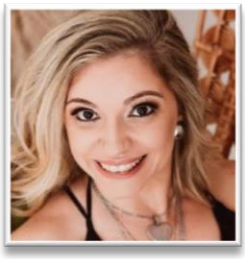
ISSN 1678-765X doi 10.20396/rdbci.v21i00.8673740

Artigo

RDBCI
Revista Digital de Biblioteconomia e Ciência da Informação
Digital Journal of Library and Information Science



Conectando a eScience à Ciência da Informação: o big metadado científico e suas funcionalidades

Luana Faria Sales Marques [1] Luís Fernando Sayão [2]

RESUMO

Introdução: No ambiente da eScience, os objetos digitais de pesquisa são caracterizados por terem um ciclo de vida complexo e longo, que depende de diferentes contextos disciplinares e perspectivas de (re)uso. Este ciclo de vida começa antes do início da pesquisa e se estende para além do final do projeto, ao longo dessa jornada, vários tipos de metadados são adicionados aos objetos, atribuídos por diferentes atores, incluindo aqueles gerados automaticamente por instrumentos científicos e ferramentas de workflow, num processo contínuo de agregação de valor aos conjuntos de dados e a outros objetos de pesquisa. Nesse contexto, os objetos digitais de pesquisa são acompanhados por uma ampla gama de metadados - com muitas funções e propriedades - que muitas vezes superam os próprios dados em volume e até em importância, configurando um "big metadado científico" de difícil organização e gestão. Objetivo: Apresentar de forma sistematizada as funções dos novos metadados a fim de apoiar a gestão de metadados e a construção de esquemas disciplinares. Metodologia: Subjacente à construção da proposta, quatro eixos dão sustentação metodológica ao estudo: histórico, pragmático, de padronização e epistemológico. Resultados: Como resultado é proposto um modelo para esquematização dos diversos elementos de metadados baseado nas suas funcionalidades, tendo como pressuposto a conexão da eScience com a Ciência da Informação estabelecida pelo big metadado. Conclusão: Conclui-se que o big metadado cria uma conexão entre a eScience e a CI, e que para além da necessidade da curadoria dos objetos de pesquisa, é necessário também uma gestão FAIR especifica para os metadados.

PALAVRAS-CHAVE

Ciência da Informação. E-Science. Big metadado. Dados científicos. Gestão de metadados Funcionalidade de metadados. Objetos digitais de pesquisa.

Correspondência dos autores

[1] Instituto Brasileiro de Informação em Ciência e Tecnologia (PPGCI)
Rio de Janeiro, RJ - Brasil
luanafsales@gmail.com

[2] Comissão Nacional de Energia Nuclear
Rio de Janeiro, RJ - Brasil
luis.sayao@cnen.gov.br

Connecting eScience to Information Science: scientific big metadata and its functionalities

ABSTRACT

Introduction: In the eScience environment, digital research objects are characterized by having a complex and long-life cycle, which depends on different disciplinary contexts and perspectives of (re)use. This lifecycle starts before the start of research and extends beyond the end of the project, along this journey, various types of metadata are added to objects, assigned by different actors, including those generated automatically by scientific instruments and workflow tools, in a

continuous process of adding value to datasets and other research objects. In this context, digital research objects are accompanied by a wide range of metadata - with many functions and properties - that often surpass the data themselves in volume and even in importance, configuring a "scientific big metadata" that is difficult to organize and manage. Objective: Systematically present the functions of new metadata to support metadata management and the construction of disciplinary schemes. Methodology: Underlying the construction of the proposal, four axes provide methodological support to the study: historical, pragmatic, standardization, and epistemological. Results: As a result, a model is proposed for schematizing the various elements of metadata based on their functionalities, based on the assumption of the connection between eScience and Information Science established by big metadata. Conclusion: It is concluded that big metadata creates a connection between eScience and CI, and that in addition to the need to curate research objects, specific FAIR management of metadata is also necessary.

KEYWORDS

Information Science. E-science. Big metadata. Metadata management. Research data. Metadata functions. Research digital object.

CRediT

- Reconhecimentos: Os autores gostariam de agradecer à Teodora Marly Gama pela preciosa revisão e normalização do artigo.
- Financiamento: Este estudo foi financiado pelas agências brasileiras Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) para as bolsas e apoio financeiro concedido. Este estudo foi financiado pelas FAPERJ através do apoio financeiro concedido no Processo SEI-260003/003239/2022.
- Conflitos de interesse: Os autores certificam que não têm interesse comercial ou associativo que represente um conflito de interesses em relação ao manuscrito.
- Aprovação ética: Não aplicável.
- Disponibilidade de dados e material: Não aplicável.
- Contribuições dos autores: Conceitualização; Curadoria de dados: SALES, L.; Análise formal: Não se aplica; Aquisição de financiamento: SALES, L.; Investigação; Metodologia: SAYÃO, L. Administração do projeto: SALES, L.; Recursos: Não se aplica; Software: Não se aplica; Supervisão: SAYÃO, L.; Validação: Não se aplica; Visualização: SAYÃO, L.; Escrita – rascunho original: SALES, L.; Escrita – revisão & edição: SAYÃO, L.

JITA: IN. Open science.

Artigo submetido ao sistema de similaridade

Submetido em: 04/06/2023 – Aceito em: 17/08/2023 – Publicado em: 30/08/2023

Editor: Gildenir Carolino Santos

1 INTRODUÇÃO

Na tentativa de identificar um ponto de partida histórico para a jornada da curadoria de dados da pesquisa, Alyssa Goodman e colaboradores (2014) remontam ao início dos anos 1600, quando Galileu Galilei (1564-1642) voltou seu icônico telescópio para Júpiter. Em seu caderno de notas, a cada noite, Galileu desenhava diagramas esquemáticos em escala de Júpiter e de alguns pontos estranhamente móveis próximos a ele, que correspondiam às luas de Júpiter. "Seu estilo claro e cuidadoso de registro e publicação não permitia apenas que Galileu compreendesse o Sistema Solar, mas permitia, ao longo dos séculos, que qualquer pessoa entendesse como Galileu o fez" (Goodman *et al.*, 2014, p.1). Para atingir esse objetivo, Galileu integrou seus dados, metadados e texto em suas anotações: os dados correspondem aos desenhos de Júpiter e de suas luas; metadados-chave são o tempo de observação, condições meteorológicas, propriedades do telescópio; e os textos compreendem a descrição do método de análise e conclusões. A abordagem integrativa dos registros científicos de Galileu – quanto à observação e análise – contribuiu decisivamente para a construção do método científico moderno. Assim como o astrônomo Tycho Brahe (1546-1601), cuja organização dos dados permitiu que Johannes Kepler (1571-1630) formulasse as leis do movimento planetário, Galileu foi pioneiro ao delinear o conceito de curadoria de dados - que hoje desempenha um papel importante na Astronomia contemporânea - estabelecendo formas e ferramentas para descrever e registrar dados (Gray *et al.*, 2002).

Passados muitos séculos da odisseia do progresso científico, na qual se sucederam paradigmas, cujos ciclos de validade podem ser caracterizados pela proposição do físico Thomas Kuhn (1962), que preconizava que uma mudança de paradigma científico ocorre porque o modo dominante de ciência não consegue dar conta de fenômenos particulares e nem responder questões-chave, exigindo assim a formulação de novas ideias (Kitchin, 2014). Jim Gray – um proeminente cientista computacional – nos convida a considerar uma nova visão sobre mudanças de paradigmas científicos que diferem, em sua concepção, da proposição de Kuhn. Ele acreditava que a maior parte da nova ciência acontece quando os dados são examinados de novas maneiras (Gray *et al.*, 2005). As transições de Gray são baseadas em avanços nas formas de dados e no desenvolvimento de novos métodos analíticos. Na sua concepção, o quarto paradigma sucede a ciência experimental, teórica e mais recentemente a computacional.

Este ponto de inflexão marca a entrada da ciência numa conformação analítica e metodológica disruptiva denominada de "quarto paradigma científico". Em uma de suas últimas palestras, em janeiro de 2007, Gray delineou a sua visão sobre esse conceito inovador, caracterizando-o como um novo enfoque metodológico orientado para a descoberta baseada na ciência intensiva em dados, além da pesquisa experimental e teórica e na simulação por computador de fenômenos naturais (Newman, 2019). Desde então, a expressão quarto paradigma tornou-se um termo genérico para se referir à abordagem interdisciplinar da pesquisa científica orientada por dados ou eScience, cujas atuações estão baseadas em dois fatores essenciais: a colaboração cientifica global em áreas-chave da ciência e em uma infraestrutura tecnológica e social avançada necessária para viabilizar essa colaboração. Nesse cenário de mudanças, novas metodologias e ferramentas analíticas passam a ser possíveis pelo avanço vertiginoso das tecnologias digitais de informação e comunicação, que, além do mais, se tornaram mais ricas e flexíveis e muito mais fáceis de usar (De Roure *et al*., 2003*)*.

