peso que desincentivam os académicos na partilha de dados: (i) o corpo docente recebe mais recompensas pela publicação de artigos e livros do que pela divulgação de dados; (ii) o esforço dos indivíduos para documentar seus dados para uso por outros é muito maior do que o esforço necessário para documentá-los apenas para seu uso e da sua equipa de pesquisa; (iii) os dados e as fontes oferecem vantagem competitiva e são essenciais para estabelecer a prioridade dos sinistros; (iv) os dados são frequentemente vistos como propriedade intelectual própria a ser controlada, sejam ou não os dados (ou suas fontes) de propriedade legal (Borgman, 2010).

A gestão de dados é morosa e trabalhosa, o que significa que não há uma relação de esforço-proveito que incentive a proatividade dos cientistas nesta área. Documentar dados, como deveriam, é muito mais difícil do que um simples resumo dos dados anexo à publicação. Um papel onde os profissionais de Informação surgem como suporte no

processo de investigação. Se os dados científicos devem ser aproveitados para comunidades maiores, as práticas científicas devem refletir-se nas bibliotecas digitais de dados, por forma a tornar a documentação e o compartilhamento de dados atraentes. Estes podem incluir mecanismos para bibliotecas digitais pessoais, atribuição e proveniência, suporte, períodos de embargo para acesso e segurança (Arzberger *et al.*, 2004; Borgman, 2004; 2007; Ribes *et al.*, 2005; Hilgartner & Brandt-Rauf, 1994). A partilha de dados implica a comunicação de algo a um conjunto de outros, potencialmente desconhecidos. Como, o que comunicar e para quem, é mais problemático do que os relatos ingénuos de colaboração científica presumem (Carlson, S., & Anderson, B., 2007). É um "*conundrum*" (Borgman, 2011), um problema intrincado e difícil, que implica lidar com grossas camadas de complexidade no que toca à natureza dos dados, investigação, inovação, bolsas de estudo, incentivos, recompensas, propriedade económica e intelectual e políticas públicas. Entre as muitas questões que se levantam na partilha de dados, há pontos desafiantes e onerosos a considerar (Bell, Hey, & Szalay, 2009; Hey, Tansley, & Tolle, 2009; Halevi & Moed, 2012; Higman & Pinfield, 2015).

Apesar de haver um consenso global quanto aos benefícios da partilha e reutilização de dados como motor de desenvolvimento científico (Fienberg *et al.*, 1985; Tenopir *et al.*, 2011), a sua efetivação de modo sistemático e efetivo ainda está longe do ideal (Borgman, 2011; Pampel & Dallmeier-Tiessen, 2014; Tenopir *et al.*, 2011). Há impedimentos metodológicos, legais, técnicos e falta de incentivos para os investigadores (Asher *et al.*, 2013; Bourne, 2010; Bourne *et al.*, 2012; Douglass, Allard, Tenopir,Wu, & Frame, 2014).

Há que equacionar práticas adequadas de compartilhamento de dados e o desenvolvimento de uma política coerente e de um quadro jurídico a nível nacional, apoiando os princípios internacionais de acesso, por forma a que os investigadores os percecionam de forma clara e sejam identificados mecanismos para tornar a comunidade ciente dos conjuntos de dados disponíveis, facilitar a sua compreensão e promover a sua reutilização efetiva. Os métodos de pesquisa baseados em dados são mais valiosos quando permitem que os académicos façam novas perguntas de novas maneiras (Borgman, 2010). Embora as estruturas nacionais para compartilhamento de dados estejam bem estabelecidas

nos Estados Unidos e na Europa, esse não é o caso em muitas outras jurisdições (Fitzgerald *et al.*, 2009; Candela *et al.*, 2015).

São apontados outros limites à partilha, nomeadamente a seleção dos dados a serem compartilhados, sob que condições, quem tem autoridade para o efeito, questões de ética, quem tem acesso aos dados e o seu uso indevido. Um dos dilemas prende-se com a pouca necessidade ou capacidade observada nos investigadores em utilizarem os dados de outros ou partilharem os seus. Daí não verem a necessidade de criar procedimentos normalizados nas suas práticas de investigação (Zimmerman, 2008).

Note-se, para além dos elementos já citados, o tratamento dos dados e políticas de partilha nas áreas das Humanidades (Poole & Garwood, 2020). Na falta de uma perspetiva externa, os estudiosos de humanidades precisam estar particularmente atentos às suposições não declaradas sobre seus dados, fontes de prova e epistemologia. Estamos apenas a começar a entender o que constitui dados nas humanidades, sem falar em como os dados diferem de estudioso para estudioso e de autor para leitor. Como Allen Renear observou: "Nas humanidades, os dados de uma pessoa são a teoria de outra" (comunicação pessoal, 22 de junho de 2009). Apesar das divergências, as fronteiras entre as ciências e as humanidades estão a desvanecer-se no que toca às práticas na eScience (Borgman, 2010). As diferentes culturas epistémicas têm práticas e métodos diferentes na forma de fazer ciência.

A curadoria e a qualidade de dados são entendidas como elementos estratégicos (Wilkinson *et al.*, 2016) no contexto de gestão de dados tendo muito autores contribuído para o aprofundamento do tema como demonstram Piccolo *et al.* (2022) no seu estudo "Qualidade de dados em gestão de dados de pesquisa".

2.2. PRINCÍPIOS FAIR

A qualidade dos conjuntos de dados de pesquisa (Batini *et al.*, 2009; Batini & Scannapieco, 2006; Hacid *et al.*, 2019) é muito importante e precisa definir os critérios para avaliar a qualidade dos conjuntos de dados. Uma boa gestão de dados é também essencial

para facilitar descobertas, inovação e reutilização para uma comunidade, depois do processo de publicação.

Em 2011 foi criada a "FORCE11[13]", uma comunidade de académicos, bibliotecários, arquivistas, editores e financiadores de pesquisas. Escreveram um manifesto que, entre outros assuntos relevantes, aponta uma revisão de artefactos de comunicação. Como ponto de partida, a sua visão envolve a criação de uma forma nova e enriquecida de publicação académica, que permita a criação e gestão de relações entre conhecimento, reivindicações e dados: (i) criação de uma infraestrutura de conhecimento que permita a partilha de componentes executáveis computacionalmente (fluxos de trabalho, código de computador e cálculos estatísticos, componentes de conteúdo cientificamente válidos); (ii) uma infraestrutura que permita que esses componentes sejam disponibilizados, revistos, referenciados e atribuídos; (iii) sistemas de recompensa que incentivem académicos e investigadores a participar e contribuir; (iv) mostrar aos académicos (também nas suas funções de autores, editores e revisores) os benefícios de uma melhor comunicação do conhecimento, usando *media* mais ricas, permitindo uma interpretação mais fácil, rápida e aprofundada das informações pelos consumidores (outros académicos, alunos e docentes, agências governamentais e não governamentais, indústria, *media* e sociedade em geral).

Essas novas e melhoradas formas de comunicação permitirão avaliações mais precisas da qualidade e do impacto do trabalho dos académicos, facilitando melhores avaliações de promoção e avaliações de propostas. Para bibliotecários e arquivistas, embora a acessibilidade *online* signifique que os acervos de bibliotecas tradicionais se tornem menos relevantes, o armazenamento, a atualização e a manutenção de dados e softwares digitais aumentarão em importância. A adaptação a essas mudanças trará novos modos de serviço aos utilizadores. A primeira iniciativa foi o DSA (Data Seal of Approval) para repositórios em 2010-2011 que avaliava os repositórios de dados.

[13] "FORCE11 is a community of scholars, librarians, archivists, publishers and research funders that has arisen organically to help facilitate the change toward improved knowledge creation and sharing. Individually and collectively, we aim to bring about a change in modern scholarly communications through the effective use of information technology" in https://force11.org/info/

Desde o final de 2016 o grupo FORCE 11 definiu os princípios **FAIR**. O acrónimo significa: **F**indable, **A**ccessible, **I**nteroperable e **R**eusable (Localizável, Acessível, Interoperável, Reutilizável). É um conjunto de princípios orientadores e práticas aceites pela comunidade, numa linguagem e procedimentos acessíveis, claros e globais tanto pelos humanos, como por sistemas computadorizados. Os princípios FAIR relacionam-se com os princípios elaborados de citações de dados definidos JDDCP (*Joint Declaration of Data Citation Principles* - FORCE11, 2013). A sua implementação no acesso aos resultados da investigação faz parte do ecossistema da Ciência Aberta. Os dados de investigação têm um ciclo de vida, cujo entendimento e partilha permite ir ao encontro dos princípios FAIR (Wilkinson, 2016).

PRINCÍPIOS FAIR[14]

1. **F**INDABLE (LOCALIZÁVEIS)

Para tornar os dados localizáveis, é recomendado:

1.1. A atribuição de um identificador único persistente aos (meta)dados;

1.2. A descrição dos dados com metadados pormenorizados.

1.3. O registo ou a indexação dos (meta)dados num recurso pesquisável.

1.4. A inclusão do identificador nos metadados.

2. **A**CCESSIBLE (ACESSÍVEIS)

Tornar os dados acessíveis significa que:

2.1. Os (meta)dados são recuperáveis através do seu identificador, mediante um protocolo de comunicações normalizado.

2.1.1. O protocolo de comunicação é aberto, gratuito e universalmente implementável.

[14]in Princípios FAIR traduzido de FAIR Principles, URL: FAIR Principles / acedido em 2022-04-26

2.1.2. O protocolo de comunicações permite um procedimento de autenticação e autorização, quando necessário.

2.2. Os (meta)dados permanecem acessíveis, mesmo se os dados já não estiverem disponíveis.

3. **I**NTEROPERABLE (INTEROPERÁVEIS)

Os dados serão interoperáveis se:

3.1. Os (meta)dados usam uma linguagem formal, acessível, partilhada e de ampla aplicabilidade para a representação do conhecimento.

3.2. Os (meta)dados usam vocabulários que seguem os princípios FAIR.

3.3. Os (meta)dados incluem referências qualificadas a outros (meta)dados.

4. **R**EUSABLE (REUTILIZÁVEIS)

Assegurar que os dados são reutilizáveis significa que:

4.1. Os (meta)dados têm uma pluralidade de atributos precisos e relevantes.

4.1.1. Os (meta)dados são disponibilizados com uma licença clara e acessível de uso dos dados.

4.1.2. Os (meta)dados são associados à sua proveniência.

4.1.3. Os (meta)dados cumprem normas relevantes da comunidade disciplinar

Desde 2016 que as plataformas Digital Science, Figshare e Springer Nature monitorizam os níveis de partilha de dados e o seu uso. O relatório anual que apresenta, intitulado "The State of Open Data[15]", já tinha destacado em 2020 que, de entre mais de 4.500 entrevistados, o número de investigadores que disseram conhecer ou ouviram falar dos princípios FAIR e Open Data está em crescimento. Em dois anos, o desconhecimento desses princípios entre os investigadores caiu de 60% para 39%.

[15] The State of Open Data is now the longest-running longitudinal study on the subject, which was created in 2016, to examine attitudes and experiences of researchers working with open data – sharing it, reusing it, and redistributing it. *in* The State of Open Data 2020 – Global Attitudes towards Open Data - Digital Science

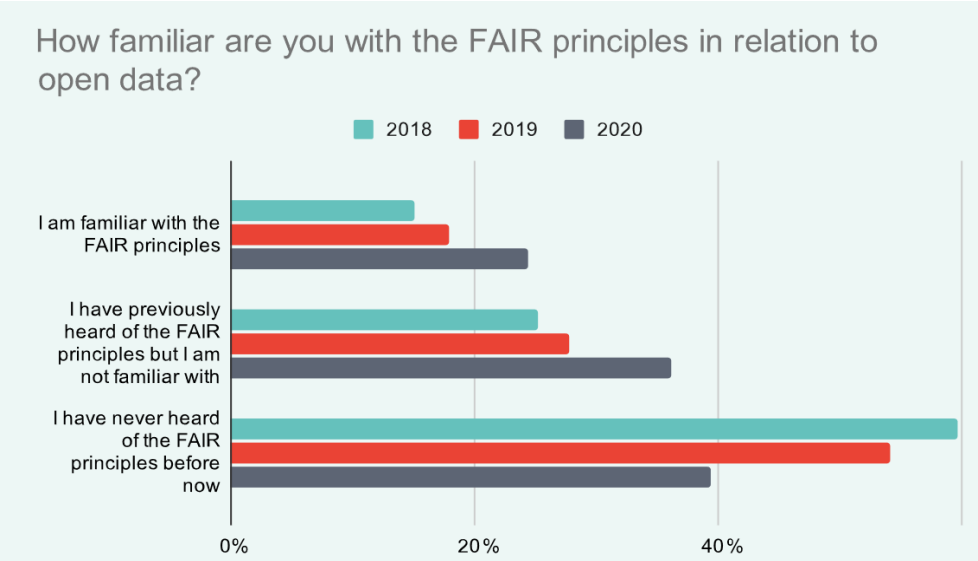


Figura 7 Grau de familiaridade com os Princípios FAIR in (van Selm, 2020)

2.3. PLANO DE GESTÃO DE DADOS (PGD)

Os primeiros indícios de um PGD, segundo pesquisas na literatura científica, referem-se a 1966, usados em projetos aeronáuticos e de engenharia complexos. O conceito evoluiu e só no século XXI é que viria a ser adotado pela comunidade científica na generalidade. (Smale et al., 2018; Miranda et al., 2021)

Um PGD descreve os dados coletados desde o início da investigação, como serão geridos, durante e após o projeto. A criação de um PGD no início de uma pesquisa permite organizar os dados, mantendo-os seguros e garantindo o acesso a quem precisa.

> Researchers can more effectively share data if they keep that objective in mind in all stages of their research. Planning to share data from the outset not only helps achieve the goal of data sharing but also may improve the quality of the research. For example, adequate documentation of data helps initial investigators as well as subsequent analysts. Data files should include the unedited raw data as well and documentation on edits, handling of nonresponse, and similar problems[...] (Fienberg *et al.*, 1985).

As motivações por detrás da elaboração de um PGD são de vária ordem. Começando por ser impulsionada como um requisito no acesso a apoios à investigação

(Antell *et al.*, 2014) serve igualmente inúmeras funções sociais. Os requisitos de agências de financiamento no que concerne os dados de pesquisa, destacam a complexidade da ciência moderna: não apenas a noção de dados, mas visões concorrentes de pesquisa, inovação e bolsa de estudos, incentivos díspares para recolher e divulgar dados e a economia e a propriedade intelectual de produtos de pesquisa e políticas públicas. A promessa de uma bolsa de estudos digital com uso intensivo de dados e tecnologia na ciência baseia-se em sistemas, serviços, ferramentas, conteúdo, políticas, práticas e recursos humanos disponíveis para descobrir, explorar e usar produtos de pesquisa (Darch & Borgman, 2014).

A maioria das pesquisas em Ciência da Informação sobre o desenvolvimento de infraestruturas de gestão de dados científicos, refere-se ao modo de apoiar o trabalho científico das comunidades que se destinam a servir (Palmer *et al.*, 2007; Ribes & Finholt, 2009). As práticas na ciência moderna dependem das práticas de geração, disseminação e análise de dados. Distinguem-se tanto pela escala massiva de produção de dados, como pela dispersão global dos recursos dos dados. Com as novas formas de instrumentação, as taxas de geração de dados aumentam rapidamente, fazendo com que os cientistas precisem de apoio na identificação e seleção de dados, úteis em contextos individuais e na preservação e curadoria de dados que tenham valor futuro, seja para os originadores ou para outros (Borgman *et al.*, 2007), sendo o PGD essencial à prossecução deste objetivo. O que é um Plano de Gestão de Dados (PGD)?

