ISSN 1678-765X

doi 10.20396/rdbci.v21i00.8673740/32407

BY

Article

RDBCI

Revista Digital de Biblioteconomia e Ciência da Informação

Digital Journal of Library and Information Science

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

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

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.

Authors' correspondence

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

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

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

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.

CRediT

- Acknowledgments: The authors would like to thank Teodora Marly Gama for the precious revision and standardization of the article.
- Funding: This study was funded by the Brazilian agencies National Council for Scientific and Technological Development (CNPq) for the scholarships and financial support granted. This study was funded by FAPERJ through the financial support granted in Process SEI-260003/003239/2022.
- Conflicts of interest: The authors certify that they have no commercial or associative interest that represents a conflict of interest in relation to the manuscript.
- Ethical approval: Not applicable.
- Availability of data and material: Not applicable.
- Authors' contributions: Conceptualization; Data curation: SALES, L.; Formal analysis: Not applicable; Acquisition of financing: SALES, L.; Research; Methodology: SAYÃO, L. Project management: SALES, L.; Features: Not applicable; Software: Does not apply; Supervision: SAYÃO, L.; Validation: Not applicable; Visualization: SAYÃO, L.; Writing – original draft: SALES, L.; Writing – revision + editing: SAYÃO, L.
- Translation: Silvia Iacovacci - MEI

JITA: IN. Open science.

Article submitted to the similarity system

Submmited: 04/06/2023 – **Accepted:** 17/08/2023 – **Published**: 17/09/2023

Editor: Gildenir Carolino Santos

1 INTRODUCTION

In an attempt to identify a historical starting point for the journey of research data curation, Alyssa Goodman and collaborators (2014) go back to the early 1600s, when Galileo Galilei (1564-1642) turned his iconic telescope towards Jupiter. In his notebook each night, Galileo drew schematic scale diagrams of Jupiter and some strangely moving points near it, which corresponded to Jupiter's moons. "His clear and careful style of recording and publishing not only allowed Galileo to understand the Solar System, but allowed, over the centuries, anyone to understand how Galileo did it" (Goodman et al., 2014, p.1). To achieve this goal, Galileo integrated his data, metadata and text in his notes: the data correspond to the drawings of Jupiter and its moons; key metadata are the time of observation, weather conditions, properties of the telescope; and the texts comprise the description of the method of analysis and conclusions. The integrative approach of Galileo's scientific records - in terms of observation and analysis - contributed decisively to the construction of the modern scientific method. Like the astronomer Tycho Brahe (1546-1601), whose organization of data allowed Johannes Kepler (1571-1630) to formulate the laws of planetary motion, Galileo was a pioneer in outlining the concept of data curation - which today plays an important role in contemporary astronomy - establishing ways and tools to describe and record data (Gray et al., 2002).

Many centuries into the odyssey of scientific progress, paradigms have followed one another, whose cycles of validity can be characterized by the proposition of physicist Thomas Kuhn (1962), who advocated that a scientific paradigm shift occurs because the dominant mode of science is unable to deal with particular phenomena or answer key questions, thus requiring the formulation of new ideas (Kitchin, 2014), Jim Gray - a prominent computer scientist - invites us to consider a new view of scientific paradigm shifts that differs, in its conception, from Kuhn's proposition. He believed that most new science happens when data is examined in new ways (Gray et al., 2005). Gray's transitions are based on advances in the forms of data and the development of new analytical methods. In his conception, the fourth paradigm succeeds experimental, theoretical and, more recently, computational science.

This turning point marks the entry of science into a disruptive analytical and methodological conformation called the "fourth scientific paradigm". In one of his last lectures, in January 2007, Gray outlined his vision of this innovative concept, characterizing it as a new methodological approach oriented towards discovery based on data-intensive science, in addition to experimental and theoretical research and computer simulation of natural phenomena (Newman, 2019). Since then, the term fourth paradigm has become a generic term to refer to the interdisciplinary approach of data-driven scientific research or eScience, whose actions are based on two essential factors: global scientific collaboration in key areas of science and an advanced technological and social infrastructure needed to make this collaboration possible. In this scenario of change, new methodologies and analytical tools are made possible by the dizzying advance of digital information and communication technologies, which have also become richer, more flexible and much easier to use (De Roure et al., 2003).