Percebe-se que, nesse cenário de transição, a pesquisa científica está cada vez mais dependente das tecnologias digitais. Porém, como desdobramento imediato desses novos condicionamentos, depende também de diferentes modos de abstração e representação dos objetos de pesquisa e de suas relações e contextos nos diversos domínios digitais ou ciberinfraestruturas de pesquisa em que são gerenciados e consumidos por seres humanos e por sistemas. Isto porque na escala e velocidade em que os dados são gerados, a sua utilização só é

possível por intermédio de métodos computacionais. Subjacente a todos esses novos desafios está a ideia de construir uma infraestrutura baseada em metadados ricos para representar os recursos presentes no ambiente de pesquisa e para apoiar a otimização de seu reúso (Mons *et al*., 2017). Todavia, isso implica na necessidade da criação de um modelo de informação de dados de pesquisa mais realista, que considere a complexidade de uma nova condição de publicação de dados que seja aberta, contextualizada e em rede, e que seja capaz de proporcionar um amplo espectro de funcionalidades que conduzam os objetos de pesquisa a níveis apropriados de "fairificação", entendida como percursos da curadoria para tornar os dados FAIR[1] (Sales; Sayão, 2022).

Essa nova configuração científica coloca outro desafio importante: para cada *dataset* que circula nessas infraestruturas sociotécnicas, são agregados inúmeros elementos de metadados que documentam e performam funções críticas sobre esses *datasets*. Esta agregação de informação constitui uma outra vertente do *big data* científico, chamado de "*big metadata*". Assim como o dado é crucial para a ciência contemporânea, o big metadado relativo a objetos de pesquisa é crucial para a descoberta, compartilhamento, reuso, crédito aos autores e reprodutibilidade dos experimentos científicos, para citar algumas funcionalidades possíveis realizadas sobre os dados e viabilizadas por elementos de metadados. "Assim, por todas as áreas, os big metadados são frequentemente mais massivos e complexos do que os dados [propriamente dito] que eles representam", resumem Sarah Bratt e colaboradores (2017, p.36).

Para os objetos digitais de pesquisa, num contexto de integração interdisciplinar preconizado pela eScience, a função dos metadados vai muito além dos processos descritivos aplicados a documentos impressos convencionais, sintetizados vagamente pela frase "metadados são dados sobre dados". A multiplicidade de funções dos metadados necessários à gestão FAIR dos objetos digitais de pesquisa, cujos princípios subsidiam as ações que os tornam encontráveis, acessíveis, interoperáveis, de forma que eles possam ser compreendidos ao longo do tempo, para reutilização em diversos contextos diferentes do projeto original, implica num volume considerável e diversificado de metadados adicionado a cada objeto digital. Os elementos de metadados vão adicionando valor aos dados ao longo do seu ciclo de vida – da geração ao arquivamento de longo prazo – por diversos agentes, incluindo *softwares* e instrumentos científicos. Entretanto, muitas vezes esses elementos estão desordenados, são ambíguos, complexos e desestruturados, e precisam de um grau de esquematização para apoiar a construção de esquemas de metadados disciplinares mais representativos.

Tentando contribuir para o equacionamento dessa outra vertente do *big data* científico, que se instala em torno dos objetos digitais de pesquisa, o presente ensaio tem como objetivo ampliar a perspectiva dos estudos sobre representação à medida que revela a sua importância dentro de um novo contexto informacional: a dos objetos digitais de pesquisa. Subjacente a esta proposta está uma questão orientadora: Que funções o metadado exerce na gestão de objetos digitais de pesquisa dentro do contexto da eScience? Para respondê-la, quatro eixos – histórico, pragmático, de padronização e epistemológico – são considerados na construção do percurso metodológico do estudo, a saber: a perspectiva histórica e fundacional da arquitetura do objeto de pesquisa iniciada por Robert Kahn e Robert Wilensky (1995); a visão pragmática do cientista da computação Jim Gray e colaboradores (2002, 2005), visão de Gray também presente em "Jim Gray on eScience: A transformed scientific method", editado por Hey, Tansley e Tolle, (2009) sobre a importância dos metadados para a proveniência e contextualização dos experimentos em ambiente eScience; a visão padronizada do OAIS, modelo conceitual do Open Archive Information System (CCSDS, 2012) sobre o objeto digital informacional e as informações de representação que constroem seu significado; a perspectiva epistemológica de Wouters (2006) e Rheinberger (1977) sobre a relação entre coisas epistêmicas, objetos epistêmicos e objetos técnicos; e a essencialidade da representação e contextualização/ descontextualização na

[1] Acrônimo para Findable, Acessible, Interoperable, Reusable.

eScience.

2 A REPRESENTAÇÃO COMO UMA DIMENSÃO CRÍTICA DA eSCIENCE

Pensando em voz alta sobre a informatização na criação do conhecimento, Paul Wouters (2006) destaca que, "eScience é uma construção discursiva na interface de práticas técnico-científicas, projeto de tecnologia de computador e política científica" (Wouters, 2006, p.7), onde práticas e tecnologias muito diferentes estão sendo integradas. Neste domínio, instrumentos científicos e simulação computacional estão criando vastos silos de dados que requerem novos métodos científicos para analisar e organizar os conjuntos de dados neles contidos. Isto se torna possível por meio do desenvolvimento em larga escala de sistemas sociotécnicos que processam, na escala necessária, essa abundância de dados e que materializam a promessa de novas descobertas em todas as áreas da ciência.

Assim sendo, a eScience passa a ser nada menos que uma revolução sobre como o conhecimento pode ser criado, tendo como alicerce conceitual a combinação de vários e diferentes desenvolvimentos como, por exemplo: o compartilhamento de recursos computacionais, como a computação em grade; o acesso distribuído e a integração de conjuntos massivos de dados; o uso de plataformas digitais integrativas, como as atuais ciberinfraestruturas de pesquisa; e de ferramentas e metodologias avançadas de análise, como a inteligência artificial e os métodos estatísticos sofisticados. Uma vez estabelecida, a eScience provoca uma intensificação na confiança em dados processados: dados capturados por sensores e instrumentos e processados por *software*, armazenados em computadores e gerenciados por estatísticas ou **metadados,** ou ambos (Newman, 2019), que realiza a aplicação de tecnologia computacional para o empreendimento da investigação científica moderna, onde a "TI encontra os cientistas", resume Gray (Hey; Tansley; Tolle, 2009, p. xviii). A força avassaladora trazida por essa revolução epistêmica afeta quase todos os domínios disciplinares, que estão sendo transformados de uma forma ou de outra; tomados em conjunto, todos os ramos das ciências – das ciências exatas às ciências sociais e humanas, e mesmo a cultura e as artes – são agora incluídos nas promessas da eScience (Wouters, 2006).

Como testemunha próxima e protagonista dessa transição, Jim Gray (Hey; Tansley; Tolle, 2009) constatou que, no contexto da eScience, estamos observando a evolução de duas vertentes de exploração científica que se desenvolvem no seio de cada disciplina, ambas gerando e consumindo massivamente dados: a **ciência computacional** e a **ciência intensiva em dados** (x-computacional e x-informática). Essas vertentes, apesar de ocorrerem no mesmo domínio disciplinar, são tão diferentes que vale a pena distingui-las. A ciência computacional está relacionada com a **simulação** de fenômenos naturais e sociais, por exemplo, a neurociência computacional simula como o cérebro trabalha. Por sua vez, a ciência intensiva em dados **coleta, faz curadoria, integra e analisa dados e informações** de muitos e diferentes experimentos. Por exemplo, enquanto a ecologia computacional simula a dinâmica dos sistemas ecológicos, a ecoinformática coleta, integra e analisa os dados e informações obtidos durante os experimentos. Newman (2019) considera que o efeito desse paradigma computacional científico é a emergência de um "pensamento computacional" que significa aplicar processos computacionais ao problema em pauta, reformulando o problema aparentemente difícil – como o comportamento humano ou um projeto de sistemas - em um problema que sabemos como resolver, por redução, incorporação, transformação ou simulação, ou seja, baseando-se em conceitos fundamentais para a ciência da computação. O pensamento computacional inclui um espectro de ferramentas mentais que refletem a amplitude da ciência da computação, conclui Wing (2006), localizando a abstração e a representação no centro da discussão. Esses dois ramos da ciência – ciência computacional e ciência intensiva em dados - estabelecem uma

interlocução dinâmica na direção da geração de novos conhecimentos por meio de poderosas infraestruturas, ferramentas e metodologias, que se pode sintetizar pelo termo "ciberinfraestrutura".