> Data Management Plans (DMPs) are a key element of good data management. A DMP describes the data management life cycle for the data to be collected, processed and/or generated by a Horizon 2020 project. As part of making research data findable, accessible, interoperable and re-usable (FAIR), a DMP should include information on:
>
> - the handling of research data during & after the end of the project
> - what data will be collected, processed and/or generated
> - which methodology & standards will be applied
> - whether data will be shared/made open access and
> - how data will be curated & preserved (including after the end of the project). (*in* Data management - H2020 Online Manual)

Um PGD é a base referencial durante o todo o processo de investigação, deve antecipar necessidade e requisitos e descrever políticas e métodos de todo o ciclo de vida dos dados.

O Horizonte 2020 foi o programa de financiamento de pesquisa e inovação da Comissão Europeia (CE) de 2014-2020 com um orçamento de quase 80 bilhões de euros. O programa foi sucedido pelo Horizonte Europa, com um orçamento de 95,5 bilhões de euros para reforçar as bases científicas e tecnológicas da CE e o Espaço Europeu da Investigação (ERA).

Nos Estados Unidos, a NSB (NATIONAL SCIENCE BOARD, 2005) apresenta requisitos para a elaboração de um PGD equivalentes à europeia, definindo PGD como um plano que descreve os dados que serão criados, como serão geridos e disponibilizados ao longo de sua vida útil, em que o PGD deve ser parte integrante de um projeto de pesquisa, desde o seu início. Deve incluir os tipos de dados a serem criados; os padrões que seriam aplicados para formato, conteúdo de metadados, etc.; disposições para armazenamento e preservação; políticas e disposições de acesso; e planos para eventual transição ou término da recolha de dados no futuro de longo prazo. Podemos encontrar variáveis nos requisitos solicitados num PGD (Shankar, 2003; Williams *et al.*, 2017), mas a linha condutora que os estrutura é semelhante, variando nas especificidades do contexto da sua criação e finalidades.

Uma boa gestão de dados de pesquisa não é um objetivo em si, mas sim o principal canal que leva à descoberta e inovação do conhecimento e à subsequente integração e reutilização de dados e conhecimento. Apesar de representar um acréscimo de trabalho e dispêndio de tempo no processo de investigação, os dividendos que daí decorrem são inegáveis. Fornece um estímulo positivo para pensar sobre como os dados gerados dentro de um projeto serão armazenados, geridos e protegidos, e possivelmente partilhados, para serem reutilizados, cumprindo o objetivo de fazer ciência de modo cooperativo. Deve fazer parte do processo de pesquisa desde o início. À medida que um projeto avança, os dados

gerados podem sofrer mudanças de tipo e volume (ERC Scientific Council, 2022). É útil considerar um PGD como uma estrutura, que deve ser mantida e modificada à medida que a pesquisa avança. Um PGD no início do ciclo de pesquisa facilitará o processo de publicação, economizará tempo, protegerá as informações e aumentará a visibilidade e o impacto dos resultados da pesquisa.

Os princípios FAIR são uma das exigências que acompanham o PGD. Para os investigadores, a mudança para dados FAIR significa que eles precisam pensar sobre quais os dados que a sua pesquisa produzirá, como esses dados serão descritos e como podem ser disponibilizados de forma a beneficiar a ciência e a sociedade em geral. Isso significa que têm de elaborar um PGD e encontrar repositórios de dados adequados.

As agências de financiamento de pesquisa estão cientes da importância da existência de uma infraestrutura e serviços para a curadoria dos dados de pesquisa e exigem o seu devido tratamento. As bibliotecas de pesquisa académica foram identificadas como locais para basear esses serviços de dados de pesquisa (RDS - Research Data Services). Os serviços de dados de pesquisa incluem planos de gestão de dados, curadoria digital (seleção, preservação, manutenção e armazenamento) e criação e conversão de metadados (Tenopir *et al.*, 2013). São necessárias estratégias adequadas de gestão e preservação de dados para os tornar disponíveis para pesquisas longitudinais e multidisciplinares, bem como para múltiplos públicos e propósitos. Os dados e informações devem ser preservados para a gestão de riscos que incluem as ameaças de perda de dados, a necessidade de proteger o trabalho intelectual dos investigadores e a usabilidade contínua dos dados que podem ser necessários para corroborar afirmações científicas (Shankar, 2003).

Nos Estados Unidos da América, a National Science Foundation[16] (NSF) exige, desde janeiro de 2011, a todas as propostas de financiamento, de qualquer tamanho ou diretoria, um PGD (Dissemination and Sharing of Research Results - NSF Data

[16] The National Science Foundation (NSF) is an independent federal agency created by Congress in 1950 "to promote the progress of science; to advance national health, prosperity, and welfare; to secure the national defense..." NSF is vital because we support basic research and people to create knowledge that transforms the future. *in* National Science Foundation

Management Plan Requirements). Este requisito incentiva fortemente a partilha e revisão pelos pares e pode definir se um projeto é financiado ou não, mediante a gestão que o investigador demonstre no PGD sobre os seus dados: quais são, como serão geridos, como serão partilhados ou não e porquê. Esta ação da NSF provocou uma ampla discussão sobre o compartilhamento de dados entre as partes interessadas em pesquisas com financiamento público. No seguimento dessa imposição local para aceder a financiamento, quase todas as instituições de pesquisa passaram a incluir a exigência de um PGD aos seus investigadores, estendendo-se a prática noutros países. A NSF não é a primeira agência de financiamento a impor a exigência de gestão de dados. Em 2003, os Institutos Nacionais de Saúde[17] (INS) implementaram uma exigência semelhante, e algumas outras agências governamentais e fundações privadas também o fizeram. No entanto, devido ao tamanho e influência da NSF, o impacto global foi mais alargado e traduzido numa prática efetiva (Antell *et al.*, 2014). Desde então, surgiu o desenvolvimento de repositórios de dados e serviços para apoiar esse tipo de atividade. Williams *et al.* (2017) numa pesquisa realizada na web, analisaram uma lista de organizações que financiaram pesquisas na instituição do autor principal para os anos civis de 2010–2014 e identificou mais de 50 documentos de requisitos de plano de compartilhamento ou gestão de dados.

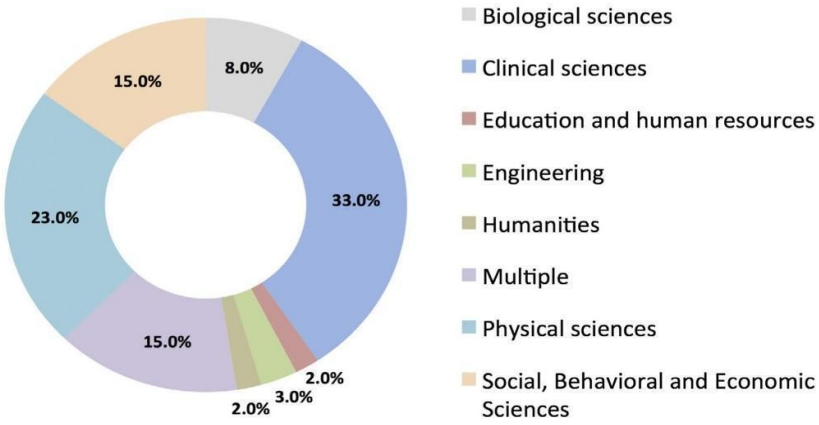


Figura 8 Disciplinas financiadas por agências com requisitos de PGD in (Williams et al., 2017)

[17] The National Institutes of Health (NIH), a part of the U.S. Department of Health and Human Services, is the nation's medical research agency — making important discoveries that improve health and save lives *in* Who We Are | National Institutes of Health (NIH)

Reportando-se também a 2014, Cox *et al.* (2017) observaram que em grande parte dos países, a maioria das instituições que responderam à pergunta sobre políticas no final de 2014 (n = 167) já tinham uma política de GDI (Gestão de Dados de Investigação) ou esperavam ter uma em 12 meses (Austrália 94%, Canadá 40%, Alemanha 100 %, Irlanda 71%, Holanda 100%, NZ 71%, Reino Unido 86%).

O que muitas vezes não é explícito nas discussões de planos de gestão de dados e requisitos de compartilhamento de dados são os conflitos de interesses e as diferentes motivações das muitas partes interessadas envolvidas. Essas motivações e interesses, bem como os incentivos e desincentivos ao compartilhamento daqueles que produzem dados de pesquisa, precisam ser trazidos para primeiro plano (Borgman, 2011).

Nas metas a atingir, comum a todos os *stakeholders*, devem estar presentes as seguintes premissas: garantir que todas as obrigações legais e expectativas da comunidade para proteger a privacidade, segurança e propriedade intelectual sejam totalmente atendidas; participar do desenvolvimento de padrões comunitários para recolha, deposição, uso, manutenção e migração de dados; trabalhar para a interoperabilidade entre comunidades e incentivar a integração interdisciplinar de dados; garantir que as decisões da comunidade sobre recolha de dados levem em consideração as necessidades dos usuários fora da comunidade; incentivar o acesso livre e aberto sempre que possível; e fornecer incentivos, recompensas e reconhecimento para cientistas que compartilham e arquivam dados (National Science Foundation, 2005). A NRS (National Research Council, 1999) enumera alguns impedimentos na concretização das metas referidas: Incompatibilidades entre sistemas de gestão e descrição de dados nas práticas de comunidades científicas diferentes; Acessibilidades diferentes a ferramentas[18], recursos e arquivos de dados extensos para armazenar conjuntos de dados da comunidade; Desafios técnicos de redes de sensores; Objetivos divergentes das organizações que produzem e distribuem bancos de dados científicos e técnicos; Custos, preços e acesso dos bancos de dados científicos e técnicos;

[18] Ferramentas disponíveis *on-line* para elaboração de PGD: ERC Data Management Plan Template; DMPonline tool; ARGOS tool; DMPTool – California Digital Library

Custos de produção e distribuição; Preços e acesso; Proteção Estatutária Mais Forte e Incentivos ao Investimento.

Em Portugal, a política de gestão e partilha de dados FCT (Fundação para a Ciência e a Tecnologia) determina que todos os resultados de investigação financiada sejam disponibilizados em acesso aberto, sempre que possível; que a gestão dos dados de investigação seja um requisito para os beneficiários de financiamento de I&D da FCT e que os dados sejam geridos de acordo com os princípios FAIR. Determina ainda como requisitos o depósito dos dados resultantes da investigação num repositório de dados de investigação fiável; o uso de identificadores persistentes e licenças normalizadas; informar das ferramentas e instrumentos necessários à validação dos resultados e incluir referência ao financiamento de acordo com as especificações previstas. Mas de acordo com Miranda *et al.* (2021) ainda há alguma dificuldade de aceitação por parte da comunidade do valor do PGD. Para uma evolução no estado de coisas sublinha a normalização do conhecimento contido no PGD como o que se apresenta no grupo de trabalho DMP Common Standards da RDA.

O grupo de trabalho do 8º Fórum de Dados define quatro componentes principais para uma estratégia institucional: i) sensibilizar/ aumentar a consciência; ii) avaliar a prontidão institucional; iii) formalizar as práticas de GDI (Gestão de Dados de Investigação; iv) definir um roteiro. Apresenta-nos as estratégias institucionais da Universidade do Minho, da Universidade de Coimbra, e as políticas de referência praticadas em Ghent University, University of Amsterdam ou University of Cambridge (Boavida et al., 2021)

2.4. METADADOS

> Finally, processing and curatorial activities generate derivative data. [...]To make data usable, it is necessary to preserve adequate documentation relating to the content, structure, context, and source (e.g., experimental parameters and environmental conditions) of the data collection – collectively called metadata. Ideally, the metadata are a record of everything that might be of interest to another researcher. For computational data, for instance, preservation of data models and specific software is as important as the preservation of data they generate. Similarly, for observational and laboratory data, hardware and instrument specifications and other contextual information are critical. Metadata is crucial to assuring that the data element is useful in the future. (National Science Board, 2005)

Uma das primeiras formas de metadados conhecidas é a obra de Pinakes, criada em 245 a.C. por Kallimachos de Cirene para catalogar a Biblioteca de Alexandria de forma sistematizada (Pomerantz, 2015). Os metadados existem desde que os humanos organizam informações apesar de atualmente ser, cada vez mais, sob a forma digital e muito mais aberta. Por mais de um século, e particularmente desde os primeiros desenvolvimentos de padrões descritivos nacionais e internacionais, a criação e gestão de metadados foi principalmente da responsabilidade dos profissionais da informação envolvidos na catalogação, classificação e indexação (Baca *et al.*, 2016).

Jack E. Myers, fundador da The Metadata Company, afirma ter cunhado o termo em 1969, na tentativa de sintetizar o conceito "sobre os dados", provavelmente emprestado da coleção de ensaios de Aristóteles sobre Metafísica (Steiner, Tobias, 2018), acrescentando o prefixo grego *τὰ μετὰ*, com a significação de "junto a, depois de, entre, com algo sobre esse mesmo algo" ao termo latino *data*. Myers registou uma marca para a palavra sem hífen - "metadados", em 1986. Apesar disso, referências ao termo aparecem em trabalhos académicos que antecedem a afirmação de Myers. Os professores do Massachusetts Institute of Technology, David Griffel e Stuart McIntosh, num artigo académico publicado em 1967, descreveram metadados como "um registo dos registos de dados" que resultam quando dados bibliográficos sobre um tópico são recolhidos de fontes discretas. Os investigadores concluíram que uma "abordagem metalinguística" ou "metalinguagem" é necessária para permitir que um sistema de computador interprete adequadamente esses dados e o seu contexto para outros dados relevantes. Ao contrário de Myers, Griffel e McIntosh trataram "meta" como um prefixo para "dados" (Kranz, sem data).

Na década de 1980, a área da Biblioteconomia utilizava expressões como descrição bibliográfica, dados catalográficos, ou simplesmente catalogação para se referir aos metadados. Os Metadados ou "dados sobre dados"[19] traduzido à letra, possibilitam a

[19] Há divergências quanto à definição de metadados. Consultar (Pomerantz, 2015, p. 26) que refuta a noção de "dados sobre dados" classificando-a como "not a useful definition of data" e "almost meaningless".

descrição, gestão e descoberta de dados e sistemas interoperáveis ou metadados que devem ajudar a contextualizar os conjuntos de dados. São dados representacionais que, adicionados à própria informação, adquirem um valor semântico para substituí-la ou representá-la. É a gestão normalizada de estruturas de organização da informação, dependentes do tipo e finalidade das fontes. Devem descrever o formato e o conteúdo do conjunto de dados, as circunstâncias da sua recolha, procedimentos usados para os manipular ou modelar, custódia, a sua qualidade, informações de preservação e descrição específica da disciplina (Shankar, 2003). Dublin *Core* é um padrão geral amplamente utilizado, originalmente desenvolvido para auxiliar na indexação de catálogos de fichas de bibliotecas físicas. Desde então, o padrão foi adaptado para metadados digitais baseados na Web. Descreve os atributos de 15 elementos de dados principais: título, criador, assunto, descrição, editora, colaboradores, data, tipo, formato, identificador, fonte, idioma, relação, cobertura e gestão de direitos. Um padrão de metadados bibliográficos semelhante é o MODS (Metadata Objects Description Schema), um esquema baseado em XML para bibliotecas, gerado pelo *Network and Standards Development Office* da Biblioteca do Congresso dos EUA, como um sucessor dos padrões *Machine-Readable Catalog* desenvolvidos na década de 1960.