Uma ciberinfraestrutura de pesquisa "é um meio que permite acesso e circulação de conhecimento distribuído, em que colaboram e se comunicam diferentes comunidades e disciplinas, rompendo fronteiras culturais, geográficas e temporais", esclarece Pérez-González (2010, p. 3); é "uma nova forma de cultura científica que se sustenta em uma robusta infraestrutura tecnológica de alto nível", completa o autor. Os dispositivos oferecidos por essa infraestrutura dão apoio a mecanismos inéditos de colaboração, baseados no acesso a uma quantidade extraordinária de dados, recursos informacionais interpretados e reutilizados por potentes ferramentas de observação, visualização e simulação.

Essas infraestruturas avançadas de pesquisa afetaram profundamente o processo longamente consolidado da metodologia científica tradicional e aumentaram enormemente o nível de produção e uso de dados em pesquisa, permitindo novos tipos de experimentos, observações, medições, análises, imagens e visualização de dados, segundo Gray (Hey; Tansley; Tolle, 2009). As mudanças metodológicas decorrentes implicam no deslocamento do ponto inicial das pesquisas, posto que a dinâmica integrativa dos fluxos de dados cria padrões, anomalias, relacionamentos e contextos que são as lentes através das quais se estudam um dado fenômeno e não uma hipótese inicial baseada em racionalidades indutivas ou dedutivas: os dados é que são o ponto de partida e não a comprovação de uma hipótese ou teoria (Newman, 2019). Entretanto, essas rupturas não se limitam somente às infraestruturas de pesquisa e à conceituação metodológica, para trabalhar nesses novos espaços de significação, uma nova geração de e-cientistas está surgindo. Eles estão criando formas inéditas de trabalhar, entendem profundamente as possibilidades das tecnologias e realizam suas pesquisas, não como um ser humano individual, mas como um nó em uma rede de humanos e máquinas (Wouters, 2006), que redefinem os padrões da comunicação científica baseados em periódicos científicos.

Nesse novo contexto científico, os seres humanos são incapazes de sincronizar a sua capacidade operativa com o escopo, escala e velocidade adequados à magnitude do *big data* científico e com a complexidade da eScience. Por conseguinte, esse fenômeno de nossos dias postula a necessidade de os seres humanos cada vez mais contarem com agentes computacionais para realizar tarefas de descoberta e integração de dados e informações em seu nome (Wilkinson *et al.*, 2016). "Os computadores são tão essenciais na simulação e no processamento de dados experimentais e observacionais que muitas vezes fica difícil traçar uma linha entre dado e análise (ou código) quando discutimos a curadoria de dados" (Goodman *et al.*, 2014, p.1), o que torna, na visão de Jeannette Wing (2006, p.33), a "interpretação de código como dados e de dado como código". Desta forma, vivemos cada vez mais "na era da descoberta de conhecimento a partir de dados baseada em agentes", como resumiram Batista *et al.* (2022, p.1). No centro desse fenômeno está a disponibilização de **metadados acionáveis por máquina**, que fornecem informações contextuais essenciais para a interpretação e reutilização dos dados em diferentes espaços de significação.

Os cenários promissores engendrados pela eScience postulam a essencialidade do acesso, compartilhamento e integração dos dados que são produzidos e consumidos por seus empreendimentos. Para que isso seja possível, é fundamental que os conjuntos de dados sejam devidamente **autodescritos** para que tanto os programas de computador quanto as pessoas possam compreendê-los e analisá-los em vários outros contextos diferentes daqueles em que foram originalmente criados**.** "Em alguns casos, os avanços vêm da análise de fontes de dados existentes de novas maneiras – "o pentaquark foi encontrado nos arquivos assim que o teórico nos disse o que procurar", exemplificam Gray e Szalay (2004, p.3).

Parece claro que a **representação** está no centro da discussão da eScience e é a base que fundamenta os sistemas sociotécnicos de informação, cujo pressuposto é encapsular as camadas que tornam o conteúdo do objeto digital de pesquisa interpretável por provedores automáticos de serviços e ferramentas de análise. Assim sendo, é necessário conhecer objetos

digitais e suas diversas faces e os papéis dos metadados (também objetos digitais), o que será visto daqui por diante.

2.1 Objeto Digital de Pesquisa como Objeto Epistêmico

"Muitos objetos da ciência […] foram criados para gerar conhecimento. Podem ser instrumentos de observação ou medição; eles próprios podem ser objetos de estudo, como amostras ou espécimes; ou podem ser **representações** ou **modelos**" (Tybjerg, 2017, p. 269, grifo nosso). Todos esses objetos são chamados de "**objetos epistêmicos**", no sentido de que eles têm um grande potencial para **gerar conhecimento**. O conceito de "objetos epistêmicos" baseia-se no trabalho de Hans-Jörg Rheinberger e de seu conceito de "coisas epistêmicas" e de instalações experimentais, anunciadas em seu notável livro publicado em 1977. No escopo do presente ensaio, o objeto epistêmico ou objeto de conhecimento, como representação abstrata, é o foco de atenção. Para isso, nos valemos do aporte teórico de Rheinberger (1977), posto que ele coloca **a representação** no cerne do empreendimento científico como sistema de significantes, interpretação e (des)contextualização. É esta abordagem que nos ajuda a compreender o papel crucial dos metadados no escopo da eScience.

Nos domínios experimentais, a representação pode ser considerada equivalente a trazer coisas epistêmicas para o existente – este é também o nosso percurso. Ao inspecionar mais de perto o sistema experimental, Rheinberger (1977) distingue dois elementos fundamentais: o primeiro ele chama de "objeto de pesquisa - o objeto científico ou a 'coisa epistêmica'. Isso compreende entidades ou processos materiais – estrutura física, reações químicas, funções biológicas – que constituem o objeto de investigação" (Rheinberger, 1977, p. 28). Em poucas palavras: **a coisa epistêmica incorpora aquilo que ainda não é conhecido** - que constitui o objeto da investigação. O segundo elemento – chamado de "**objeto técnico**" - é o conjunto de condições experimentais onde os objetos de pesquisa estão inseridos. É por meio desse arranjo "que os objetos de investigação se tornam entrincheirados e articulados entre si em um domínio mais amplo de prática epistêmica e cultura material, incluindo instrumentos, dispositivos de inscrição, modelos de organismo, e os teoremas flutuantes ou conceitos de fronteira ligados a eles" (Rheinberger, 1977, p.29).

É por meio dessas condições técnicas que **o contexto institucional** passa para o trabalho de bancada, enfatiza o autor. Isso acontece em termos de instalações de medição locais, fornecimento de materiais, tradições de pesquisa, fluxo de trabalho de laboratório e habilidades acumuladas por pessoal técnico por longos períodos. A diferença entre condições experimentais (objeto técnico) e coisa epistêmica é, portanto, **funcional** em vez de **estrutural**. "As condições técnicas determinam o campo de **possíveis representações** de uma coisa epistêmica; e a coisa epistêmica suficientemente estabilizada se transforma em repertório técnico de arranjo experimental", como Rheinberger resume em seus argumentos (Rheinberger, 1977, p.29, grifo nosso). Portanto, um objeto epistêmico pode ser considerado uma máquina geradora de perguntas, enquanto o produto técnico é uma máquina de responder perguntas.

Rheinberger (1977) também nos esclarece sobre o papel que os objetos epistêmicos desempenham no espaço de representação criado nas atividades científicas, trazendo a ideia de descontextualização que se torna um conceito relevante no âmbito da curadoria de objetos de pesquisa. "O que há de significativo na representação enquanto inscrição é que as coisas podem ser representadas fora de seu contexto original e local e inseridas em outros contextos. É o tipo de representação que importa", afirma Rheinberger (1997, p.106). Wouters (2006, p. 11) destaca o interesse especial de muitos projetos de eScience, cujo cerne visa a descontextualização de objetos e posterior contextualização direta e em qualquer contexto. Ele pergunta e ao mesmo tempo responde: "Como isso é possível?" (Wouters, 2006, p.11). Isso é possível por meio de **metadados** que devem descrever o significado do objeto de pesquisa para que outras máquinas e humanos possam fazer (re)uso desses objetos em contextos impensáveis

no momento da produção do objeto. “Metadados são representações do contexto original de objetos epistêmicos que permitem que novos contextos possam ser criados para que esses objetos gerem novas questões”, comenta Wouters (2006, p.11). Para que isso aconteça é necessário agregar camadas de representação em diferentes espaços de significação por audiências inesperadas.

Os requisitos dos sistemas de informação de pesquisa atuais pressupõem a ideia de **interpretação de conteúdo por seres humanos e sistemas**. A representação da informação sustenta o potencial interpretativo dos objetos de pesquisa em contextos diferentes e novos, ou em novos espaços de significação, o que pode ser explicado pelo constructo teórico de Rheinberger (1977) de objeto epistêmico e objeto técnico. Contudo, o modelo pioneiro de arquitetura de objetos digitais, proposto em 1995 por Robert Kahn e Robert Wilensky, era agnóstico no sentido de não considerar os conteúdos que transportava, porém seus elementos continuam subjacentes às arquiteturas mais avançadas, como o FDO – FAIR Digital Object Architecture (Santos, 2020). É o que será visto a seguir.

2.2 Objeto de Dados como Objeto Informacional

Os padrões - incluindo os protocolos da Internet - são formas comuns de conhecimento codificado que circulam entre comunidades para garantir uniformidade e semelhança em processos ou produtos através do espaço e do tempo (Lischer-Katz, 2017*)*. É o caso do OAIS – -, uma norma internacional ISO (ISO 14721), que estabelece uma relação técnica e aplicável entre objeto de **dado**, objeto **informacional** e **conhecimento,** que aplicaremos para compor nossa proposição.

No contexto do Modelo OAIS, **a informação** é “definida como qualquer tipo de **conhecimento** que pode ser **intercambiado**, e essa informação é sempre expressa (ou seja, representada) por algum tipo de **dado**” (CCSDS, 2012, p. 2-3). Essa definição exige que o destinatário dos sinais ou padrões (ou seja, dados) seja capaz de decodificá-los e entender o que é comunicado; isso requer que o destinatário da mensagem – seja humano ou sistema - tenha conhecimento contextual e tácito adequado para decodificar os sinais, símbolos ou padrões, e então entender a mensagem que eles representam. Assim, uma vez recebida a mensagem, é necessário um certo nível de conhecimento para processá-la, interpretá-la e entendê-la. O Modelo OAIS utiliza o conceito de “base de conhecimento” para qualificar esse tipo de conhecimento. Mais formalmente, pode-se dizer que uma pessoa ou sistema – por exemplo, um agente computacional – possui uma **Base de Conhecimento**, que permite compreender as informações recebidas. Se um destinatário ainda não possui conhecimento suficiente para entender as informações, os dados precisam ser acompanhados de informações de representação – isto é, as informações que mapeiam os dados em conceitos mais significativos e/ou engendram uma contextualização que lhes conferem significado – de uma forma que seja compreensível usando a base de conhecimento do destinatário. Esta categoria de informação está incluída no processo de comunicação e pode ser, no domínio da pesquisa científica, por exemplo, um livro de códigos, um dicionário, um caderno de laboratório ou de campo, um projeto, um manual, anotações e mais uma infinidade de documentos. Nesse sentido, **dados** quando interpretados usando sua **Informação de representação** produzem **um objeto informacional** (CCSDS, 2012), mais formalmente: o **Objeto Informacional** é composto por um **Objeto de Dado** – que pode ser físico ou digital – e pela **Informação de Representação,** que permite a interpretação completa do dado.

Por exemplo, a saída de um instrumento científico digital é expressa pela sequência de bits (os dados) que representa, neste exemplo, uma tabela ASCII de números; quando esses bits são combinados com informações de representação, eles são convertidos em informações mais significativas, como números que fornecem as coordenadas de um local na Terra medido em

graus de latitude e longitude. Para transformar a sequência de bits em informação significativa, a Informação de Representação deve conter dois tipos de informação: a primeira descreve o formato, ou conceitos de estrutura de dados, que serão aplicados às sequências de bits e que por sua vez resultam em mais valor de significado, como caracteres, números, gráficos, *arrays*, tabelas, visualização, para citar alguns. Este tipo de informação é referido como **Informação estrutural** do objeto informacional de representação. Mas raramente são suficientes, em muitos casos é necessário um segundo tipo de informação: esta informação adicional é referida como **Informação Semântica**. Por exemplo, onde o Objeto Digital é descrito como uma sequência de caracteres de texto, deve ser fornecida a informação adicional relativa a qual idioma ele estava sendo expresso. Assim, o objetivo da informação de representação do objeto é converter a sequência de bits em informações mais significativas.

Há um percurso importante percorrido para que essas ideias pudessem ser incorporadas às arquiteturas e modelos de objetos digitais. Para que se possa compreender com a amplitude necessária o conceito de objeto de pesquisa, é necessário primeiro compreender as ideias que o antecederam e as que vão determinar o futuro dos sistemas de informação para a pesquisa. É o que será visto a seguir.

2.3 Objeto Digital como Arquitetura

O conceito de objeto digital foi introduzido por Robert Kahn e Robert Wilensky em um artigo clássico publicado em 1995 – *A framework for distributed digital objects service* – que foi reimpresso em 2006. Os autores descrevem o "aspecto fundamental de uma infraestrutura que é aberta em sua arquitetura e que suporta uma grande e extensível classe de **serviços de informação digital distribuída**" (Kahn; Wilensky, 2006, p.1). Eles também definem as entidades básicas que devem estar presentes nesse sistema, no qual as informações, na forma de objeto digital, são armazenadas, acessadas, disseminadas e gerenciadas. O modelo estabelece ainda convenções de nomes para identificar e localizar objetos digitais, bem como descrevem um serviço para usar nomes de objetos para localizá-los e disseminá-los, e um protocolo de acesso (Kahn; Wilensky, 1995, 2006).

No caminho histórico, conceitual e técnico iniciado por Kahn e Wilensky , os elementos estruturais por eles delineados são atualmente colocados no contexto dos Princípios Orientadores FAIR, que visam tornar dados localizáveis, acessíveis, interoperáveis e reutilizáveis. Esses princípios assumem globalmente um papel proeminente como um arcabouço para a sustentabilidade dos dados de pesquisa e para a construção apropriada de sistemas de curadoria. Além disso, a abordagem FAIR sempre considera a ideia de "acionabilidade por máquina", entendida como a capacidade de sistemas computacionais realizarem serviços sobre os dados sem intervenção humana (De Smedt; K*oureas; Wittenbuger*, 2020; Schwardmann, 2020). Este cenário parece possibilitar a realização de uma Internet FAIR de Dados e Serviços (IFDS), cujo ponto central é o conceito *de FAIR Digital Object* (FDO), em português, Objeto Digital FAIR - um tipo de Objeto Digital que está no contexto *FAIR Digital Object Framework* (FDOF). O FDOF, como o próprio nome diz, é um *framework* que define um modelo para **representar objetos** em um ambiente digital, e um conjunto de recursos para fornecer suporte fundamental para os princípios FAIR (Santos, 2020).

Um Objeto Digital FAIR (FDO) é definido formalmente por Luiz Olavo Bonino da Silva Santos (2020) como uma **sequência de bits** que representa uma unidade de informação acionável por máquina, **identificada** por um identificador globalmente único, persistente e resolvível com comportamento de resolução previsível, descrito por registros de **metadados** – que também são Objetos Digitais FAIR -, e classificados pelo **sistema de tipagem FDOF**. A partir dessa perspectiva, um FDO é uma unidade acionável estável que agrupa informações suficientes para permitir a **interpretação** e **o processamento confiáveis** dos dados nele contidos.

O modelo pioneiro de arquitetura de objetos digitais proposto em 1995 por Robert Kahn e Robert Wilensky - que tinha como objetivo fundamentar uma rede de bibliotecas digitais de relatórios de computação (Sayão, 2009) - era agnóstico no sentido de não considerar os conteúdos que transportava. No artigo, Kahn e Wilensky (1995) enfocam os aspectos de rede da infraestrutura, "ou seja, aqueles para os quais o conhecimento do **conteúdo do objeto digital não é necessário.** A definição dos aspectos baseados em conteúdo da infraestrutura não é propositalmente abordada [...]", confirmam os autores (Kahn; Wilensky, 1995, p.118). Em contraste, os requisitos dos sistemas de informação de pesquisa atuais pressupõem a ideia de **interpretação de conteúdo**, nessa direção, o FDO é uma unidade acionável estável que agrupa informações suficientes para permitir a interpretação e o processamento confiáveis dos dados nele contidos (De Smedt; K*oureas; Wittenbuger*, 2020). Essa característica coloca a **representação** no centro da discussão da eScience e é a base que fundamenta os sistemas tecno-sociais de informação, cujo pressuposto é encapsular as camadas que tornam o conteúdo do objeto digital interpretável por provedores automáticos de serviços e ferramentas de análise. Compreendida a ideia de Objeto Digital, tem-se elementos suficientes para a compreensão do que vem a ser um Objeto Digital de Pesquisa, na esfera da eScience.