Um padrão mais recente, o schema.org, é baseado na colaboração de software de código aberto que fornece uma coleção de esquemas de metadados voltados para dados estruturados da Internet, e-mail e outras formas de dados digitais (Kranz, sem data).

3 INICIATIVAS EM LITERACIA DOS DADOS DE INVESTIGAÇÃO

O Digital Curation Centre (DCC) é um centro de especialização líder mundial em curadoria de informações digitais com foco na capacitação, capacidade e competências para gestão de dados de pesquisa e representa um papel importante na literacia da gestão de dados, disponibilizando ferramentas teórico-práticas na curadoria de dados. É um exemplo de uma iniciativa não institucional, que proporciona eventos e worshops sobre o tema [20], em estreita colaboração com Libraries, learning resources and research | Jisc, o que demonstra até que ponto a importância dos dados de investigação penetrou no tecido social das comunidades, para além do meio académico.

Mas obviamente que o meio académico é o ambiente mais fértil para as iniciativas em literacia dos dados de investigação e onde têm vindo a proliferar. A prática de olhar para os que têm sucesso com iniciativas desta natureza, permite a aprendizagem de modelos e recomendações para as bibliotecas que pretendem evoluir na gestão de dados e incrementar as competências *core* dos seus profissionais. Esforços anteriores em literacia de dados já foram realizados com bastante sucesso (exemplos da Scholars' Lab and Research Computing Lab / University of Virginia; The Science Data Literacy project at Syracuse University; The Purdue University Libraries (Carlson & Johnston, 2015). Yoon e Schultz (2017) apontam como exemplos de sucesso na prestação de serviços de gestão de dados de pesquisa e partilha das suas experiências e conhecimentos: a Universidade de Purdue, o Massachusetts Institute of Technology (MIT), Universidade de Cornell e Universidade da Califórnia.

Em 2010, a iniciativa em literacia informacional de dados, ou, como é nomeada na obra de Carlson & Johnston "Data Information Literacy (DIL)" levanta questões fundamentais a que tenta dar resposta, para deste modo fornecer linhas de orientação a outras iniciativas, no meio académico. Note-se a relevância dada ao papel dos profissionais de informação, como suporte explícito e implícito, em todo o processo de investigação.

[20] Consultar International Data curation Education Action (IDEA) Working Group A Report from the Second Workshop of the IDEA e Training | DCC

> We developed the Data Information Literacy (DIL) project to answer two overarching questions. First, what data management and curation skills are needed by future scientists to fulfill their professional responsibilities and take advantage of collaborative research opportunities in *e-Science* and technology-driven research environments? Second, how can academic librarians apply their expertise in information retrieval, organization, dissemination, and preservation to teaching these competencies to students? By answering these questions our goals were to build a foundation in the library community for teaching DIL competencies, to teach students DIL competencies appropriate to their discipline, and to develop a robust process for librarians to develop DIL curricula and programming (Carlson & Johnston, 2015).

Do projeto DIL derivaram as seguintes propostas de competências *core* a serem introduzidas: introdução a bases de dados e formatos de dados; descoberta e aquisição de dados; gestão e organização de dados; conversão de dados e interoperabilidade, garantia da qualidade; metadados.; curadoria e reutilização de dados; culturas de prática; preservação de dados; análise de dados; visualização de dados; ética, incluindo citação de dados. Todos estes tópicos são desenvolvidos com recomendações detalhadas, sendo um excelente guia de boas práticas.

Tomamos como outro exemplo a publicação de Rice & Haywood (2011) que reporta as iniciativas no contexto de gestão de dados de investigação, na Universidade de Edimburgo[21]. Sucintamente, sublinhamos alguns dos pontos chave das ações levadas a cabo: (i) elaboração de diretrizes para a equipa de pesquisa no planeamento e gestão de dados de investigação, disponíveis na página WEB da universidade; (ii) formação de pós-graduados e investigadores em início de carreira, incorporando boas práticas nas suas atividades de investigação e assim contribuírem para mudanças culturais; (iii) desenvolvimento de políticas para esclarecer expectativas e responsabilidades entre a instituição e sua equipa de pesquisa; (iv) análise de lacunas do serviço. A equipa da Biblioteca de Dados e o gestor do projeto Data Asset Framework (DAF) formularam um conjunto de páginas WEB, com o propósito de criar diretrizes gerais para a equipa de pesquisa. Estas diretrizes foram incorporadas no site dos Serviços de Informação em setembro de 2009 e ganharam alguma atenção entre os profissionais da comunidade de curadoria digital. A captação da atenção dos utilizadores teve por base algumas estratégias direcionadas, que podem ser indicadores de sucesso: criação de páginas curtas, com definições e políticas de financiadores; explicação das vantagens de uma boa gestão de dados; lista de verificação para o planeamento; documentação e metadados; e

[21] "The University of Edinburgh is a research-led Higher Education Institution (HEI) with nearly 27,000 students whose mission is "the creation, dissemination and curation of knowledge." It has been ranked in the top five of UK universities by volume of 4star "world-leading" research in the UK's 2008 Research Assessment Exercise and twentieth in the world according to the 2009 Times Higher Education-QS World University Rankings. Within the University, Information Services (IS) provides support for the research and education activities of schools across three colleges" *in* (Rice & Haywood, 2011)

armazenamento, segurança e criptografia (com referências cruzadas às informações existentes no site). Outro grupo de páginas abordou conceitos básicos de preservação digital, partilha de dados e depósito de dados num repositório, além de contatos-chave para obter ajuda dentro da instituição. (Rice & Haywood, 2011)

Em 2013 a LERU (League of European Research Universities) produziu o LERU Roadmap for Research Data, oferecendo orientação às universidades europeias para enfrentarem os desafios e oportunidades relacionadas com os dados de investigação. O grupo da RDA (Research Data Alliance) divulgou, em 2016, um documento com o objetivo de apoiar os bibliotecários a envolverem-se na gestão de dados de investigação. O documento sugere 23 recursos de informações úteis organizados em oito categorias: 1) Recursos de aprendizagem, 2) Dados de referência e disseminação, 3) Planos de Gestão de Dados, 4) Literacia de Dados, 5) Metadados, 6) Citação de Dados, 7) Licenças e Privacidade, 8) Preservação Digital, 9) Repositórios de Dados. (Príncipe & Silva, 2018).

Em Portugal, no âmbito das políticas de Ciência aberta, a Comissão Europeia ao definir uma política de dados abertos para o programa Horizonte 2020 tornou obrigatório o depósito dos dados de investigação necessários para validar os resultados apresentados em publicações científicas, alertando para a necessidade de uma definição de planos para a gestão dos dados produzidos. A existência dos Fóruns de Gestão de Dados[22], que já vão na sua 8ª edição, espelham como a comunidade expressa a necessidade de obter mais conhecimento e desenvolver competências sobre o tema.

> Em Portugal, a Fundação para a Ciência e a Tecnologia adotou em 2014, no quadro mais amplo da política de Acesso Aberto, orientações que encorajam os investigadores a disponibilizarem os dados resultantes dos projetos de I&D em bases de dados de Acesso Aberto apropriadas, incentivando ainda os investigadores a promoverem e/ou participarem em iniciativas nacionais e internacionais que procurem as formas mais adequadas de partilha de dados nas diferentes áreas do conhecimento.(Príncipe & Silva, 2018)

[22] O Fórum GDI é um espaço de debate e partilha de ideias, projetos e boas práticas de Gestão de Dados de Investigação que procura juntar gestores de repositórios digitais e data centers, técnicos de informação, bibliotecas, arquivos e curadoria de dados, especialistas de informática, investigadores, cientistas de dados e gestores de ciência de instituições de investigação e organismos de financiamento de ciência. Resulta do desenvolvimento de projetos e infraestruturas de informação e dados científicos, no âmbito da FCT-FCCN, do projeto RCAAP e de várias instituições de investigação e ensino superior e no contexto da construção da Política Nacional de Ciência Aberta. *In* http://forumgdi.rcaap.pt/ .

A plataforma NAU é um serviço desenvolvido e gerido pela Unidade FCCN da Fundação para a Ciência e a Tecnologia (FCT) que permite a criação de cursos em formato MOOC (Massive Open Online Course), ou seja, cursos abertos e acessíveis a todos, produzidos por entidades reconhecidas e relevantes na sociedade, que contam com a participação de milhares de pessoas. Realizaram um curso "O Essencial da GDI", destinado inicialmente a investigadores e alunos de doutoramento, fornecendo as competências básicas na gestão de dados. No entanto referem que poderá alargar-se a outros profissionais, o que vemos com bom grado. (Príncipe *et al.*, 2021)

> "To be information literate, a person must be able to recognize when information is needed and have the ability to locate, evaluate and use effectively the needed information" (American Library Association, 1989).

3.1. COMPETÊNCIAS *CORE* PARA OS PROFISSIONAIS DE INFORMAÇÃO

"When scientists are convinced that librarians understand why communication is the essence of science, then, librarians will find that they will have some enthusiastic, collaborating scientists on their hands." (Garvey, 1979)

As tecnologias e dispositivos de produção e armazenamento de informação fazem parte do quotidiano de quase todas as sociedades atuais. Conjuntos de dados de grande dimensão e heterogeneidade, e tarefas intensivas de grande escala computadorizadas são uma realidade da investigação científica atual, não só nas ciências naturais, como nas humanidades, que estão despertos, cada vez mais, para a necessidade de ter representações digitais de coleções e acervos, anexando metadados descritivos prescritos.

A curadoria, atividade desde sempre ligada aos bibliotecários, é agora uma realidade mais complexa e predominantemente tecnológica. Na década de 90, já Palmer (1996) afirmava "before information professionals can begin to improve existing services in research libraries, they need to understand the information work involved in the research processes of contemporary researchers" sublinhando o necessário e, felizmente, crescente carácter cooperativo da investigação na ciência e oferta de serviços intermediários, auxiliares na transferência e tradução de informações entre comunidades científicas. A adaptação e

reformulação das competências *core* dos profissionais de informação é visível e tem vindo a trilhar o seu caminho, ainda que paulatinamente.

Pretende-se que o bibliotecário adquira novas competências ao compreender as dinâmicas de um processo de investigação, numa perspetiva interdisciplinar, e tenha contacto com diferentes ambientes, conteúdos e modelos de colaboração. Para os profissionais da informação é essencial uma constante atualização de competências e conhecimentos, e uma aproximação efetiva à investigação. Só neste contacto direto se encara o bibliotecário como elemento integrante do processo de investigação. A velocidade na nossa sociedade, fruto das evoluções tecnológicas, em constante e rápido desenvolvimento, imprime um novo perfil a todos os *stakeholders* do processo da ciência.

Kling & McKim (2000) já mostravam preocupação com a questão do desenvolvimento da comunicação académica afirmando que este não deve ser baseado em noções vagas de publicação o que pode levar a uma confusão entre as diversas atividades e interesses das comunidades intelectuais. Outros autores reforçam esta linha de pensamento, concluindo que para haver um apoio efetivo dos investigadores, o desenvolvimento de bibliotecas digitais precisava levar em consideração os muitos tipos de informações de pesquisa e os seus papéis divergentes, em diferentes campos de pesquisa (Agre, 1995; Borgman, 2000; Palmer (2005)

O apoio dos bibliotecários não se limita à receção de dados para depósito. A sua experiência em padrões, práticas e tecnologias de documentação, sendo a documentação de dados uma parceria inevitável entre académicos e profissionais da informação, permite que, juntos, cumpram os requisitos do universo dos dados de investigação. Fornecem uma infraestrutura intuitiva, que permite aos investigadores descobrir, utilizar e reutilizar dados, evitando a duplicação de tempo e esforço gastos na criação desnecessária de novos conjuntos de dados. Uma interpretação dos curadores que recolham, descrevam e conectam dados é a ideia do papel de proxy da comunidade (referenciado no NSF Long-lived Data Collections Report), onde os curadores tentam entender a ontologia do domínio e definir padrões, trabalhando em parceria com os cientistas.

As bibliotecas acompanham a evolução tecnológica, cujo percurso de desenvolvimento tende a acelerar, provocando alterações em todo o tecido social, sendo simultaneamente reflexo e precipitante sociais, na medida em que estão no cerne dos movimentos globalizados do conhecimento e da comunicação da ciência.

Pela revisão da literatura percebemos que há divergência de opiniões sobre os tipos de especialização relevantes e que papéis são considerados necessários ao acompanhamento do ciclo dos dados de investigação, decorrente da heterogeneidade da formação e papéis dos profissionais de informação, na sua correlação com o universo dos dados de investigação. Felizmente, como nos dava conta Griffin (2013) "But even given this uncertainty, it is reassuring to see that bodies of opinion on critical technological issues and best practices are converging."

O 4º paradigma estabeleceu-se, e com ele os dados passaram a ser o centro nevrálgico das novas metodologias de pesquisa. Faltam, no entanto, competências e conhecimentos, por parte de investigadores e académicos, na gestão de dados que permitam a sua reutilização e preservação a longo trecho.

A literacia em dados configura-se como um requisito essencial para qualquer atividade relacionada com dados de investigação. No contato com dados de investigação, seja qual for o ator, terá de ter capacidade de processar, classificar e filtrar grandes quantidades de informações, para as quais são necessárias competências para pesquisar, filtrar, processar, criar e sintetizar informações (Schneider, 2013; Koltay, 2015; Vilar & Zabukovec, 2019).

As bibliotecas assumem-se agora como um agente ativo e crítico que proporciona um conjunto de ferramentas e conhecimentos que facilitam e agilizam a investigação e o acesso ao financiamento da ciência. Alterou-se a forma de utilizar, armazenar e disseminar a informação e, consequentemente, alteraram-se os papéis das bibliotecas e dos seus profissionais. A função de gestão da ciência da Biblioteca, aliada a um fortalecimento da relação com investigadores e corpo académico em geral, posiciona a Biblioteca num patamar de reconhecimento do seu valor, expresso no apoio à investigação e comunicação científicas (Amante, 2014; Akers *et al.*, 2016).

Mas houve alterações de cenário noutras áreas, igualmente importantes: pedagogia, bolsa de estudos e tecnologias, expectativas econômicas e responsabilidades sociais. Não existe uma estratégia comum para os bibliotecários apoiarem a gestão de dados de pesquisa no ensino superior, mas Nitecki & Davis (2017) acentuam os valores profissionais e fortes compromissos para melhorar a sociedade que levaram os bibliotecários a abraçar a gestão de dados de investigação, apesar de ser um território desconhecido para a maioria. O ceticismo recente sobre o valor da pesquisa na opinião pública, reforça o papel crítico dos profissionais da informação para garantir o acesso a informações oficiais e preservar a produção de pesquisa e o registo

cultural. Entende-se deste modo que os profissionais de informação no ensino superior têm a oportunidade de alargar o seu escopo para além da mera gestão de recursos, contribuindo e tendo uma voz ativa em estratégias inspiradoras para o bem social, ao promover colaborações com movimentos como a ciência aberta (Nitecki & Davis, 2017).

Surgem amiúde estudos de caso que possibilitam ter uma visão pragmática e fundamentada, pela observação das práticas e iniciativas realizadas por Bibliotecas, em diferentes países. É fundamental uma visão atenta dos exemplos de outras instituições, mas a uniformização de competências *core* deve, no entanto, ter em atenção a especificidade de cada instituição. As competências *core* necessárias às instituições diferem, dependendo dos tipos e características das instituições. "In the absence of an understanding in the profession of the data librarian role, library policy and strategic planning might not accurately reflect the experience of 'frontline' data librarians" (Federer, 2018).

Os serviços de dados de pesquisa são definidos como serviços que abordam todo o ciclo de vida dos dados, incluindo PGD, curadoria digital (seleção, preservação, manutenção e armazenamento) e criação e conversão de metadados. Apesar de inúmeras discussões sobre os possíveis papéis do profissional de informação face aos serviços de pesquisa de dados (Council on Library and Information Resources, 2008; Association of Research Libraries, 2006; Hey and Hey 2006; Gold, 2007), Tenopir *et al.* (2013) apontavam para a o facto de não ser contemplada a preparação individual do profissional. Concluiram "[...]An implication is that library based RDS are important opportunities for increased alignment between library services and the university research mission."

Pretende-se que os bibliotecários assumam uma postura ativa perante as comunidade e organizações que servem, redefinindo-se o perfil tradicional e o reconhecimento profissional de novas competências (Borges & Casado, 2017). A complexidade e as linguagens de programação, associadas à prática de análise de dados deverão aliar-se às competências de gestão dos bibliotecários, na premissa da preservação e a avaliação de dados apoiadas na tecnologia.

Perante um cenário heterogéneo de teorias, opiniões, recomendações e observações, que estratégias afinal devem ser estabelecidas e delineadas para as Bibliotecas e os seus profissionais para o cumprimento da sua missão secular, ao serviço do conhecimento humano? Deverão munir-se de competências generalistas, fornecendo serviço de dados em várias áreas académicas, transversalmente, numa interação transdisciplinar com o corpo académico ou

investirem na especialização em determinada área, servindo o processo científico com um conhecimento aprofundado e especializado de determinada área, ficando ao serviço apenas de utilizadores daquela área?

A literatura fornece pistas e reflexões sobre quais as competências *core* adequadas aos profissionais de informação no contexto da eScience (Federer, 2018; Tenopir *et al.*, 2013; Yoon & Schultz, 2017; Nitecki & Davis, 2017), apontando algumas práticas exigíveis ao novo perfil dos bibliotecários (Kenan, 2016; Cox et al., 2012; Semeler et al., 2019), nomeadamente: adquirirem conhecimentos de competências interpessoais e comportamentais; competências na comunicação escrita formal e redação de documentação técnica e estudos de caso; capacidade de adaptação e proatividade na busca de atualização profissional; conhecimento do ecossistema da organização onde se inserem, aumentando a sua compreensão do que os dados de pesquisa significam para os investigadores e o ponto de vista destes em relação à gestão de dados de investigação; conhecimento das políticas de agências de financiamento; conhecimento específico sobre o uso de dados (tipos de dados, padrões e esquemas de metadados, diplomas legais e regulamentares, preservação de dados); conhecerem a tecnologia da informação, lógica de programação dos computadores e suas linguagens (Python, Structured Query Language, Java e eXtensible Markup Language); conhecerem o funcionamento de softwares científicos usados para transformar dados; entenderem os fundamentos das ferramentas de recuperação de informações; conhecerem o design e estruturas de bancos de dados; entenderem a base dos estudos métricos sobre a informação; design centrado no usuário, ferramentas de processamento de linguagem natural, Internet das Coisas, e grandes dados.

Não se exige que o profissional da informação se torne um cientista de dados em toda a sua plenitude, obviamente, mas é imprescindível munir-se dos conhecimentos base nestas áreas. Deste modo, Semeler *et al*. (2019) sugerem as competências que os bibliotecários devem procurar ter, tradicionalmente atribuídas aos cientistas de dados, mantendo o foco em gerar e gerir novos serviços de informação, baseados na coleta e análise de dados. Falamos de avaliação (inclusive o seu valor económico) e retenção de dados, advocacia, promoção, marketing, conscientização, coordenação de práticas entre unidades e instituições, habilidades de negociação e habilidades de gestão de reclamações e expectativas. O papel do bibliotecário entende-se então, como facilitador em todas as etapas da investigação científica e participação nos processos de tomada de decisão.

3.2 BOAS PRÁTICAS EM OBSERVAÇÃO

Reportando-se ao ano de 2014, Cox *et al.* (2017) observaram a popularidade, em todos os países, de conferências e workshops. Os webinars mostraram-se mais populares na Nova Zelândia, Austrália e Canadá. A colaboração com programas académicos apresentou a menor aceitação, com uma média de apenas 25% dos inquiridos a destacar esta opção; foi particularmente baixo no Reino Unido. Propuseram uma estratégia para a aprendizagem e atualização dos profissionais em serviços de dados na forma de cursos de curta duração, com base no trabalho e práticas da instituição onde se inserem (Cox *et al.*, 2017).

Iniciativas como DigCurv[23] exploraram a disponibilização de formação profissional para curadores digitais nos setores de bibliotecas, arquivos, museus e património cultural, necessária para desenvolver novas competências, essenciais para a gestão a longo prazo das coleções digitais (Cox *et al.*, 2017).

O crédito e reconhecimento são incentivos sempre dignos de nota. Estimulam o meio académico e tomam a forma de "marketing" na sua divulgação. São formas de, ainda que indiretamente, impulsionar o movimento de literacia de dados de investigação. São exemplos disso os prémios e bolsas da "Association of College and Research Libraries" (ACRL) apresentados anualmente e possíveis graças ao generoso apoio corporativo, que permite que a ACRL honre o que há de melhor em biblioteconomia académica e de pesquisa. (DFREE, 2020)

As bibliotecas geram novos modelos de comunicação académica, tendo uma forte influência na disseminação do conhecimento. A sua estrutura institucional e competências de curadoria, aliadas a novas competências centradas nos dados, configuram-se como uma infraestrutura de conhecimento que viabiliza a manutenção de grandes conjuntos de dados, a disponibilização de diferentes versões para diferentes propósitos, criação de links de metadados, informações de proveniência, preservação e armazenamento a longo prazo e todas as tarefas associadas à curadoria que garantem a acessibilidade. A interoperabilidade no contexto internacional é essencial para que os repositórios institucionais possam crescer. Os dados de pesquisa podem ser categorizados, classificados e ter um valor agregado. Distinguir o

[23] Digital Curator Vocational Education Europe Project- DigCurV, was a project funded by the European Commission's Leonardo da Vinci programme to establish a curriculum framework for vocational training in digital curation launched today. [...] DigCurV brought together organisations from Europe, Canada and the USA with a strong track record of international work in the field of digital libraries and digital preservation. *in* DigCurV – Digital Curator Vocational Education Europe.

tipo de dados é crucial para os tomadores de decisão na preparação, armazenamento e preservação de dados, para acesso futuro. Grifin (2013) sugere algumas medidas: (i) preservar a documentação relativa ao conteúdo, estrutura, contexto e proveniência (exemplo de parâmetros experimentais e condições) das coleções de dados – ou seja metadados(registo de tudo o que pode interessar a outro investigador); (ii) preservar modelos de dados e software específico dos dados computacionais; (iii) descrever o hardware e especificações de instrumentação dos dados observacionais e de laboratório; (iv) preservar e armazenar dados em múltiplas formas para uma maior interoperabilidade com sistemas e redes; (v) alargar o espectro de acessibilidade ao colocar em repositórios ligados a conjuntos de dados diversos e heterogéneos; (vi) tratar os dados de investigação no início do seu ciclo de vida, para criar coleções altamente funcionais que cresçam naturalmente e se liguem a outros recursos (Griffin, 2013).

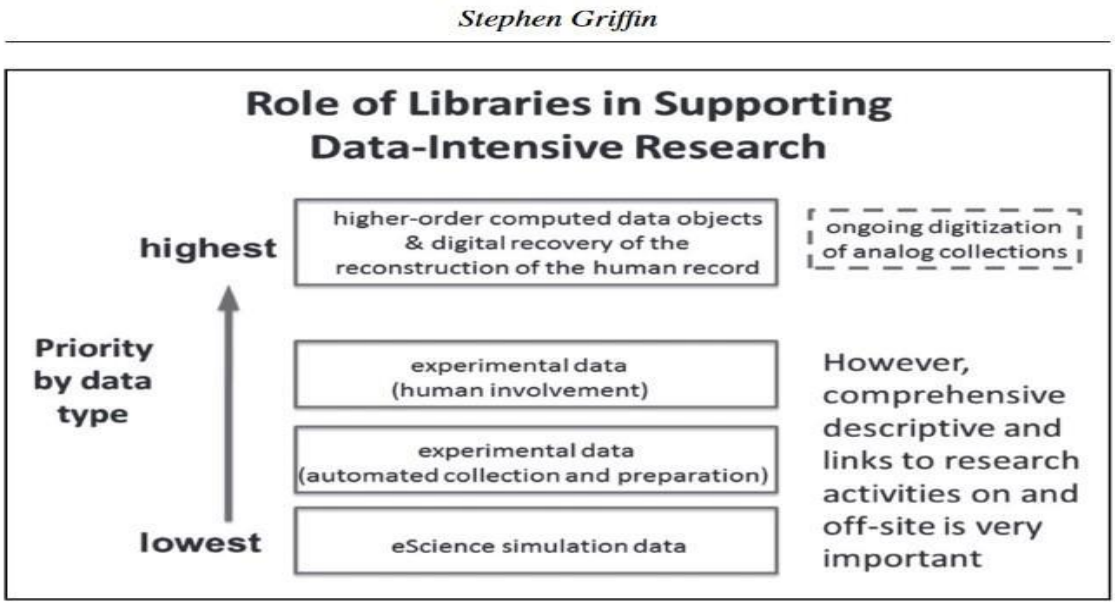


Figura 9 O papel dos bibliotecários na investigação intensiva de dados in Griffin (2013)

É necessária uma atitude proativa dos profissionais em identificar novas oportunidades de intervenção e colaboração, seja na implementação de sessões de formação e desenvolvimento de novos serviços. Para que seja possível a gestão de dados ao longo do seu ciclo de vida, os investigadores e profissionais da informação deverão ter competências que lhes permitam organizar os resultados da pesquisa de forma a lidar com o grande volume de dados gerados e determinar quais os conjuntos de dados que precisam ser preservados; Aplicar práticas eficazes para descrever, organizar, descobrir e aceder a conjuntos de dados (diferentes

da tradicional curadoria de recursos de informação bibliográfica); Permitir a comunicação e práticas colaborativas entre todos os envolvidos na geração dos dados; Obedecer às diretrizes do governo e de outras agências de financiamento que exigem PGD (Nitecki & Davis, 2017).

A literatura sugere que as bibliotecas sejam intervenientes em todas as fases do ciclo de vida da investigação, na obtenção de financiamento, planeamento dos projetos de investigação, na formação de investigadores, na integração dos profissionais de informação nas equipas de investigadores e na análise do impacto das publicações. No entanto, Revez (2019) não encontra uma concretização empírica dessas sugestões no caso português. Sugere, portanto, a recomendação dos autores revistos, para que se estabeleça um maior diálogo entre investigadores e profissionais da informação, sublinhando que, face a algum desconhecimento por parte dos investigadores no que toca ao apoio que estes profissionais podem dar, é importante que estes últimos apresentem e se auto proponham, com as competências técnicas necessárias a um apoio robusto e fiável. É necessário construir pontes de confiança. Numa perspetiva mais recente, Amante & Inácio (2021) sugerem que em Portugal:

> Os profissionais da informação, sobretudo no contexto das BES (Bibliotecas do Ensino Superior), parecem estar particularmente aptos para a utilização e exploração de ferramentas de gestão de ciência, uma vez que possuem conhecimentos, competências e aptidões particulares no âmbito da promoção do acesso à informação que se agrupam em quatro dimensões essenciais: a publicação e aconselhamento académico; a pesquisa e recuperação de informação e compreensão do trabalho académico; a formação; e a curadoria da produção científica. (Amante & Inácio, 2021)

Bryant *et al.* (2017) propõem três funções principais para a biblioteca: aumento da conscientização do valor dos dados entre os investigadores; prestação de serviços de armazenamento e preservação de dados dentro da instituição, por meio de repositórios institucionais; e desenvolvimento de uma nova vertente profissional na forma de biblioteconomia de dados. Para além destas funções revela-se extremamente importante o nível de interação entre bibliotecários e investigadores/ docentes, binómio explorado há muito, em extensa literatura (Amante, 2012). Sublinhamos um dos três níveis (*networking*; coordenação; colaboração) de interação mais benéfico para ambos os grupos, presentes na proposta de Raspa&Ward (2000): a colaboração. Os níveis propostos assentam na perspetiva da duração e intensidade da interação, distribuição das tarefas e a partilha de objetivos comuns:

> Librarians are particularly suited for **collaborative** enterprise. They practice daily the kind of listening that requires them to translate for students and instructional faculty the questions they bring to the library. Librarians are on the edges of research, on the first threshold where the researcher has an idea or a hunch about something but needs a guide to navigate the waters of inquiry. Librarians know issues change shape, even vanish, depending on where the inquirer situates him- or herself in the search process. (Raspa & Ward, 2000)

Frequentemente, o trabalho do curador de dados não é reconhecido nas publicações e isso precisa ser abordado. O cientista de dados ou curador de dados desempenha um papel fundamental no processo de publicação académica e merece ser reconhecido e recompensado adequadamente. Há também espaço para uma maior convergência de conhecimento de assunto com habilidades de biblioteca e armazenamento, como parte do desenvolvimento e treinamento profissional (Lyon, 2007). Federer (2018) propõe uma taxonomia de competências e conhecimentos baseando-se nas listas existentes da biblioteca, adicionando algumas relacionadas com dados.

"The new knowledge ecologies will necessarily involve transformations of the research process: traditional institutions will adapt or die; new forms will come into being." (Edwards, 2013).

O apoio e a participação na investigação são decisivos para os investigadores e para as organizações que fazem ciência. É importante a observação das práticas e experiências já desenvolvidas, mas "almeja-se o discernimento da influência que as bibliotecas exercem sobre o processo científico" (Borges & Casado, 2017).

Nos processos de gestão de dados de investigação é essencial a colaboração entre profissionais de informação e investigadores. Se por um lado os profissionais de informação têm competências na produção de descrições abrangentes dos seus conjuntos de dados, por outro, falta-lhes a especificidade do domínio científico a tratar que, só o investigador poderá colmatar. Também se passa o inverso: o investigador não tem as competências, nem o tempo para assimilar e concretizar o tratamento dos dados das suas investigações (Swan & Brown, 2008). Urge, portanto, uma maior cooperação entre estes atores. Não esqueçamos os conjuntos de dados de cauda longa perdidos por não serem administrados de modo a serem preservados adequadamente, pesquisáveis e reutilizáveis por outros. São, por excelência, os serviços de apoio da biblioteca (munidos de competências a este nível) a investigadores, pela adoção de padrões e boas práticas de gestão de dados ao longo do ciclo de vida que viabilizam uma cauda longa com "vida útil".