3 OS OBJETOS DIGITAIS DE PESQUISA NA ESFERA DA ESCIENCE: O BIG METADADO

Nesses ambientes científicos digitalmente avançados, grande parte das ações sobre os objetos de pesquisa, como análises e visualização, são raramente performadas sobre os objetos propriamente dito, mas sobre abstrações e representações apropriadas aos objetivos da pesquisa, como modelos e representações gráficas, ou sobre as representações na forma de metadados, apoiados por instrumentos semânticos, como vocabulários controlados e ontologias. Assim sendo, observa-se que a pesquisa contemporânea está atrelada a diferentes modos de abstração e representação dos objetos de pesquisa e de suas relações contextuais. Isto coloca em pauta a necessidade de se inserir nas infraestruturas de pesquisa modelos de representação de objetos de pesquisa baseados em metadados ricos e semanticamente bem estruturados, que sejam interpretáveis por seres humanos e máquinas. Esta condição cria uma conexão essencial entre o *big data* científico e o **fenômeno do big metadado**,

Conforme discutido anteriormente, a ciência de dados empreende motivada pela excepcional disponibilidade de dados digitais e de novas competências computacionais que viabilizam soluções baseadas em dados. Essas ideias são também centrais para a realização do quarto paradigma científico, conforme preconizado por Jin Gray (Hey; Tansley; Tolle, 2009) para explicar as oportunidades sem precedentes que se vislumbra para a ciência orientada por dados. Na sua concepção, os **metadados são componentes vitais** para a realização da eScience, embora o significado dos metadados seja frequentemente negligenciado ou interpretado de forma limitada, na medida em que é considerada somente a sua face meramente descritiva ou "apenas dado sobre dado", deixando oculta toda a sua complexidade representacional, semântica e funcional. Não obstante, "Nessa nova ecologia informacional, os metadados podem atrair um novo olhar da pesquisa se forem entendidos como **big metadado"**, enfatiza Jane Greenberg (2017, p.25).

Além de uma associação com a diversidade e o tamanho do *big data*, o big metadado reflete a ampla gama de atividades de ciclo de vida de dados encontradas entre projetos, configurações e sistemas. A concepção de metadados como dados estruturais que suportam funções associadas a um objeto digital, e a escala, diversidade e complexidade dessas funções dependerão da natureza intrínseca do objeto digital, do ambiente onde ele está inserido – por exemplo, negócios, governo ou pesquisa científica – e das idiossincrasias de seu ciclo de vida. Porém, é, no nível meta do ciclo de vida dos dados, que está o ciclo de vida dos metadados, que gera o big metadado (Grenberg, 2017).

Contextualizando o fenômeno do big metadado a partir do exemplo de uma grande plataforma de dados, como o GenBank, que está habilitada por uma ciberinfrastrutura avançada de pesquisa, cujos processos produzem enormes quantidades de dados e, associados a esses dados, geram também uma grande quantidade de metadados, "posto que cada *dataset* [produzido] inclui seus próprios metadados, nos agora temos não somente um *big data* científico mas também um *big metadado"* (Bratt *et a*l., 2017, p.1). Assim como os dados de pesquisa são essenciais para a ciência contemporânea, big metadados agregados a esses objetos de pesquisa são também essenciais para a descoberta de outros dados, compartilhamento, reutilização, crédito de autores e reprodutibilidade de pesquisa e, mais recentemente, com o surgimento das metodologias e ferramentas da ciência de dados, o big metadado está sendo usado como objeto das *analytics* e oferecem *insights* e direcionamentos importantes no âmbito dos empreendimentos científicos.

Como parte da história recente da ciência de dados, já no início da década de 2010 começa aparecer na literatura o termo "big metadado" como um desafio que se instala no âmbito do *big data*. Nesse contexto, Smith e colaboradores (2014, p.1) já alertavam que os ecossistemas de *big data* da época careciam de uma abordagem que considerasse os princípios de **gestão de metadados: "**Na maioria dos casos, os ecossistemas de *big data* surgiram sem um tipo de suporte para gerenciamento de metadados amplamente reconhecido como essencial em sistemas empresariais tradicionais". Isso se tornou um obstáculo para que grandes organizações compartilhassem dados e códigos de preparação e análise de dados, para integrar dados e assegurar que códigos analíticos fizessem suposições compatíveis com os dados que foram utilizados. Esse problema se torna presente também no big dado científico, revelando os contornos de um **big metadado científico**, onde a gestão de metadados deve se tornar parte essencial do ciclo de vida da curadoria dos dados e de sua fairificação.

Assim sendo, o big metadado no domínio científico pode ser identificado como a disponibilidade massiva de metadados de diversas categorias e tipos que são agregados aos dados e a outros objetos de pesquisa e que desempenham várias funcionalidades que são relevantes para as diversas vertentes do ciclo de vida dos dados, como recuperação, interoperabilidade, reúso e diversos tipos de análise. Essa complexidade, intensificada pela eScience, implica na necessidade de metodologias e ações voltadas para a organização, classificação e esquematização dos elementos de metadados chamadas coletivamente de **gestão de metadados**.

O quadro 1 abaixo é uma interpretação destes autores do esquema já clássico dos 5Vs, proposto inicialmente por Marr (2014) para o *big data*, aplicados ao fenômeno do big metadado que se incorpora aos desafios da ciência orientada por dados.

Quadro 1. 5 Vs aplicados ao big metadado científico

V	BIG METADADO CIENTÍFICO
VOLUME	O grande volume de metadados para descrever e registrar os processos científicos, bem como as inúmeras funcionalidades desempenhadas por eles no decorrer do ciclo de vida da pesquisa, confirmam a existência do big metadado. Algumas vezes, o volume de metadados é menor ou igual à extensão dos dados que eles descrevem; outras vezes, devido à complexidade das atividades do ciclo de vida dos dados, os metadados excedem o tamanho dos dados que estão sendo descritos ou rastreados.
VELOCIDADE	No âmbito dos experimentos científicos, os metadados são gerados por processos automáticos por meio de instrumentos, sensores remotos, por códigos ou por ferramentas de *workflow* em grande velocidade. Cada vez mais estão presentes nas bancadas dos laboratórios, plataformas automatizadas para a coleta/geração organização e análise de dados e metadados. Há, entretanto, os elementos de

V	BIG METADADO CIENTÍFICO
	metadados assinalados por meios intelectuais por pesquisadores, informáticos e profissionais de informação.
VARIEDADE	A **variedade** dos metadados reflete a ampla diversidade de formatos, ciclos, modelos, estrutura e tipos de metadados presentes no universo científico. Há uma desigualdade clara, porém natural, no ecossistema de metadados que registram os diversos procedimentos, sistemas e processos dos laboratórios. Por exemplo: metadados descritivos, administrativos, técnicos, estruturais, de preservação, de processamento, que vão além da descrição e desempenham diversas funções, como recuperação, acesso e interoperabilidade. A demanda por metadados específicos aplicados a domínios disciplinares intensificam essa variedade, que é ampliada pelo extenso e diversificados tipos de ciclos de vida de dados e metadados, que mesmo utilizando o mesmo padrão/esquema de metadados têm práticas diferentes de implementação.
VERACIDADE	A **veracidade** do big metadado científico é estabelecida pelas boas práticas para o assinalamento dos metadados como forma de agregar valor aos dados e a outros objetos de pesquisa; contribuem fortemente para a veracidade, a proveniência dos esquemas de metadados, seu nível de padronização, a conexão com os processos científicos, com as práticas e idiossincrasias das comunidades disciplinares, a completude dos elementos de metadados e a precisão dos instrumentos terminológicos – vocabulários, taxonomias, ontologias.
VALOR	O **valor** final do big metadado está no seu apoio à interpretação apropriada dos dados e no reúso em distintos espaços de significação, ao longo do tempo e do espaço, por seres humanos e por agentes computacionais. Isto gera valor também para os dados. O valor do big metadado está também associado ao seu papel como matéria prima primária para os métodos de análise, ou seja, os metadados são considerados não como representação de conteúdos, mas como o próprio objeto de análise. O valor dos metadados está associado ainda a sua qualidade, medida por parâmetros, tais como: granularidade, *timeliness*, acurácia, completude e proveniência (meta-metadados).

Fonte: Elaborado pelos autores, inspirado na estrutura proposta por Marr (2014).

São várias as dimensões relevantes que exteriorizam a interface entre a eScience e a CI. Nesta seção, pôde-se pautar algumas conexões mais diretamente relacionadas ao nosso estudo: o fenômeno do big metadado científico e de sua gestão, e as funcionalidades ampliadas dos metadados necessárias para a realização das ações da ciência dos dados sobre objetos digitais de pesquisa como, por exemplo, a acionalidade por programas. Entretanto, outras conexões também podem ser feitas se olhadas sob outras perspectivas, é o que vai ser visto a seguir.

4 A SINERGIA ENTRE A eSCIENCE E A CIÊNCIA DA INFORMAÇÃO: O PROTAGONISMO DOS METADADOS

Além do fenômeno do Big Metadado e sua gestão, percebe-se a necessidade de inclusão também dos novos modelos em rede de socialização e de conceitos de veículos de comunicação científica como essenciais para a construção do pilar da cooperação da eScience. Nesse ambiente, "algumas comunidades científicas já estão experimentando novas formas de representação de conhecimento, tais como as **nano-publicações**, que são basicamente declarações em alguma forma de linguagem semântica como RDF, ampliada por metadados suficientes", confirmam De Smedt, K*oureas* e Wittenbuger (2020, p.14), colocando ênfase no

metadado como elo de conexão entre a eScience e a CI. Observa-se, porém, que essa perspectiva vem se consolidando desde os primórdios da eScience, conforme introduzida por Jim Gray.