Na prática, vamos tendo o feedback empírico da colaboração entre investigadores e bibliotecários. Na opinião de Mariëtte van Selm[24](2020), bibliotecária na Universidade de Amsterdão, com experiência de 8 anos de apoio aos académicos na gestão de dados de investigação, " Libraries are – or should be – the spider in the web of all data support" e acrescenta:

> Yes, researchers who want to become a little more self-sufficient in handling their data will have to learn some things. But no researcher has to know or be everything. Every conversation between a researcher and a data librarian will make the both of them a little wiser: the librarian learns more about the research that is going on at their institution, the researcher learns what to take into account and why. And all it costs is a little time. (van Selm, 2020)

Sublinha como é pouco realista e subentende-se, injusto, pretender que um investigador tenha, em simultâneo, competências tão variadas como: especialista na sua própria área, na segurança da informação, jurista, bibliotecário, arquivista e clarividente, que seja em suma, "several *personas* in one". Segundo este relatório, os investigadores ainda não consideram as bibliotecas como a sua primeira escolha na ajuda com o universo dos dados, como podemos observar na Figura 10.

[24] Mariëtte van Selm is a historian, holding a PhD in Theology, who worked at several research institutes in the Netherlands before starting at the Library of the University of Amsterdam (UvA) and Amsterdam University of Applied Sciences (AUAS) ten years ago. She developed and now coordinates research data support at the Library, is manager of the UvA/AUAS Research Data Management Programme and is a member of the Advisory Board of the National Coordination Point Research Data Management (LCRDM) in the Netherlands. As of this year, she's also a member of the ResearchIT team at UvA/AUAS' IT department. She holds the questionable honour of being the first UvA employee to register an ORCID ID. https://orcid.org/0000-0003-3711-4282 *in* (Science *et al.*, 2020)

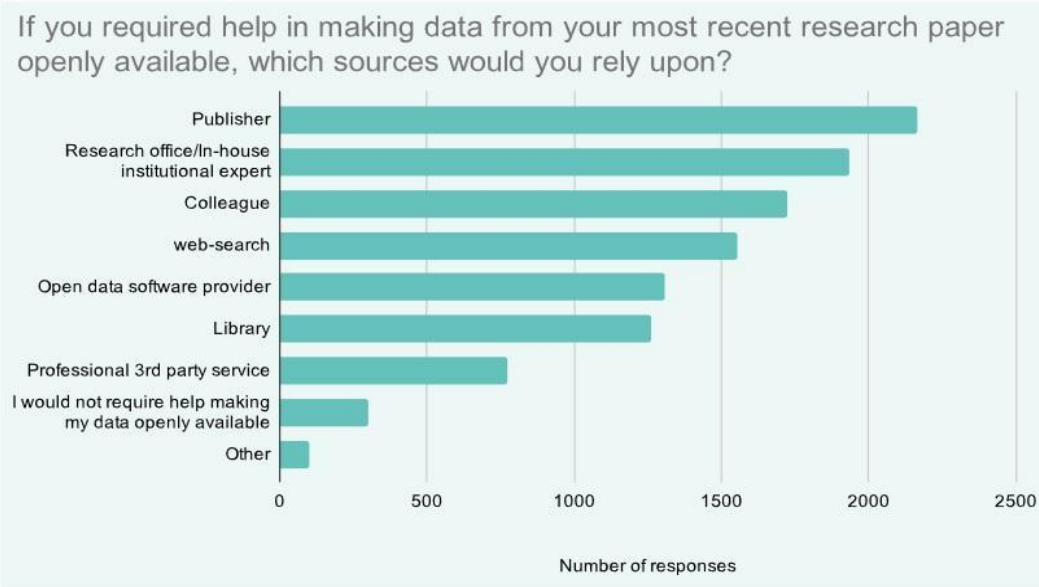


Figura 10 A quem os investigadores pedem ajuda para lidar com os dados? (van Selm, 2020)

A literatura também aponta a necessidade de combinar diferentes tipos de dados estatísticos para medir a eficiência do trabalho feito nas bibliotecas e o possível impacto que eles podem ter nas suas organizações. Também visa utilizar o conceito de *data warehousing* como ferramenta para unir diferentes tipos de dados estatísticos na análise. (Laitinen & Saarti, 2012).

Os relatórios sobre o desempenho das bibliotecas, como o apresentado por Lyon (2017), são essenciais para explorar as funções, direitos, responsabilidades e relacionamentos de instituições, centros de dados e outros *stakeholders* importantes que trabalham com dados.

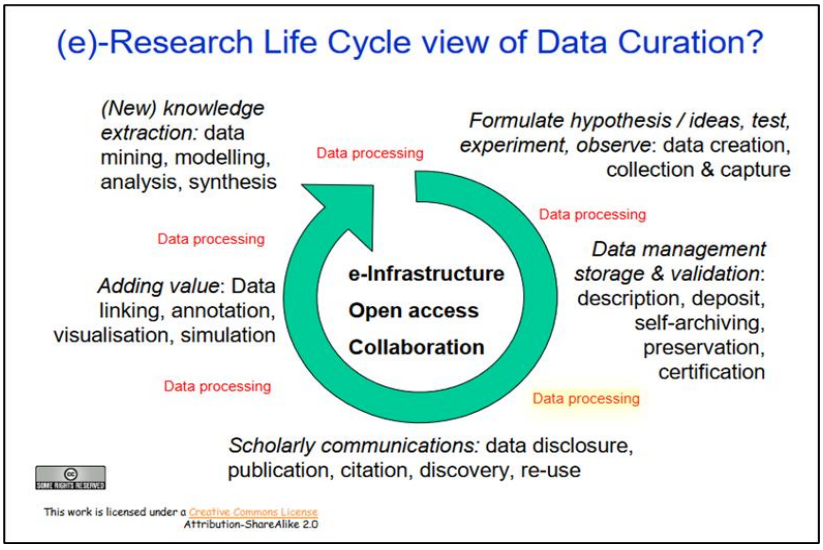


Figura 11 e-Research Life Cycle and data curation in (Lyon, 2007)

É necessária uma coordenação estratégica entre domínios, nos níveis mais altos das organizações de financiamento de pesquisa, para implementar a infraestrutura e os serviços para gerir, eficazmente, o dilúvio de dados crescente. Deve haver um consenso sobre os padrões de dados da comunidade. É necessário identificar se há o uso de identificadores persistentes para dados, controlo de versão de conjuntos de dados complexos e a derivação de modelos para anotação de conjuntos de dados.

No aspeto humano das atividades de curadoria de dados, muitos investigadores parecem desconhecer a gama de questões associadas às melhores práticas de gestão, o que reflete a necessidade crescente de programas coordenados de ensino, treino e competências para equipar a comunidade de pesquisa. (Lyon, 2007).

> A criação de metadados e de dados ligados (LOD), as questões da interoperabilidade, a prestação de novos serviços, a colaboração em serviços de descoberta, a participação em projetos de publicação, incluindo revistas em Acesso Aberto, a gestão de repositórios institucionais, a preservação ou, de um modo mais amplo, a curadoria dos dados, entendida como a gestão ativa da preservação ao longo do ciclo de vida do objeto, são apenas algumas das preocupações assumidas pela bibliotecas, particularmente as universitárias, na contemporaneidade. (Borges, 2017)

A colaboração requer competência de escuta, sempre atento às diferenças subtis de suposições, teorias, definições e métodos. As lições e competências aprendidas com essas parcerias podem melhorar a erudição de todos os participantes. A curadoria de dados agregando valor por meio de documentação, padronização, migração para novos formatos – é essencial para o uso e reutilização de dados a longo prazo. A política pública de planos de gestão e partilha de dados, por sua vez, depende do devido cuidado e curadoria dos dados da pesquisa científica e tecnológica. Assim, estudos de dados e práticas de dados têm implicações para a política social, bem como para o trabalho cooperativo (Borgman *et al.*, 2012).

Plataformas de tecnologia comuns também são importantes para alcançar interoperabilidade e sustentabilidade, e podem ser aproveitadas como investimentos em projetos. Mais especificamente: (i) Na pesquisa básica, a conceção de projetos de última geração, equipamentos, recolha de dados, bancos de dados e ferramentas analíticas estão

sempre em revisão; (ii) Equipamentos estáveis, dados e ferramentas analíticas são usados para calibrar fundos estáveis contra os quais novos tipos de dados podem começar a ser reconhecidos e o longo processo de avaliação de dados pode prosseguir; (iii) As comunidades de pesquisa básica sabem como construir conhecimento no contexto dessa mistura de estabilidade e instabilidade em projeto, equipamento, dados e análise e desenvolvem diferentes abordagens para responder a perguntas compartilhadas. A curadoria de dados para pesquisa básica deve acomodar esse trabalho diário contínuo com estabilidade e instabilidade, inclusive em conjuntos de dados de grande escala. Para fazer isso, precisamos entender as práticas básicas de pesquisa com dados (Wallis *et al.*, 2012).

Todas as práticas de investigação passam, atualmente, pelo mundo digital. Para uma prossecução de todo o processo de pesquisa há que disponibilizar recursos e estes custam dinheiro: computadores, software, equipa e conteúdo. É necessário um investimento de fundo em plataformas e padrões técnicos comuns. As instituições de investigação e ensino superior devem elaborar estratégias com vista ao cumprimento dos requisitos dos órgãos financiadores da ciência, seguindo três etapas fundamentais: 1) compreender a sua posição atual; 2) definir onde se quer estar no futuro; 3) traçar um programa de atividades para fazer essa transição.

As bibliotecas de instituições de investigação e ensino superior são parceiros estratégicos que devem ter um papel ativo na definição de políticas e estratégias para a gestão de dados, na dinamização de sistemas para repositórios de dados e serviços de apoio ao ciclo de vida de dados, com particular enfoque nos planos de gestão de dados e na documentação de conjuntos de dados, que podem assumir a coordenação da agenda que deverá expandir e fortalecer o papel institucional na gestão dos dados de investigação (Príncipe & Silva, 2018).

Na conceção de uma estratégia integrada para garantir sistemas e serviços de suporte à gestão de dados científicos nas instituições de investigação e ensino superior desenvolvidos de forma coerente, devem seguir-se três etapas fundamentais: 1) compreender a sua posição atual; 2) definir onde se quer estar no futuro; 3) traçar um programa de atividades para fazer essa transição. (Príncipe & Silva, 2018)

CONCLUSÃO

O acesso aberto e a ciência aberta constituem o cenário em que se movimentam os *stakeholders* da investigação científica atualmente. O aumento exponencial de dados, (*Big Data, Data Deluge, Long Tail*) caracterizado pelos 3V's (volume, variedade e velocidade), marcou a paisagem, não só no modo de comunicar a ciência, mas em todo o processo de investigação, sem possibilidade de retorno. A base tecnológica atual derrubou barreiras geográficas, temporais, políticas e sociais na comunicação da ciência. Ergueu, no entanto, barreiras de outra natureza, expressas na necessidade de criar infraestruturas tecnológicas e humanas, para lidar com o fenómeno emergente. Falamos dos dados, nas suas mais diversas concretizações, que surgem como o elemento basilar de atividades nas várias áreas da sociedade civil atual, e com uma pertinência ainda mais marcante no processo científico.

Numa incursão inicial sobre o tema, percebemos a importância de uma contextualização histórica dos conceitos correlacionados com dados e com a sua gestão nas comunidades científicas. A sistematização da evolução conceptual na Ciência da Informação na sua ligação intrínseca com a investigação científica, pela lente de autores consagrados na área, permitiu um entendimento mais consolidado do fenómeno do dilúvio de dados e o "efeito borboleta" a nível global.

A publicação como face visível da investigação científica, e a sua repercussão cumpre as aspirações de visibilidade e certificação entre pares, já presentes nos cientistas do século XVII, patentes no exemplo da formação da *Royal Society of London*. Atualmente, o processo de suporte que coadjuva todas as etapas de uma investigação, antes, durante e à *posteriori*, complexificou-se num registo paralelo à evolução tecnológica, mas nem sempre à mesma velocidade. Com o advento da evolução tecnológica e inovação nos processos de instrumentação e computação, capazes de gerar quantidades de dados gigantescas, os investigadores correm o risco de descurar a preservação dos seus dados de investigação, por falta de uma planificação na sua gestão. Os profissionais de informação têm um papel essencial neste processo, pelas competências de curadoria e pesquisa inerentes à profissão. A diferença face ao passado "tradicional" das Bibliotecas está no elemento tecnológico. As competências tradicionais dos bibliotecários não foram eliminadas neste novo paradigma, mas reinventaram-se e parecem começar a afirmar-se como um processo mais natural e intuitivo, ainda que haja muito trabalho de campo a fazer, incluindo formação em constante atualização.

Foi neste cenário que a presente dissertação foi trabalhada, espelhando as teorias e práticas em curso mais relevantes na literatura científica sobre a temática, numa perspetiva temporal evolutiva, com o propósito de lançar novas pistas na prossecução e estabelecimento normativo de práticas normativas de competências *core* na gestão de dados de investigação, adequadas às exigências de cada e no fortalecimento da cooperação entre investigadores e profissionais de informação a bem da ciência.

Os paradigmas da ciência foram evoluindo da sua faceta empírica, teórica e computacional até à exploração de dados capturados ou gerados pela simulação, vinculado ao 4º paradigma/ ciência intensiva de dados.

A dissertação teve como objetivo geral identificar o perfil necessário ao profissional da informação para se afirmar no contexto do universo da Ciência Aberta e dos dados de investigação. Para o cumprimento deste objetivo identificaram-se, como objetivos específicos, o mapeamento da literatura internacional que incide quer nesta matéria quer na identificação das ações levadas a cabo nas bibliotecas e dos atores que têm desenvolvido ações de literacia de dados.

Começámos por compreender e contextualizar a derivação do 4º paradigma, explorando o percurso do seu desenvolvimento e trajetória, num estudo exploratório partindo de uma visão geral para o particular, abordando, discutindo os conceitos envolvidos em cada um dos capítulos. Assim, sobre o 4º paradigma da ciência e a ciência intensiva de dados que revolucionaram a forma de fazer e comunicar ciência, aprofundámos os conceitos de: **i)** dilúvio de dados e Big Data; **ii)** exemplos de ferramentas úteis, como a mineração de dados, para lidar com os 3V's dos dados de investigação; **iii)** e, por fim, tentámos demonstrar a importância e relevância no progresso da ciência, da chamada Cauda Longa, enquanto parente "menor" no universo científico, percebendo que tem tanta ou mais importância, do que projetos de grande amplitude e visibilidade. No segundo capítulo passámos a uma fundamentação teórica, que julgamos necessária, da (in)definição de dados de investigação e propostas para uma taxonomia. Necessária, na medida em que só com um entendimento lato e bem consolidado do que são dados de investigação, nas suas diversas circunstâncias, permite uma seleção, preservação, manutenção e armazenamento adequado. Dentro deste capítulo percorremos os princípios a respeitar no tratamento dos dados de investigação, fruto, por um lado, dos requisitos impostos pelas agências de financiamento da investigação científica e por outro, pela observação dos benefícios da aplicação dessas premissas, que incluem: **i)** Boas práticas no uso,

reutilização e partilha dos dados de investigação, os conceitos que lhes são inerentes e as dificuldades que ainda impedem uma total concretização destes princípios basilares da partilha na ciência; **ii)** Uma breve contextualização histórica dos princípios FAIR, como e porque surgiram, o seu impacto no tratamento e classificação dos dados e a exigência pelas agências de financiamento do seu cumprimento na elaboração do Plano Geral de Dados. **iii)** Definição de Plano geral de Dados, os seus pressupostos e mapeamento da sua aplicação a nível internacional e nacional, expresso com exemplos de algumas instituições de referência. **iv)** Por fim, dentro deste ponto terminamos com a contextualização do surgimento do conceito de metadados, os "dados dentro dos dados" e da importância da informação que transportam anexada aos dados de primeira linha.