A partir da perspectiva pragmática conferida pelo seu trabalho como cientista da computação, atuando no campo da Astronomia Virtual, Jim Gray e colaboradores (2002) constataram há décadas o que parece hoje ser essencial para a curadoria: "os dados são incompreensíveis e, portanto, inúteis, a menos que haja uma descrição detalhada e clara de como e quando foram coletados e como os dados derivados foram produzidos" (Gray *et al*., 2002, p.5). Para tanto, os dados devem ser cuidadosamente documentados e publicados de forma que permitam fácil acesso e **processamento automático,** abrindo dessa forma a possibilidade de que ferramentas computacionais genéricas e pessoas possam entendê-los e reutilizá-los. Assim sendo, agregar informações a um objeto de pesquisa digital torna-se a principal responsabilidade da curadoria digital. Essas adições e associações ocorrem em todos os pontos do ciclo de vida da curadoria, corrobora Hunter (2006)**.**

Na percepção de Gray, um conjunto de dados científicos de valor contínuo, uma vez publicado, deve permanecer disponível **para sempre**, dando suporte a uma escala variada de reprodutibilidade e verificabilidade e a novas descobertas. No entanto, é preciso compreender que os pesquisadores que examinarão esses dados mais tarde não saberão, explicitamente, os detalhes de como os dados foram coletados e processados. Para entender os dados, esses pesquisadores precisarão saber "(1) como os instrumentos foram projetados e construídos; (2) quando, onde e como os dados foram coletados; e (3) uma descrição cuidadosa das etapas de processamento que levaram aos produtos derivados finais**"** (Gray *et al.,* 2002, p.1). Os autores destacaram ainda que esses produtos derivados são os principais objetos de **investigação e análise científica** dos **dados**. Sendo assim, para interpretar os dados, no cenário atual e futuro, os pesquisadores precisam de informações expressas principalmente por **metadados.** Essas autodescrições agregadas aos dados são centrais para todos os cenários postulados pela eScience.

Do ponto de vista prático e cientificamente contextualizado de Jim Gray e de seus colaboradores, os metadados podem ser entendidos como: "a informação descritiva sobre os dados que explica os atributos mensurados, seus nomes, unidades, precisão, acurácia, layout dos dados e, idealmente, muito mais" (Gray *et al*., 2005, p.3). Os autores ainda enfatizam que os metadados mais importantes devem registrar a **linhagem dos dados** que descrevem, como os dados foram medidos, adquiridos ou computados. Estendendo as características dos metadados, Gray observa que metadados de qualidade se tornam centrais para o compartilhamento interdisciplinar de dados e para as ferramentas de análise e visualização. Os metadados devem, idealmente, registrar tudo o que deve ser de interesse do pesquisador, incluindo modelos de dados, equipamentos especiais, especificação de instrumentação, linhagem de dados e muito mais, consolidam Jim Gray e colaboradores (2002).

Para além das questões mais tecnológicas, os metadados desempenham um papel importante em pautas sensíveis para o ciclo de comunicação científica, como a revisão por pares e reprodutibilidade dos experimentos. Isto porque os materiais da eScience têm sido objeto de escrutínio em relação às principais questões que abordam o processo de integridade e ética científica. Por exemplo, a reprodutibilidade de muitos experimentos pode não ser viável por vários motivos, incluindo as dificuldades de se reconstruir os aparatos e os ambientes experimentais. Portanto, um exame minucioso dos conjuntos de dados, documentados por metadados ricos e publicados, pode ser exigido por periódicos acadêmicos para apoiar a presunção de reprodutibilidade. Além disso, o valor de fornecer conjuntos de dados acompanhados de informações metodológicas claras e transparentes usadas nos experimentos, é essencial para manter a ética profissional e a integridade das pesquisas e de suas conclusões (Bohle, 2013).

Para que os metadados protagonizem esses novos papéis no contexto da ciência contemporânea, que vão muito além da descrição bibliográfica – que continua essencial - é

necessária uma ampla variedade de tipos, propriedades e funções de metadados que precisam estar organizados para fundamentar a construção de esquemas que atendam a requisitos disciplinares. É o que será visto a seguir.

5 FUNCIONALIDADE DOS METADADOS: MUITO ALÉM DA DESCRIÇÃO BIBLIOGRÁFICA

No âmbito da pesquisa científica, metadado é um termo guarda-chuva que designa as informações estruturadas e atributos que, adicionados aos dados, conferem a eles proveniência, contexto e potencial de compreensão e interpretabilidade, elementos que são críticos para ampliar as possibilidades de reúso dos dados. Visto dessa forma, metadado pode ser entendido como um meio de adicionar valor ao dado de pesquisa, ampliando o seu potencial de conduzir informação e conhecimento no espaço e no tempo. "Além de rotulagem e de categorização, os metadados podem ser pensados mais universalmente como uma linguagem de valor agregado que serve como uma camada integradora em um sistema de informação", completa Jane Greenberg (2017, p.22). Essa abstração conecta o objeto de pesquisa a um conjunto de funcionalidades importantes, que apoiam o gerenciamento de dados no contexto de um sistema de informação – como é, por exemplo, um repositório digital -, como identificação, recuperação, preservação, níveis de contextualização e proveniência, ou ainda permissões de reúso.

Por seu lado, os objetos digitais de pesquisa são caracterizados por terem um ciclo de vida complexo e longo, que depende de diferentes contextos disciplinares e perspectivas de (re)uso em domínios variados. Este ciclo de vida começa antes do início da pesquisa e se estende indefinidamente para além do final do projeto, quando os dados precisam ser arquivados por longo prazo em sistemas confiáveis. Ao longo desta jornada, vários tipos de metadados são adicionados aos objetos, atribuídos por diferentes partes interessadas, incluindo aqueles gerados automaticamente por instrumentos científicos (Wittenburg *et al*., 2018) e por ferramentas de fluxo de trabalho laboratorial – software de *workflow*-, em um processo contínuo de agregação de valor aos conjuntos de dados e a outros objetos de pesquisa. Esses metadados, idealmente, devem ser compreendidos tanto por humanos quanto por computadores. Especificamente, no que concerne aos recursos compatíveis com FAIR, dados, metadados e serviços devem atender aos requisitos de serem acionáveis por máquina sem supervisão humana, sempre que possível, principalmente para atingir os objetivos de uma Internet FAIR de Dados e Serviços (Mons *et al.*, 2017). Por conseguinte, esses objetos digitais de pesquisa requerem uma ampla gama de metadados - com muitas funções e propriedades - que **muitas vezes** superam os próprios dados em volume e até em importância, configurando, em algumas situações, um big metadado. Fica claro, portanto, que para os objetos digitais de pesquisa, a função dos metadados vai muito além dos processos descritivos aplicados a documentos impressos convencionais, sintetizados vagamente pela frase "metadados são dados sobre dados". As funções associadas aos metadados de um objeto digital de pesquisa e a sua escala, diversidade e complexidade dependem da natureza do objeto digital, do ambiente onde ele está inserido, das suas componentes estruturais e semânticas e das peculiaridades disciplinares que determinam o seu ciclo de vida.

Em face das condições de contorno exigidas pela eScience, é essencial se estabelecer diretrizes para a construção, gestão e aplicação de metadados ricos que sustentem a reprodutibilidade e reúso dos objetos de pesquisa. Isto é especialmente relevante em domínios disciplinares específicos e verticalizados onde é crítico se descrever experimentos complexos, que envolvem múltiplos processos, que precisam ser altamente contextualizados. Esse fato ganha ainda uma nova perspectiva no âmbito da ciência aberta, onde não somente os resultados finais – dados e publicações - devem ser descritos, mas também todo o aparato para atingí-los,

como modelos, códigos computacionais, algoritmos, métodos de laboratório, equipamentos, *workflow*, etc. que, por fim, têm um caráter muito disciplinar.

Ross Harvey (2010) alinha algumas das informações que o curador de dados e vários outros agentes – incluindo instrumentos científicos e software, como sistemas de *workflow* de laboratórios – vão agregando aos dados durante o seu ciclo de vida na forma de metadados e documentação. Essas informações permitem que os dados possam ser efetivamente gerenciados, acessados e reusados, agora e no futuro. A partir deste ponto, propõe-se, como resultado deste ensaio, uma sistematização das funcionalidades dos metadados, para os objetos digitais de pesquisa (Quadro 2).

Quadro 2. Funcionalidade dos metadados

CATEGORIA	FUNÇÃO
REPRESENTAÇÃO	**Codifica e** contribui para **organizar o conhecimento** de um domínio disciplinar, de uma forma que é relevante para o campo de pesquisa e é familiar para a comunidade de pesquisa.
	Transforma objetos de dados em **objetos de informação** (CCSDS, 2012) ou em **objetos epistêmicos** (Rheinberger, 1977).
	Mostra o que é necessário para **representar no objeto de pesquisa,** no padrão exigido pelos usuários (HARVEY, 2010).
DESCRIÇÃO	**Identifica** o objeto de pesquisa de forma unívoca, global e persistente (considerando o identificador persistente como parte dos elementos do esquema de metadados).
	Localiza o objeto de pesquisa.
	Esclarece o que é o objeto de pesquisa.
	Registra a **informação bibliográfica** sobre o objeto de pesquisa permitindo que ele seja **referenciado e citado** segundo os padrões pertinentes; confere **crédito aos diversos autores** dos objetos.
	Aponta as **propriedades e estruturas técnicas** dos elementos que compõem o objeto digital de pesquisa.
GESTÃO	Apoia a **gestão** completa do **ciclo de vida** do objeto de pesquisa.
	Registra a **proveniência do esquema de metadados** (informado pelos metametadados) e o grau de gestão aplicados ao esquema corrente.
RECUPERAÇÃO	Apoia a **formulação de consultas** com nível apropriado de granularidade e precisão tomando em conta as características disciplinares.
	Apoia a **encontrabilidade** e o **acesso** aos objetos digitais.
	Apoia a **seleção** e a **avaliação** dos objetos recuperados.
RELAÇÃO	Mantém **links confiáveis** para o objeto.
	Fornece **link** do objeto de pesquisa para outros objetos relacionados (artigos de periódicos, software, datasets, etc.) para tornar visível o

CATEGORIA	FUNÇÃO
	ecossistema **onde o objeto está localizado e o seu relacionamento com outros objetos de pesquisa.**
INTERPRETABILIDADE	Amplia o **nível de interpretabilidade** dos objetos de pesquisa em diferente espaço de significação para humanos e agentes computacionais, agora e no futuro.
	Facilita o **reúso** dos objetos de pesquisa pelos seus próprios criadores e por outros pesquisadores.
	Apoia a **interlocução com outras coleções e com diferentes sistemas** por meios automatizados.
PROVENIÊNCIA E CONTEXTUALIZAÇÃO	Registra a **história do objeto de pesquisa** (proveniência, rastreabilidade e linhagem).
	Relata **os processos, parâmetros, variáveis, metodologias, códigos e instrumentos** que foram relevantes para coletar/gerar, processar e analisar o objeto de pesquisa.
QUALIDADE	Indica as **ações de garantia e controle de qualidade** aplicados aos objetos (exemplo, flags para dados faltantes, discrepantes etc.).
INTEROPERABILIDADE	Habilita a **acionabilidade por máquina** dos objetos de pesquisa.
REVISÃO POR PARES	**Informa os revisores de artigos convencionais e de artigos de dados** sobre os processos de obtenção, processamento, análises níveis de qualidade e o potencial de reprodutibilidade dos objetos de pesquisa.
PRESERVAÇÃO	Apoia as estratégias de **preservação de longo prazo.**
	Informa as **dependências técnicas** do objeto de pesquisa.
	Fornece **informações de representação** semântica e estrutural que permite interoperabilidade com o futuro. A reconfiguração do objeto no futuro.
	Compõe os **pacotes de informação de arquivamento** (AIP/OAIS).
CONFIANÇA	Apoia a presunção de **autenticidade e confiabilidade** dos objetos de pesquisa.
	Registra as ações aplicadas para garantir a **integridade** dos objetos, como *hash* e *chercksun*.
PERMISSÕES	Anuncia as **licenças** associadas ao objeto.
	Informa sobre o **grau de sensibilidade** do objeto.
	Informa sobre as **permissões de acesso.**
	Aponta as **ações que podem ser realizadas** sobre o objeto de pesquisa.
COPYRIGHT	Informa sobre os direitos de **propriedade intelectual** associados ao objeto.

CATEGORIA	FUNÇÃO
ANOTAÇÃO	Agrega comentários colaborativos **de outros pesquisadores** diferentes dos autores.
ADMINISTRAÇÃO	Identifica as **pessoas/equipes que operacionalizam** o ciclo de vida dos dados: quem coleta, quem indexa, quem garante a segurança física etc.
	Identifica os ***stakeholders* envolvidos**, por exemplo: financiadores, instituições parceiras.
	Identifica quem é o **responsável pela gestão e preservação** dos objetos de pesquisa.

Fonte: Elaborado pelos autores (2023).

Acredita-se que esta estrutura pode contribuir para a gestão de metadados que torne objetos digitais de pesquisa em objetos informacionais ou epistêmicos. No entanto, para desempenhar as suas funções, os elementos de metadados precisam ainda de outra camada de representação que corresponde à codificação dos conteúdos por meio de instrumentos/esquemas terminológicos padronizados, como vocabulários controlados, tesauros, ontologias, taxonomias e outras estruturas classificatórias. Estas instâncias estão fora do contexto do presente artigo, mas também se colocam como uma oportunidade de estudo no âmbito da Ciência da Informação, especialmente no contexto da Organização do Conhecimento, aproximando mais uma vez essa área da eScience.

6 CONCLUSÃO

O quarto paradigma científico, concretizado pela eScience, é, em pouquíssimas palavras, uma abordagem metodológica que leva estes autores a *insigths* e a novas descobertas baseadas em processos avançados de integração e análise da abundância de dados. Conforme observa Newman (2019), isto não é propriamente novo, na medida em que dados crescentemente complexos e em grandes quantidades podem ser considerados parte do impulso para o empiricismo na ciência do século XIX e do início do século XX. “O que é novo - e ainda não claro – é como este novo paradigma transformará os modos fundamentais de pesquisar e atuar na ciência e tecnologia do século XXI” (Newman, 2019, p.525). Gim Gray (Hey; Tansley; Tolle, 2009) sintetiza esta transformação, que ainda se corporifica de forma pragmática, afirmando que a eScience é onde a computação encontra os cientistas. Assim sendo, uma dimensão unificadora da eScience, essencial para a realização dos seus sonhos e promessas, é a infraestrutura de tecnologia da informação e comunicação que está à sua volta (Wouters, 2006). Ao que tudo indica, as transformações na ciência vão se suceder no ritmo vertiginoso da evolução da computação.

No fim desse ensaio, o que se conclui é que existe um papel fundamental da Ciência da Informação e da Biblioteconomia na revolução da descoberta computacional que está subjacente à eScience.

Como visto no decorrer do trabalho, o que conecta a eScience à Ciência da Informação são as várias dimensões relevantes da representação: epistemológicas, tecnológicas e sociológicas - remetendo à ideia do pensamento computacional sobre a abstração e ações que são performadas pelas diversas funções que estendem o conceito longamente consolidado de metadado.

Este percurso pode ser mensurado pela capacidade de interpretação e de acionalidade dos metadados por agentes computacionais no exercício de prover serviços informacionais

avançados, como análises integrativas e visualização. A interpretação automática é a chave para a internet do futuro: a Internet FAIR de Dados e Serviços, cujo elemento principal são os Objetos Digitais FAIR, dotados de uma arquitetura própria cuja chave são metadados ricos. Neste sentido, vale lembrar que FAIR, antes de tudo é metadados, basta uma rápida leitura dos princípios para perceber que o conceito de metadados está expresso em quase todos os 15 princípios.

No decorrer do presente estudo, pôde-se observar a confluência de ideias em torno da relevância da representação e da abstração para a realização dos empreendimentos da eScience. Simulações, modelagem, visualização, virtualização dos objetos de pesquisa físicos (como amostras, herbários etc.) algoritmos, assim como os metadados estão no coração do abstracionismo epistêmico dos processos de descoberta da eScience. Pelo lado da CI, questões referentes à representação sempre acompanharam o seu amplo espectro de estudo e aplicação, revelando-se fundamental também para a melhoria da comunicação entre pessoas, instituições e, mais recentemente, entre máquinas. No contexto específico da Biblioteconomia, metadados sempre foram elementos-chave para a catalogação e a indexação de livros e de outros documentos contidos nas bibliotecas. No contexto da Documentação, a partir do conceito de documento preconizado por Suzanne Briet (1951), em seu tratado *Qu`est-ce que la documentation?* foi colocada em pauta a importância da representação, para que um objeto qualquer pudesse se tornar um documento. Jin Gray na sua visão fundacional, pôde prever a importância da conexão entre eScience e CI, que hoje é intensificada ainda mais pelo big metadado: "Os astrônomos provavelmente reinventarão muitos dos **conceitos** já desenvolvidos nas **comunidades** de **bibliotecas** e **museus**. Os bibliotecários pensaram profundamente sobre essas questões e faríamos bem em aprender com a experiência deles" (Hey; Tansley; Tolle, 2009).

Os resultados práticos da presente pesquisa ampliam a perspectiva dos estudos sobre representação à medida que revela a sua importância dentro de um novo contexto informacional: a dos objetos digitais de pesquisa e de seu acúmulo de metadados, denominado de big metadado, que como visto no quadro 1, estabelece uma interlocução com os 5Vs do *big data*, quando eles próprios se tornam fontes para as novas metodologias de análise. O ensaio tenta ainda mostrar que os metadados passam a ser também objetos de gestão dentro das ciberinfraestruturas de pesquisa, consolidando a ideia de "gestão de metadados".