Finda esta exposição dos principais conceitos e movimentos apurados na revisão da literatura, avançámos para o último ponto, que reúne as condições de concretização do objetivo geral: delinear o perfil dos profissionais de informação. Num cenário já com alguma estrutura conceptual e testemunhos da sua faceta empírica, com abertura para as divergências e especificidades de autores, atores e instituições, que abre uma discussão pública e salutar, pareceu-nos estar em posição de avançar para as competências *core*, tão necessárias na investigação científica, resultando em novas perspetivas de atuação, por parte dos investigadores face aos profissionais de informação, essenciais, como já demonstrado, no progresso da ciência. Este capítulo apresenta algumas iniciativas em literacia de dados de investigação, numa mostra de exemplos concretos no panorama internacional e nacional, destacando algumas das suas especificidades. Estas iniciativas são essenciais à investigação e podem ser uma incumbência natural, dos profissionais de informação. O apoio tradicional dado pelo bibliotecário na pesquisa e recolha de informação, passa agora para uma vertente digital e tecnológica, de todos os elementos referidos anteriormente no tratamento dos dados, que lhes estão associados**. i)** Para a concretização de iniciativas robustas de literacia em dados de investigação, os profissionais que julgamos mais habilitados para o efeito e cuja missão profissional se adequa, indubitavelmente, a ações de formação no meio académico (entre outras) são os bibliotecários, comummente chamados de profissionais de informação, pelo alargamento do escopo da sua atuação, aos dias de hoje. Apresentámos os pontos chave de competências a adquirir e as correntes mais significativas validadas pela revisão da literatura, com propostas e alguns exemplos, das competências *core* para os profissionais de informação. **ii)** Formem-se os profissionais de informação nestas novas competências e veremos o fluxo dinâmico de competências a fluir na instituição, com todos os benefícios daí decorrentes, numa

relação calibrada pela cooperação e entendimento dos atores envolvidos. Terminamos este ponto com os novos rumos possíveis que se abrem no horizonte da ciência, para os profissionais da informação e das Bibliotecas de Investigação. **iii)** Nesse sentido e sempre apoiados na narrativa da revisão da literatura, conseguimos reunir um conjunto de sugestões, propostas e reflexões que visam a evolução nas incumbências dos profissionais da informação enquanto elementos integrantes da investigação, a nova realidade das Bibliotecas, necessariamente ligadas ao mundo digital, tendo como pano de fundo os dados de investigação, em todo o seu ciclo de vida e enriquecendo e facilitando o processo de comunicação e disseminação da ciência.

Finda a descrição dos pontos chave da presente dissertação chegámos às seguintes conclusões:

1) O 4º paradigma configurou, sem margem para dúvida, uma nova conceção de fazer ciência, alterando as visões e as práticas na sua comunicação e disseminação. O dilúvio de dados emergente da evolução tecnológica e de computação se por um lado levantou problemas, pela rapidez e volume dos dados daí resultantes, por outro abriu novos horizontes na ciência. Entre eles destacamos a disseminação *on line* da literatura científica e a interoperabilidade daí decorrente; a criação de infraestruturas de informação comuns, comunicação e publicação; a convergência semântica de ferramentas de dados, cruzando fronteiras disciplinares e epistemológicas; a estimulação da colaboração entre pares e outros atores presentes no processo de investigação, como os profissionais da informação.

A literatura comprova uma passada mais lenta, na norma vigente da eScience, nas ciências não naturais referindo, no entanto, algumas evoluções no sentido de aproximar as práticas nas diferentes culturas epistémicas.

O conceito de *Big Data* veio para ficar e de acordo com a literatura continuará a crescer. A preocupação de não ser possível acompanhar a produção de dados, de não ter uma evolução nas infraestruturas suficientemente robusta e bem oleada, ainda se mantém e não parece ser um assunto resolvido. Com agrado observamos que há ferramentas e conceitos de outras áreas disciplinares (computação, economia, etc.) que dão força à noção de interdisciplinaridade e cooperação.

Sublinhamos a Cauda Longa pela expressão que dá, finalmente, aos projetos de investigação de menor envergadura e aos seus investigadores ou pequenas equipas que, não

obstante, são preciosos no progresso científico, permitindo que esses dados sejam disponibilizados para uso futuro ao invés de ficarem eternamente perdidos, sem qualquer rentabilização.

2) Apesar das muitas divergências e *nuances* na definição de dados de investigação, a literatura sugere uma confluência nos pontos essenciais, sendo um tema ainda muito debatido, mas já suficientemente uniformizado para criar padrões normativos e suficientemente amplo para abranger todas as realidades, inclusive as que não têm reflexo na realidade física, como os dados do ciberespaço. Verificou-se um consenso cada vez mais generalizado, sobre o uso, reutilização e partilha dos dados. Aferiu-se dos seus benefícios, já comprovados, e parece uma probabilidade concretizável que, mais cedo ou mais tarde, o mundo se torne um espaço de partilha global natural e inequívoca, fiel aos princípios FAIR, base de apoio preciosa na elaboração dos PGD, fundamentais na organização do trabalho científico. Os avanços relatados pelas instituições que investiram nestas boas práticas são por demais evidentes. O prestígio individual e coletivo, das instituições e seus profissionais é amplamente aumentada. As mentalidades são sempre algo muito difícil de mudar. Daí que os requisitos das agências de financiamento para a criação de PGD, cumprindo com os princípios FAIR sejam um marco muito positivo. Assim como as políticas estabelecidas por instituições governamentais que reforçam a mudança daqueles que se mantêm reféns de "outros tempos". A Ciência Aberta cobre todos estes elementos numa união profícua, rumo ao conhecimento humano, que afeta todo o tecido social, sem exceção.

3) É um fato que há muito para fazer. Nas bibliotecas mais expressivas no panorama internacional já podemos conferir estas propostas a serem colocadas em práticas com resultados muito satisfatórios. Em Portugal parece haver ainda um longo processo. Os investigadores ainda não encaram o profissional de informação como um elemento integrante das suas investigações. A literatura refere alguns exemplos a nível internacional de comportamentos semelhantes. Com base nestes estudos ficam várias recomendações que podem e devem ser parte das estratégias governamentais e das instituições do ensino superior. São necessárias mais iniciativas em literacia dos dados de investigação, dadas pelas instituições aos seus profissionais de informação, para que estes possam cumprir o papel formativo para o corpo académico e finalmente, serem reconhecidos no seu papel de suporte e colaboração, e parte da tomada de decisões. Descurar, e o termo não é inocente, a competência de curadoria destes profissionais, é assumir o risco de perda dos dados por erros técnicos de seleção, preservação, manutenção e

armazenamento dos dados. Para além de retirar tempo e disponibilidade ao investigador para a sua investigação.

Pensamos que a presente dissertação contribui para uma visão global das práticas no contexto da Ciência Aberta, consubstanciada na revisão da literatura, crucial nas preocupações atuais do processo científico, com foco no papel das Bibliotecas e profissionais de informação como elementos ativos e intervenientes no processo de tomada de decisões.

Como pista para investigação futura pensamos que seria útil apurar a situação em Portugal através de uma abordagem empírica, aplicando outras formas de recolha de informação, como por exemplo um inquérito em forma de questionário ao Grupo de Trabalho de Bibliotecas do Ensino Superior, entrevista a um/a elemento representativo/a de uma entidade reconhecida por ações de literacia em dados de investigação no ensino superior, e um mapeamento da aplicação de PGD e das práticas já instituídas nas Bibliotecas do Ensino Superior em território nacional.

REFERÊNCIAS BIBLIOGRÁFICAS

Akers, K. G., Sarkozy, A., Wu, W., & Slyman, A. (2016). ORCID Author Identifiers: A Primer for Librarians. Medical Reference Services Quarterly, 35(2), 135–144. https://doi.org/10.1080/02763869.2016.1152139

Amante, M. J. (2012). Relações entre bibliotecários e docentes no Ensino Superior: 62.

Amante, M. J. (2014). The librarian and knowledge manager: The case of repositories. 12.

Amante, M. J., & Inácio, A. (2021). Profissionais de informação para as bibliotecas do século XXI: Desafios para a gestão da informação científica e Ciência Aberta. Sob a lente da ciência aberta: olhares de Portugal, Espanha e Brasil, 2021, ISBN 978-989-26-2022-0, págs. 221-250, 221–250. https://dialnet.unirioja.es/servlet/articulo?codigo=7805462

Antell, K., Foote, J. B., Turner, J., & Shults, B. (2014). Dealing with Data: Science Librarians' Participation in Data Management at Association of Research Libraries Institutions. College & Research Libraries, 75(4), 557–574. https://doi.org/10.5860/crl.75.4.557

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P., & Wouters, P. (2004). An International Framework to Promote Access to Data. Science (New York, N.Y.), 303, 1777–1778. https://doi.org/10.1126/science.1095958

Baca, M., Gilliland, A. J., Gill, T., Woodley, M. S., & Whalen, M. (2016, julho 20). Introduction to Metadata [InteractiveResource]. Getty Research Institute, Los Angeles. http://www.getty.edu/publications/intrometadata

Banning, E. B. (2020). What Are Data? Measurements and Errors. Interdisciplinary Contributions to Archaeology, 5–16. Scopus. https://doi.org/10.1007/978-3-030-47992-3_1

Bates, M. J. (2005). Information and Knowledge: An Evolutionary Framework for Information Science. Information Research: An International Electronic Journal, 10(4). https://eric.ed.gov/?id=EJ1082014

Bates, M. J., & Maack, M. N. (Eds.). (2009). Information. Em Encyclopedia of Library and Information Sciences, Third Edition (0 ed., pp. 2347–2360). CRC Press. https://doi.org/10.1081/E-ELIS3-120045519

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM Computing Surveys, 41(3), 16:1-16:52. https://doi.org/10.1145/1541880.1541883

Batini, C., & Scannapieco, M. (2006). Data Quality: Concepts, Methodologies and Techniques. https://doi.org/10.1007/3-540-33173-5

Bell, C., Hey, T., & Szalay, A. (2009). Beyond the Data Deluge (Computer Science). Science (New York, N.Y.), 323, 1297–1298. https://doi.org/10.1126/science.1170411

Berger, C. (sem data). Big Data Analytics with Oracle Advanced Analytics In-Database Option. 42.

Bethel, E. W., Greenwald, M., Kleese van Dam, K., Parashar, M., Wild, S., & Wiley, H. S. (2016). Management, Analysis, and Visualization of Experimental and Observational Data—The Convergence of Data and Computing: https://escholarship.org/uc/item/3p80p9bs

Boavida, C., Pais, C., Rodrigues, E., Pereira, F., Borba, F., Cardoso, J., Noro, J., & de Coimbra, U. (2021). Workshop: Políticas e estratégias institucionais para GDI. 64.

Borges, M. M. (2017). Reflexos da Tecnologia Digital no Processo de Comunicação da Ciência. Em Faculdade de Filosofia e Ciências - FFC- Campus de Marília, M. J. V. Jorente, & D. L. Padrón, Una Mirada a La Ciencia de La Información desde Los Nuevos Contextos Paradigmáticos de La Posmodernidad (pp. 179–196). Faculdade de Filosofia e Ciências. https://doi.org/10.36311/2017.978-85-7983-904-7.p179-196

Borges, M. M., & Casado, E. (2021). Sob a lente da Ciência Aberta: Olhares de Portugal, Espanha e Brasil. https://doi.org/10.14195/978-989-26-2022-0

Borges, M. M., & Freitas, M. C. V. de. (2020). Relatos de experiência de Gestão da Informação na Universidade de Coimbra. Componentes curriculares do eixo temático da gestão na pós-graduação em Ciência da Informação, no Brasil, Espanha e Portugal, 215–240. https://estudogeral.sib.uc.pt/handle/10316/93314

Borges, M. M., & Sanz Casado, E. (2009). A ciência da informação criadora do conhecimento Vol. I. Imprensa da Universidade de Coimbra. https://doi.org/10.14195/978-989-26-0319-3

Borgman, C. (2010). Research Data: Who Will Share What, with Whom, When, and Why? SSRN Electronic Journal. https://doi.org/10.2139/ssrn.1714427

Borgman, C. (2011). The Conundrum of Sharing Research Data. Journal of the American Society for Information Science and Technology, 63. https://doi.org/10.2139/ssrn.1869155

Borgman, C., & Furner, J. (2005). Scholarly Communication and Bibliometrics. ARIST, 36, 2–72. https://doi.org/10.1002/aris.1440360102

Borgman, C. L. (2007). Scholarship in the digital age. MIT Press. http://archive.org/details/scholarshipindig00borg

Borgman, C. L. (2010a). The Digital Future is Now: A Call to Action for the Humanities. Digital Humanities Quarterly, 003(4).

Borgman, C. L. (2010b). The Digital Future is Now: What the Humanities can Learn from eScience. https://escholarship.org/uc/item/15q7d9p3

Borgman, C. L. (2018). Open Data, Grey Data, and Stewardship: Universities at the Privacy Frontier. Berkeley Technology Law Journal, 33(2), 365.

Borgman, C. L. (2019). The Lives and After Lives of Data. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.9a36bdb6

Borgman, C. L., Darch, P. T., Sands, A. E., Pasquetto, I. V., Golshan, M. S., Wallis, J. C., & Traweek, S. (2015). Knowledge infrastructures in science: Data, diversity, and digital libraries. International Journal on Digital Libraries, 16(3–4), 207–227. https://doi.org/10.1007/s00799-015-0157-z

Borgman, C. L., Darch, P. T., Sands, A. E., Wallis, J. C., & Traweek, S. (2014). The Ups and Downs of Knowledge Infrastructures in Science: Implications for Data Management. https://doi.org/10.1109/JCDL.2014.6970177

Borgman, C. L., Goshen, M. S., Sands, A. E., Wallis, J. C., Cummings, R. L., Darch, P. T., & Randles, B. M. (2016). Data Management in the Long Tail: Science, Software and Service. The International Journal of Digital Curation, 11(1), 128–149. https://doi.org/10.2218/ijdc.v11i1.428

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007). Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. International Journal on Digital Libraries, 7(1–2), 17–30. https://doi.org/10.1007/s00799-007-0022-9

Borgman, C. L., Wallis, J. C., & Mayernik, M. S. (2012). Who's Got the Data? Interdependencies in Science and Technology Collaborations. Computer Supported Cooperative Work (CSCW), 21(6), 485–523. https://doi.org/10.1007/s10606-012-9169-z

Borgman, C. L., Wallis, J. C., Mayernik, M. S., & Pepe, A. (2007). Drowning in data: Digital library architecture to support scientific use of embedded sensor networks. Proceedings of the 2007 Conference on Digital Libraries - JCDL '07, 269. https://doi.org/10.1145/1255175.1255228

Borgman, C. L., Wofford, M. F., Golshan, M. S., & Darch, P. T. (2020). Collaborative Qualitative Research at Scale:Reflections on 20 years of Acquiring Global Data and Making Data Global. https://escholarship.org/uc/item/3081t2jm

Borgmann, A. (2000). Holding On to Reality: The Nature of Information at the Turn of the Millennium. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/H/bo3640475.html

Borko, H. (1968). Information science: What is it? https://doi.org/10.1002/ASI.5090190103

Borko, H., International Federation for Documentation, & Kungl.Tekniska hogskolan (Eds.). (1971). System analysis, an approach to information: FID/TM tutorial report.

Borralho, A., Fialho, I., & Cid, M. (2014). A Triangulação Sustentada de Dados como Condição Fundamental para a Investigação Qualitativa. http://dspace.uevora.pt/rdpc/handle/10174/13777

Bradford, S. C. (1948). Documentation. Crosby Lockwood.

Bryant, R., Clements, A., Feltes, C., Groenewegen, D., Huggard, S., Mercer, H., Missingham, R., Maliaca Oxnam, Rauh, A., & Wright, J. (2017). Research Information Management: Defining RIM and the Library's Role. https://doi.org/10.25333/C3NK88

Bryant, R. E., Katz, R. H., & Lazowska, E. D. (2008). Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society. 7.

Buckland, M. K. (1991). Information as thing. Journal of the American Society for Information Science, 42(5), 351–360. https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<351::AID-ASI5>3.0.CO;2-3

Candela, L., Castelli, D., Manghi, P., & Tani, A. (2015). Data journals: A survey. Journal of the Association for Information Science and Technology, 66(9), 1747–1762. https://doi.org/10.1002/asi.23358

Capurro, R., & Hjørland, B. (2003). The concept of information. Annual Review of Information Science and Technology, 37(1), 343–411. https://doi.org/10.1002/aris.1440370109

Capurro, R., & Hjorland, B. (2007). The concept of information as we use in everyday. Perspectivas em Ciência da Informação, 12, 148–207.

Carlson, J., & Johnston, L. (Eds.). (2015). Data information literacy: Librarians, data, and the education of a new generation of researchers. Purdue University Press.

Carlson, S., & Anderson, B. (2007). What Are Data? The Many Kinds of Data and Their Implications for Data Re-Use. Journal of Computer-Mediated Communication, 12(2), 635–651. https://doi.org/10.1111/j.1083-6101.2007.00342.x

Case, D. (2003). Looking for Information—A Survey of Research on Information Seeking, Needs, and Behavior. Inf. Res., 8.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, 36(4), 1165–1188. https://doi.org/10.2307/41703503

Chen, S., & Liu, X. (2004). The contribution of data mining to information science. J. Information Science, 30, 550–558. https://doi.org/10.1177/0165551504047928

Cox, A., Kennan, M., Lyon, L., & Pinfield, S. (2017). Developments in research data management in academic libraries: Towards an understanding of research data service maturity. Journal of the Association for Information Science and Technology, 68. https://doi.org/10.1002/asi.23781

Cox, A. M., Kennan, M. A., Lyon, L., Pinfield, S., & Sbaffi, L. (2019). Maturing research data services and the transformation of academic libraries. Journal of Documentation, 75(6), 1432–1462. https://doi.org/10.1108/JD-12-2018-0211

Cox, A. M., Pinfield, S., & Smith, J. (2016). Moving a brick building: UK libraries coping with research data management as a «wicked» problem. Journal of Librarianship and Information Science, 48(1), 3–17. https://doi.org/10.1177/0961000614533717

Cox, A. M., & Tam, W. W. T. (2018). A critical analysis of lifecycle models of the research process and research data management. Aslib Journal of Information Management, 70(2), 142–157. https://doi.org/10.1108/AJIM-11-2017-0251

Cox, A., Verbaan, E., & Sen, B. (2012). Upskilling Liaison Librarians for Research Data Management. Ariadne, 70. http://www.ariadne.ac.uk/issue/70/cox-et-al/

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. Library Review, 65(3), 122–135. https://doi.org/10.1108/LR-06-2015-0061

Delserone, L. M. (2008). At the Watershed: Preparing for Research Data Management and Stewardship at the University of Minnesota Libraries. Library Trends, 57(2), 202–210. https://doi.org/10.1353/lib.0.0032

Drucker, J. (2011). Humanities Approaches to Graphical Display. 22.

East, H. (1983). Documentation—Bradford,sc. Journal of Information Science, 7(3), 127–129.

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. Social Studies of Science, 41(5), 667–690. https://doi.org/10.1177/0306312711413314

Elragal, A. (2014). ERP and Big Data: The Inept Couple. Procedia Technology, 16. https://doi.org/10.1016/j.protcy.2014.10.089

ERC Scientific Council. (2022). Open Research Data and Data Management Plans Information for ERC grantees. European Research Council. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf

Fadiya, S. O., Saydam, S., & Zira, V. V. (2014). Advancing big data for humanitarian needs. Procedia Engineering, 78, 88–95.

Federer, L. (2016). The Medical Library Association Guide to Data Management for Librarians. Rowman & Littlefield.

Federer, L. (2018). Defining data librarianship: A survey of competencies, skills, and training. Journal of the Medical Library Association, 106(3). https://doi.org/10.5195/jmla.2018.306

Federer, L. M., Lu, Y.-L., & Joubert, D. J. (2016). Data literacy training needs of biomedical researchers. Journal of the Medical Library Association : JMLA, 104(1), 52–57. https://doi.org/10.3163/1536-5050.104.1.008

Fienberg, S. E., Martin, M. E., & Straf, M. L. (1985). Sharing research data. Washington, D.C. : National Academy Press. http://archive.org/details/sharingresearchd0000unse

Fitzgerald, A., Fitzgerald, B., & Pappalardo, K. (2009). The Future of Data Policy. Em The fourth paradigm: Data-Intensive Scientific Discovery (pp. 201–208).

Fox, P., & Hendler, J. (2009). Semantic eScience: Encoding Meaning in Next-Generation Digitally Enhanced Science. Em The Fourth paradigm: Data-Intensive Scientific Discovery.

Freitas, M. C. V. de, & Corujo, L. M. N. (2021). Arquivistas, cientistas e dados abertos: Uma equação complexa? Sob a lente da ciência aberta: olhares de Portugal, Espanha e Brasil, 2021, ISBN 978-989-26-2022-0, págs. 189-220, 189–220. https://dialnet.unirioja.es/servlet/articulo?codigo=7805463

Furner, J. (2003). Little Book, Big Book: Before and After Little Science, Big Science: A Review Article, Part I. Journal of Librarianship and Information Science - J LIBR INF SCI, 35, 115–125. https://doi.org/10.1177/0961000603352006

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Garvey, W. D. (1979). Communication, the essence of science: Facilitating information exchange among librarians, scientists, engineers, and students. Oxford ; New York : Pergamon Press. http://archive.org/details/communicationess0000garv

Glaser, B. (2007). All Is Data | Grounded Theory Review. Http://Groundedtheoryreview.Com, 6(2). http://groundedtheoryreview.com/2007/03/30/1194/

Gomes, J. C., Pimenta, R. M., & Schneider, M. (2019). MINERAÇÃO DE DADOS NA PESQUISA EM CIÊNCIA DA INFORMAÇÃO: DESAFIOS E OPORTUNIDADES. XX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2019, 22.

Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., Di Stefano, R., Gil, Y., Groth, P., Hedstrom, M., Hogg, D. W., Kashyap, V., Mahabal, A., Siemiginowska, A., & Slavkovic, A. (2014). 10 Simple Rules for the Care and Feeding of Scientific Data. PLoS Computational Biology, 10(4), e1003542. https://doi.org/10.1371/journal.pcbi.1003542

Gray, J. (2007). Jim Gray on eScience: A Transformed Scientific Method. Em S. T. Tansley & K. Tolle (Eds.), The Fourth Paradigm: Dataintensive scientific discovery. Microsoft Research, 2009.

Griffin, S. (2013). New Roles for Libraries in Supporting Data-Intensive Research and Advancing Scholarly Communication. International Journal of Humanities and Arts Computing, 7(supplement), 59–71. https://doi.org/10/gf867m

Hacid, H., Sheng, Q. Z., Yoshida, T., Sarkheyli, A., & Zhou, R. (2019). Data Quality and Trust in Big Data: 5th International Workshop, QUAT 2018, Held in Conjunction with WISE 2018, Dubai, UAE, November 12–15, 2018, Revised Selected Papers. Springer.

Halevi, G., & Moed, H. (2012). The Technological Impact of Library Science Research: A Patent Analysis.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. https://doi.org/10.1016/C2009-0-61819-5

Hansen, C., Johnson, C., Pascucci, V., & Silva, C. (Eds.). (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery (p. 288).

Hashem, I., Yaqoob, I., Anuar, N., Mokhtar, S., Gani, A., & Khan, S. (2014). The rise of "Big Data" on cloud computing: Review and open research issues. Information Systems, 47, 98–115. https://doi.org/10.1016/j.is.2014.07.006

Hayes, R. M. (1969). Information Science in Librarianship. 19(1–4), 216–236. https://doi.org/10.1515/libr.1969.19.1-4.216

Heidorn, P. B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. Library Trends, 57(2), 280–299.

Heidorn, P. B. (2011). The Emerging Role of Libraries in Data Curation and E-science. Journal of Library Administration, 51(7–8), 662–672. https://doi.org/10.1080/01930826.2011.601269

Heidorn, P. B., Stahlman, G. R., & Chong, S. (2015). Datasphere at the Biosphere II: Computation and data in the wild. https://www.ideals.illinois.edu/handle/2142/73759

Hey, T., & Hey, J. (2006). E-Science and its implications for the library community. Library Hi Tech, 24(4), 515–528. https://doi.org/10.1108/07378830610715383

Hey, T., & Trefethen, A. (2002). The UK e-Science *Core* Programme and the Grid. Future Generation Computer Systems, 18, 1017–1031. https://doi.org/10.1016/S0167-739X(02)00082-1

Hey, T., & Trefethen, A. (2003). The Data Deluge: An e-Science Perspective (pp. 809–824). https://doi.org/10.1002/0470867167.ch36

Hey, T., & Trefethen, A. (2005). Cyberinfrastructure for e-Science. Science, 308. https://doi.org/10.1126/science.1110410

Hiebel, G., Goldenberg, G., Grutsch, C., Hanke, K., & Staudt, M. (2021). FAIR data for prehistoric mining archaeology. International Journal on Digital Libraries, 22(3), 267–277. https://doi.org/10.1007/s00799-020-00282-8

Higgins, S. (2008). The DCC Curation Lifecycle Model. International Journal of Digital Curation, 3(1), 134–140. https://doi.org/10.2218/ijdc.v3i1.48

Higman, R., Bangert, D., & Jones, S. (2019). Three camps, one destination: The intersections of research data management, FAIR and Open. Insights-the Uksg Journal, 32, 18. https://doi.org/10.1629/uksg.468

Higman, R., & Pinfield, S. (2015). Research data management and openness The role of data sharing in developing institutional policies and practices. Program-Electronic Library and Information Systems, 49(4), 364–381. https://doi.org/10.1108/PROG-01-2015-0005

Hilgartner, S., & Brandt-Rauf, S. I. (1994). Data Access, Ownership, and Control: Toward Empirical Studies of Access Practices. Knowledge, 15(4), 355–372. https://doi.org/10.1177/107554709401500401

Holbrook, J. B., & Frodeman, R. (2011). Peer review and the ex ante assessment of societal impacts. Research Evaluation, 20(3), 239–246. https://doi.org/10.3152/095820211X12941371876788

Huang, M.-H., & Huang, M.-J. (2018). An analysis of global research funding from subject field and funding agencies perspectives in the G9 countries. Scientometrics, 115(2), 833–847. https://doi.org/10.1007/s11192-018-2677-y

Hunt, S. (sem data). Research Guides: Data Visualization Services Toolkit for Libraries: Home. Obtido 10 de maio de 2022, de https://libguides.umn.edu/c.php?g=1056829&p=7678485

Jeffery, K. G., Asserson, A., Houssos, N., Brasse, V., & Jörg, B. (2014). From Open Data to Data-intensive Science through CERIF. Procedia Computer Science, 33, 191–198. https://doi.org/10.1016/j.procs.2014.06.032

Johnston, S. F. (2018). Alvin Weinberg and the Promotion of the Technological Fix. Technology and Culture, 59(3), 620–651. https://doi.org/10.1353/tech.2018.0061

Kanfer, A. G., Haythornthwaite, C., Bowker, G. C., Bruce, B. C., Burbules, N. C., Porac, J., & Wade, J. (2000). Modelling distributed knowledge processes in next generation multidisciplinary alliances: Academia/Industry Working Conference on Research Challenges, AIWORC 2000.

Proceedings - Academia/Industry Working Conference on Research Challenges 2000, 83–91. https://doi.org/10.1109/AIWORC.2000.843277

Karasti, H., Baker, K., & Halkola, E. (2006). Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network. Computer Supported Cooperative Work (CSCW). https://doi.org/10.1007/s10606-006-9023-2

Karasti, H., Baker, K. S., & Millerand, F. (2010). Infrastructure Time: Long-term Matters in Collaborative Development. Computer Supported Cooperative Work (CSCW), 19(3), 377–415. https://doi.org/10.1007/s10606-010-9113-z

Kling, R., & McKim, G. (2000). Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication. Journal of the American Society for Information Science, 51(14), 1306–1320. https://doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1047>3.0.CO;2-T

Knorr-Cetina, K. (Karin). (1999). Epistemic cultures: How the sciences make knowledge. Cambridge, Mass. : Harvard University Press. http://archive.org/details/epistemicculture0000knor

Koltay, T. (2015). Data literacy: In search of a name and identity. Journal of Documentation, 71(2), 401–415. https://doi.org/10.1108/JD-02-2014-0026

Koltay, T. (2016). Data governance, data literacy and the management of data quality. Ifla Journal-International Federation of Library Associations, 42(4), 303–312. https://doi.org/10.1177/0340035216672238

Koltay, T. (2017). Data literacy for researchers and data librarians. Journal of Librarianship and Information Science, 49(1), 3–14. https://doi.org/10.1177/0961000615616450

Koltay, T. (2019). Accepted and Emerging Roles of Academic Libraries in Supporting Research 2.0. Journal of Academic Librarianship, 45(2), 75–80. https://doi.org/10.1016/j.acalib.2019.01.001

Kuhn, T. S. (1970). The structure of scientific revolutions ([2d ed., enl). University of Chicago Press.

Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. International Journal of Information Management, 34(3), 387–394. https://doi.org/10.1016/j.ijinfomgt.2014.02.002

Laitinen, M., & Saarti, J. (2012). A model for a library-management toolbox: Data warehousing as a tool for filtering and analyzing statistical information from multiple sources. Library Management, 33(4/5), 253–260. https://doi.org/10.1108/01435121211242290

Lakatos, I., & Musgrave, A. (1980). A Crítica e o Desenvolvimento do Conhecimento. Ciência e filosofia, 2, 157–162. https://doi.org/10.11606/issn.2447-9799.cienciaefilosofi.1980.107354

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. Studylib.Net. https://studylib.net/doc/8647594/3d-data-management--controlling-data-volume--velocity--an...

Latour, B., & Woolgar, S. (1986). Laboratory Life: The Construction of Scientific Facts. Princeton, N.J.: Princeton Univers ity Press. https://books.google.pt/books/about/Laboratory_Life.html?id=XTcjm0flPdYC&redir_esc=y

Laure Perrier & Leslie Barnes. (2018). Developing Research Data Management Services and Support for Researchers: A Mixed Methods Study. Partnership: The Canadian Journal of Library and Information Practice and Research, 13(1). https://doi.org/10.21083/partnership.v13i1.4115

Lynch, M. E., & Woolgar, S. (Eds.). (1990). Representation in Scientific Practice. MIT Press.