Neste sentido, à medida que a Ciência da Informação e a Biblioteconomia se convergem às áreas eScience, o protagonismo dos metadados se torna cada vez mais presente, gerando novos desafios que devem se configurar em estudos futuros, especialmente no que tange à construção de esquemas de metadados disciplinares FAIR que atendam as necessidades de representação e funcionalidades das ciências x-computacional e x-informática, localizadas no contexto da eScience. Assim, a presente pesquisa continua à medida que novos estudos estão sendo desenvolvidos sobre "autoria de metadados", vocabulários FAIR e compatibilização semântica e que devem ser publicados em breve.

REFERÊNCIAS

BATISTA, D. *et al*. Machine actionable metadata model. **Scientific Data**, London, v. 9, n. 1, 2022. Disponível em: https://go.nature.com/3CsMpd9 . Acesso em: 20 fev. 2023.

BOHLE, S. **What is E-science and how should it be managed?** 2013. Disponível em: https://bit.ly/3Je1raz . Acessso em: 04 jul. 2022.

BRATT, S. E. *et al*. Big data, big metadata and quantitative study of science: A workflow model for big scientometrics. **Proceedings of the Association for Information Science and**

Technology, v. 54, n. 1, 2017. Disponível em: https://bit.ly/43Ul9Qu . Acesso em: 04 jul. 2022.

BRIET, S. **Qu'est-ce que la documentation**. Paris: EDIT, 1951.

CCSDS - CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEM. **Reference Model for an Open Archival Information System (OAIS)**. Washington, DC: CCSDS, 2012. (Recommended Practice CCSDS 650.0-M-2. Magenta book). Disponível em: https://public.ccsds.org/pubs/650x0m2.pdf. Acesso em: 30 set. 2019.

DE ROURE, D. *et al*. The Semantic Grid: a future e-Science infrastructure. *In*: BERMAN, F.; FOX, G.; HEY, A. J. G. (ed.). **Grid Computing. Making the Global Infrastructure a Reality**. Chichester, West-Sussex, UK: John Wiley & Sons, 2003. p. 437-470.

DE SMEDT, K.; KOUREAS, D.; WITTENBUGER, P. FAIR Digital Objects for Science: From data pieces to actionable knowledge units. **Publications**, Basel, v. 8, n. 21, 2020. Disponível em: https://bit.ly/3CrUnU7. Acesso em: 06 jan. 2023.

GOODMAN, A. *et al*. Ten simple rules for the care and feeding of scientific data. **PLoS Computer Biology**, Bethesda, v. 10, n. 4, 2014. Disponível em: https://ury1.com/lB3fB. Acesso em: 29 jul. 2022.

GRAY, J. *et al*. **Online scientific data curation, publication, and archiving**. Redmont, WA: Microsoft Corporation, 2002. Disponível em: https://ury1.com/5KUT8. Acesso em: 29 jul. 2022.

GRAY, J. *et al*. **Scientific data management in the coming decade**. Redmont, WA: Microsoft Corporation, 2005. Disponível em: https://ury1.com/rAhgO. Acesso em: 19 jul. 2022.

GRAY, J.; SZALAY, A. **Where the rubber meets the sky**: bridging the gap between database and science. Redmont, WA: Microsoft Corporation, 2004. Disponível em: https://arxiv.org/abs/cs/0502011. Acesso em: 22 mar. 2023.

GREENBERG, J. Big metadata, smart metadata, and metadata capital: toward greater synergy between data science and metadata. **Journal of Data and Information Science,** Beijing, v. 2, n. 3, 2017. Disponível em: https://urx1.com/qcMN3. Acesso em: 19 jul. 2022.

HARVEY, R. **Digital Curation**: a how-to-do-it manual. New York, NY: Neal-Schuman Publishers, 2010.

HEY, T.; TANSLEY, S.; TOLLE, K. (ed.). Jim Gray on eScience: A transformed scientific method. *In*: HEY, T.; TANSLEY, S.; TOLLE, K. (ed.). **The fourth paradigm**: Data-intensive scientific discovery. Redmond: Microsoft Research, 2009. p. xvii-xxxi. Disponível em: bit.ly/3Cv2f7e. Acesso em: 29 jul. 2022.

HUNTER, J. **Scientific Models** – A user-oriented approach to the integration of scientific data and digital libraries. 2006. Disponível em: https://urx1.com/YS02G. Acesso em: 20 mar. 2023.

KAHN, R.; WILENSKY, R. **A framework for distributed digital objects service**. 1995. Disponível em: https://urx1.com/78YyW. Acesso em: 06 dez. 2022.

KAHN, R.; WILENSKY, R. A framework for distributed digital objects service. **International Journal on Digital Libraries**, Berlin, v. 6, n. 2, p. 115–123, 2006. Disponível em: https://ury1.com/yxnq5. Acesso em: 06 dez. 2022.

KITCHIN, R. Big data, new epistemologies and paradigm shifts. **Big Data & Society**, v. 1, n. 12, 2014. Disponível em: https://l1nq.com/4DYs5. Acesso em: 29 jul. 2022.

KUHN, T. S. **The Structure of scientific revolutions**. Chicago: University of Chicago Press, 1962.

LISCHER-KATZ, Z. Studying the materiality of media archives in the age of digitization: forensics, infrastructures and ecologies. **First Monday**, Chicago, v. 22, n. 1, 2017. DOI http://dx.doi.org/10.5210/fm.v22i1.7263. Disponível em: https://urx1.com/tDjon. Acesso em: 06 dez. 2022.

MARR, B. **Big data**: The 5 Vs everyone must know. 2014. Disponível em: https://bit.ly/42M05dT. Acesso em: 04 jul. 2022.

MONS, B. *et al*. Cloudy, increasingly FAIR: revisiting the FAIR Data guiding principle for the European Open Science. **Information Service & Use**, Clifton, v. 37, n. 1, p. 49-56, 2017. Disponível em: https://l1nq.com/AeVRd. Acesso em: 22 mar. 2023.

NEWMAN, W. Big Data – Building software: some thoughts on the future of building science. **Creative Education**, v. 10, n. 3, p. 524-34, 2019.

PÉREZ-GONZÁLEZ, L. Modelo/s de coste para la preservación de los datos científicos em la e-ciencia. *In*: JORNADAS DE GESTIÓN DE LA INFORMACIÓN, 12., 2010, Madrid. **Anales** [...]. Madrid: SEDIC, 2010. Disponível em: http://eprints.rclis.org/8555/1/Perez.pdf. Acesso em: 06 jul. 2022.

RHEINBERGER, H-J. **Toward a history of epistemic things**: Synthesizing proteins in the test tube. California: Stanford University Press, 1977.

SALES, L. F.; SAYÃO, L. F. Plataformas de gestão de dados de pesquisa: expandindo o conceito de repositórios de dados. **Palabra clave**, v. 12, n. 1, p. 171-171, 2022.

SANTOS, L. O. B. da S. (ed.). **FAIR digital object framework documentation**. (Working Draft). 2020. Disponível em: https://fairdigitalobjectframework.org/. Acesso em: 06 jul. 2022.

SAYÃO, L. F. Afinal, o que é biblioteca digital?. **Revista USP**, n. 80, p. 6-17, 2009

SCHWARDMANN, U. Digital objects – FAIR Digital Objects: Which services are required? **Data Science Journal**, London, v.19, n.1, 2020. Disponível em: https://l1nq.com/nHRen. Acesso em: 20 mar. 2023.

SMITH, K. *et al*. "Big Metadata": The need for principled metadata management in big data ecosystems. *In*: WORKSHOP ON DATA ANALYTICS IN THE CLOUD - DANAC'14, 2014. **Proceedings** […]. Snowbird, UT: ACM, 2014. p. 1-4.

TYBJERG, K. Exhibiting epistemic objects. **Museum & Society**, Leicester, v.15, n. 3, p. 269-286, 2017.

WILKINSON, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, London, v. 3, n. 1, 2016. Disponível em: https://ury1.com/Xg5zY. Acesso em: 22 mar. 2023.

WING, J. M. Computational Thinking. **Communications of the ACM**, New York, v. 49, n. 3, p. 33-35, 2006. Disponível em: https://l1nq.com/Wbl99. Acesso em: 20 maio 2023.

WITTENBURG, P. *et al*. **Digital objects as drivers towards convergence in data infrastructures**. 2018. Disponível em: https://urx1.com/XXJZD. Acesso em: 06 jul. 2022.

WOUTERS, P. What is the matter with e-science? – thinking aloud about informatisation in knowledge creation. **Pantaneto Forum**, n. 23, July 2006. Disponível em: https://urx1.com/J0QjK. Acesso em: 06 jan. 2023.