Lyon, L. (2007). Dealing with Data: Roles, Rights, Responsibilities and Relationships Consultancy Report.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. (2011). Big Data: The Next Frontier for Innovation, Comptetition, and Productivity.

Mayernik, M. S. (2011). Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators. https://doi.org/10.2139/ssrn.2042653

McDonough, A. M. (1963). Information economics and management systems. —. New York : McGraw-Hill. http://archive.org/details/informationecono0000mcdo

Medina-Smith, A., Tryka, K. A., Silcox, B. P., & Hanisch, R. J. (2016). Librarians and Scientists Partner to Address Data Management: Taking Collaboration to the Next Level. Digital library perspectives, 32(3), 142–152. https://doi.org/10.1108/DLP-08-2015-0010

Mejia, C., & Kajikawa, Y. (2018). Using acknowledgement data to characterize funding organizations by the types of research sponsored: The case of robotics research. Scientometrics, 114(3), 883–904. https://doi.org/10.1007/s11192-017-2617-2

Merton, R. K. (1968). Social theory and social structure. New York, Free Press. http://archive.org/details/socialtheorysoci00mert

Miranda, P., Rehemtula, S., Cardoso, J., Correia, A., & Pereira, F. (2021). Planos de Gestão de Dados: Formação para Research Support Staff. 123.

Mitchum, R. (2012). Unwinding the 'Long Tail' of Science | Computation Institute. Computation Institute. https://voices.uchicago.edu/compinst/blog/unwinding-long-tail-science/

National Research Council. (1999). A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases. The National Academies Press. https://doi.org/10.17226/9692

National Science Foundation. (2005). Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. https://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

Newton, M. P., Miller, C. C., & Bracke, M. S. (2010). Librarian Roles in Institutional Repository Data Set Collecting: Outcomes of a Research Library Task Force. Collection Management, 36(1), 53–67. https://doi.org/10.1080/01462679.2011.530546

Nitecki, D. A., & Davis, M. E. (sem data). Changing Landscapes: New Roles for Academic Librarians. 11.

Paisley, W. (1968). Information Needs and Uses. Annual Review of Information Science and Technology, 3, 1–30.

Palmer, C., Cragin, M., Heidorn, P., & Smith, L. (2007). Data curation for the long tail of science: The case of environmental sciences.

Palmer, C. L. (1996). Information work at the boundaries of science: Linking library services to research practices. Library Trends, 45(2), 165–191.

Palmer, C. L. (2005). Scholarly work and the shaping of digital access. Journal of the American Society for Information Science and Technology, 56(11), 1140–1153. https://doi.org/10.1002/asi.20204

Panneerselvam, J., Liu, L., & Hill, R. (2015). Chapter 1—An Introduction to Big Data. Em B. Akhgar, G. B. Saathoff, H. R. Arabnia, R. Hill, A. Staniforth, & P. S. Bayerl (Eds.), Application of Big Data for National Security (pp. 3–13). Butterworth-Heinemann. https://doi.org/10.1016/B978-0-12-801967-2.00001-X

Piccolo, D. M., Wolf Tadini, A. V., Teixeira, H. D., Botega, L. C., Goncalves Sant'Ana, R. C., Santarem Segundo, J. E., & Vesu Alves, R. C. (2022). Data quality in research data management: A bibliometric study. Em Questao, 28(1), 159–184. https://doi.org/10.19132/1808-5245281.159-184

Piorun, M., Kafel, D., Leger-Hornby, T., Najafi, S., Martin, E., Colombo, P., & LaPelle, N. (2012). Teaching Research Data Management: An Undergraduate/Graduate Curriculum. Journal of eScience Librarianship, 1(1). https://doi.org/10.7191/jeslib.2012.1003

Piwowar, H. A., & Chapman, W. W. (2010). Public sharing of research datasets: A pilot study of associations. Journal of Informetrics, 4(2), 148–156. https://doi.org/10.1016/j.joi.2009.11.010

Piwowar, H., Day, R., & Fridsma, D. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. PloS one, 2, e308. https://doi.org/10.1371/journal.pone.0000308

Pomerantz, J. (2015). Metadata. MIT Press.

Poole, A. H., & Garwood, D. A. (2020). Digging into data management in public-funded, international research in digital humanities. Journal of the Association for Information Science and Technology, 71(1), 84–97. https://doi.org/10.1002/asi.24213

Price, D. J. de S. (Derek J. de S. (1986). Little science, big science—And beyond. New York : Columbia University Press. http://archive.org/details/littlesciencebig00pric

Primich, T. (2010). A Semester-Long Seminar in Statistical Visualization for Undergraduates as Taught by a Science and Engineering Librarian. Science & Technology Libraries, 29(3), 181–188. https://doi.org/10.1080/0194262X.2010.497702

Príncipe, P., Correia, A., Moura, P. C. M., & Rodrigues, E. (2018). Estratégia Institucional para a gestão de dados de investigação na UMINHO: O papel dos SDUM. Actas do Congresso Nacional de Bibliotecários, Arquivistas e Documentalistas, 13, Article 13. https://publicacoes.bad.pt/revistas/index.php/congressosbad/article/view/1812

Príncipe, P., & Furtado, F. R. C. D. (2017). Relatório do 2º Fórum GDI [Report]. http://repositorium.sdum.uminho.pt/

Príncipe, P., Moura, P., & Pereira, F. (2021). Relatório do 7.º Fórum GDI [Report]. http://repositorium.sdum.uminho.pt/

Príncipe, P., Rodrigues, E., Vieira, A., Correia, A., Carvalho, José, & Moura, P. (2021). O Essencial da Gestão de Dados de Investigação. NAU Site. http://www.nau.edu.pt/en/course/o-essencial-da-gestao-de-dados-de-investigacao/

Príncipe, P., & Silva, D. (2018). Gestão de Dados de Investigação: O papel das Bibliotecas em Portugal – estratégias, serviços e competências. Actas do Congresso Nacional de Bibliotecários, Arquivistas e Documentalistas, 13, Article 13. https://publicacoes.bad.pt/revistas/index.php/congressosbad/article/view/1886

Principe, P., Silva, D., Sanches, T., Lopes, S., Pereira, A. A., Lopes, C., Antunes, M. L., Carvalho, M., Vargues, M. M., Saraiva, P. S., Aurindo, M. J., Martins, T. A., Amante, M. J., Cunha, T., Guerreiro, D., Carvalho, M. de, Pireza, I., Gonçalves, A., Carvalho, C., … Correia, M. A. (2020). Recomendações para as Bibliotecas do Ensino Superior de Portugal 2020-2022. Zenodo. https://doi.org/10.5281/zenodo.3841363

Ramos, P. C. C. da S. (2014). The impact of rewards on the success of crowdfunding campaigns. https://repositorio.ucp.pt/handle/10400.14/15934

Raspa, R., & Ward, D. (2000). The Collaborative imperative: Librarians and faculty working together in the information universe. Chicago : Association of College and Research Libraries. http://archive.org/details/collaborativeimp00rich

Read, K. B., LaPolla, F. W. Z., Tolea, M. I., Galvin, J. E., & Surkis, A. (2017). Improving data collection, documentation, and workflow in a dementia screening study. Journal of the Medical Library Association, 105(2), 160–166. https://doi.org/10.5195/jmla.2017.221

Regan, C. J. (2012). Review: The Fourth Paradigm by Tony Hey, Stewart Tansley, and Kristin Tolle. InterActions: UCLA Journal of Education and Information Studies, 8(1). https://doi.org/10.5070/D481011836

Revez, J. M. R. (2019). O papel das bibliotecas na investigação científica: Perceções, comportamento informacional e impacto. https://estudogeral.sib.uc.pt/handle/10316/87349

Ribes, D., Baker, K. S., Millerand, F., & Bowker, G. C. (2005). Comparative interoperability project: Configurations of community, technology, organization. Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, 65–66. https://doi.org/10.1145/1065385.1065399

Ribes, D., & Finholt, T. (2009). The Long Now of Technology Infrastructure: Articulating Tensions in Development. Journal of the Association for Information Systems, 10(5). https://doi.org/10.17705/1jais.00199

Rice, R., & Haywood, J. (2011). Research Data Management Initiatives at University of Edinburgh. International Journal of Digital Curation, 6(2), 232–244. https://doi.org/10.2218/ijdc.v6i2.199

Rodrigues, S. P. (2011). A Web 2.0 nos arquivos portugueses. https://repositorio.ucp.pt/handle/10400.14/8818

Saracevic, T. (1999). Information science. Journal of the American Society for Information Science, 50(12), 1051–1063. https://doi.org/10.1002/(SICI)1097-4571(1999)50:12<1051::AID-ASI2>3.0.CO;2-Z

Schneider, R. (2013). Research Data Literacy. Em S. Kurbanoğlu, E. Grassian, D. Mizrachi, R. Catts, & S. Špiranec (Eds.), Worldwide Commonalities and Challenges in Information Literacy Research and Practice (Vol. 397, pp. 134–140). Springer International Publishing. https://doi.org/10.1007/978-3-319-03919-0_16

Shankar, K. (2003). Scientific data archiving: The state of the art in information, data, and metadata management. https://escholarship.org/uc/item/33x8j76m

Shannon, C., & Weaver, W. (1949). The Mathematical Theory of Communication. 131.

Shorish, Y. (2015). Data Information Literacy and Undergraduates: A Critical Competency. College & Undergraduate Libraries, 22, 97–106. https://doi.org/10.1080/10691316.2015.1001246

Steiner, Tobias. (2018). Metadaten und OER: Geschichte einer Beziehung. https://doi.org/10.25656/01:15741

Surkis, A., LaPolla, F. W. Z., Contaxis, N., & Read, K. B. (2017). Data Day to Day: Building a community of expertise to address data skills gaps in an academic medical center. Journal of the Medical Library Association, 105(2), 185–191. https://doi.org/10.5195/jmla.2017.35

Swan, A., & Brown, S. (2008). The Skills, Role and Career Structures of Data Scientists and Curators: An Assessment of Current Practice and Future Needs.

van Selm, M. (2020). If a Researcher Would Meet a Librarian... (The State of Open Data 2020, pp. 4–6) [Report]. Digital Science. https://doi.org/10.6084/m9.figshare.13227875.v2

Veiga, V. S. de O., Silva, C. H., & Borges, M. M. (2021). Modelo de fatores que influenciam no comportamento de compartilhamento de dados de pesquisa (MFDADOS). Em Sob a lente da ciência aberta: Olhares de Portugal, Espanha e Brasil (pp. 153–188). Imprensa da Universidade de Coimbra. https://doi.org.10.14195/978-989-26-2022-0

Vilar, P., & Zabukovec, V. (2019). Research data management and research data literacy in Slovenian science. Journal of Documentation, 75(1), 24–43. https://doi.org/10.1108/JD-03-2018-0042

Wallis, J. C., Mayernik, M. S., Borgman, C. L., & Pepe, A. (2010). Digital libraries for scientific data discovery and reuse: From vision to practical reality. Proceedings of the 10th annual joint conference on Digital libraries, 333–340. https://doi.org/10.1145/1816123.1816173

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PloS One, 8(7), e67332. https://doi.org/10.1371/journal.pone.0067332

Wallis, J. C., Wynholds, L. A., Borgman, C. L., Sands, A. E., & Traweek, S. (2012). Data, data use, and inquiry: A new point of view on data curation. https://escholarship.org/uc/item/24x7c6pq

Wang, J., & Shapira, P. (2011). Funding acknowledgement analysis: An enhanced tool to investigate research sponsorship impacts: the case of nanotechnology. Scientometrics, 87(3), 563–586. https://doi.org/10.1007/s11192-011-0362-5

Wang, X., Duan, Q., & Liang, M. (2021). Understanding the process of data reuse: An extensive review. Journal of the Association for Information Science and Technology, 72(9), 1161–1182. https://doi.org/10.1002/asi.24483

Wang, X., Liu, D., Ding, K., & Wang, X. (2012). Science funding and research output: A study on 10 countries. Scientometrics, 91(2), 591–599. https://doi.org/10.1007/s11192-011-0576-6

Weinberg, A. M. (1961). Impact of Large-Scale Science on the United States. Science, 134(3473), 161–164. https://doi.org/10.1126/science.134.3473.161

Wellisch, H. (1972). From Information Science to Informatics—Terminological Investigation. Journal of Librarianship, 4(3), 157–187. https://doi.org/10.1177/096100067200400302

Wersig, G. (1976). Terminology of documentation = Terminologie de la documentation = Terminologie der Dokumentation = Terminologii͡a v oblasti dokumentat͡sii = Terminología de la documentación: A selection of 1200 basic terms published in English, French, German,

Russian, and Spanish. Paris : The Unesco Press. http://archive.org/details/terminologyofdoc0000wers

Wersig, G. (1993). Information science: The study of postmodern knowledge usage. Information Processing & Management, 29(2), 229–239. https://doi.org/10.1016/0306-4573(93)90006-Y

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18

Williams, M., Bagwell, J., & Nahm Zozus, M. (2017). Data management plans, the missing perspective. Journal of Biomedical Informatics, 71, 130–142. Scopus. https://doi.org/10.1016/j.jbi.2017.05.004

Wilms, K. L., Stieglitz, S., Ross, B., & Meske, C. (2020). A value-based perspective on supporting and hindering factors for research data management. International Journal of Information Management, 54, 102174. https://doi.org/10.1016/j.ijinfomgt.2020.102174

Witt, M. (2016). 23 Things: Libraries for Research Data. https://doi.org/10.15497/RDA00005

Wynholds, L. A., Wallis, J. C., Borgman, C. L., Sands, A., & Traweek, S. (2012). Data, data use, and scientific inquiry: Two case studies of data practices. https://doi.org/10.1145/2232817.2232822

Wynholds, L., Fearon, D., Borgman, C. L., & Traweek, S. (2011). Slides for When use cases are not useful: Data practices, astronomy, and digital libraries. https://escholarship.org/uc/item/6fx563fc

Yang, C., Zhang, X., Zhong, C., Liu, C., Pei, J., Ramamohanarao, K., & Chen, J. (2014). A spatiotemporal compression based approach for efficient big data processing on Cloud. Journal of Computer and System Sciences, 80(8), 1563–1583. https://doi.org/10.1016/j.jcss.2014.04.022

Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. International Journal of Information Management, 36(6, Part B), 1231–1247. https://doi.org/10.1016/j.ijinfomgt.2016.07.009

Yi, K., Chen, T., & Cong, G. (2018). Library personalized recommendation service method based on improved association rules. Library Hi Tech, 36(3), 443–457. https://doi.org/10.1108/LHT-06-2017-0120

Yoon, A., & Schultz, T. (2017). Research Data Management Services in Academic Libraries in the US: A Content Analysis of Libraries' Websites. College & Research Libraries, 78. https://doi.org/10.5860/crl.78.7.920

Zhao, D. (2010). Characteristics and impact of grant-funded research: A case study of the library and information science field. Scientometrics, 84(2), 293–306. https://doi.org/10.1007/s11192-010-0191-y

Zhu, Y. (2020). Open-access policy and data-sharing practice in UK academia. Journal of Information Science, 46(1), 41–52. https://doi.org/10.1177/0165551518823174

Zhu, Y., & Xiong, Y. (2015). Towards Data Science. Data Science Journal, 14(0), 8. https://doi.org/10.5334/dsj-2015-008

Zimmerman, A. (2008). Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists.

Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. JASIST, 58, 479–493. https://doi.org/10.1002/asi.20508