In this transition scenario, scientific research is increasingly dependent on digital technologies. However, as an immediate consequence of these new conditions, it also depends on different ways of abstracting and representing research objects and their relationships and contexts in the various digital domains or research cyberinfrastructures in which they are managed and consumed by humans and systems. This is because at the scale and speed at which data is generated, its use is only possible through computational methods. Underlying all these new challenges is the idea of building an infrastructure based on rich metadata to represent the resources present in the research environment and to support the optimization of their reuse (Mons et al., 2017). However, this implies the need to create a more realistic research data information model that considers the complexity of a new condition of data publication that is open, contextualized and networked, and that is capable of providing a broad spectrum of functionalities that lead research objects to appropriate levels of "fairification", understood as curatorial paths to make data FAIR[1] (Sales; Sayão, 2022).

This new scientific configuration poses another important challenge: for each dataset circulating in these socio-technical infrastructures, countless metadata elements are added that document and perform critical functions on these datasets. This aggregation of information is another aspect of scientific big data, known as "big metadata". Just as data is crucial to contemporary science, big metadata relating to research objects is crucial to the discovery, sharing, reuse, crediting of authors and reproducibility of scientific experiments, to name a few of the possible functionalities performed on data and made possible by metadata elements. "Thus, in all areas, big metadata is often more massive and complex than the data [itself] that it represents," summarize Sarah Bratt and collaborators (2017, p.36).

For digital research objects, in a context of interdisciplinary integration advocated by eScience, the function of metadata goes far beyond the descriptive processes applied to conventional printed documents, loosely summarized by the phrase "metadata is data about data". The multiplicity of metadata functions necessary for the FAIR management of digital research objects, whose principles support the actions that make them findable, accessible, interoperable, so that they can be understood over time, for reuse in various contexts other than the original project, implies a considerable and diverse volume of metadata added to each digital object. Metadata elements add value to data throughout its life cycle - from generation to long-term archiving - by various agents, including software and scientific instruments. However, these elements are often disorganized, ambiguous, complex and unstructured, and need a degree of schematization to support the construction of more representative disciplinary metadata schemes.

Trying to contribute to the equation of this other aspect of scientific big data, which is installed around digital research objects, this essay aims to broaden the perspective of studies on representation as it reveals its importance within a new informational context: that of digital research objects. Underlying this proposal is a guiding question: What role does metadata play in the management of digital research objects within the context of eScience? To answer it, four axes - historical, pragmatic, standardization and epistemological - are considered in the construction of the study's methodological path, namely: the historical and foundational perspective of research object architecture initiated by Robert Kahn and Robert Wilensky (1995); the pragmatic vision of computer scientist Jim Gray and collaborators (2002, 2005), Gray's vision also present in "Jim Gray on eScience: A transformed scientific method", edited by Hey, Tansley and Tolle, (2009) on the importance of metadata for the provenance and contextualization of experiments in an eScience environment; the standardized view of OAIS, the conceptual model of the Open Archive Information System (CCSDS, 2012) on the informational digital object and the representational information that constructs its meaning; the epistemological perspective of Wouters (2006) and Rheinberger (1977) on the relationship between epistemic things, epistemic objects and technical objects; and the essentiality of representation and contextualization/decontextualization in eScience.

2 REPRESENTATION AS A CRITICAL DIMENSION OF eSCIENCE

Thinking aloud about computerization in knowledge creation, Paul Wouters (2006) points out that, "eScience is a discursive construction at the interface of techno-scientific practices, computer technology design and science policy" (Wouters, 2006, p.7), where very different practices and technologies are being integrated. In this domain, scientific instruments

[1] Acronym for Findable, Accessible, Interoperable, Reusable.

and computer simulation are creating vast silos of data that require new scientific methods to analyze and organize the data sets they contain. This is made possible by the large-scale development of socio-technical systems that process this abundance of data on the scale required and which hold out the promise of new discoveries in all areas of science.

As such, eScience becomes nothing less than a revolution in how knowledge can be created, with its conceptual foundation being the combination of various and different developments such as: the sharing of computational resources, such as grid computing; distributed access to and integration of massive data sets; the use of integrative digital platforms, such as today's research cyberinfrastructures; and advanced analysis tools and methodologies, such as artificial intelligence and sophisticated statistical methods. Once established, eScience brings about an intensified reliance on processed data: data captured by sensors and instruments and processed by software, stored in computers and managed by statistics or metadata, or both (Newman, 2019), which realizes the application of computational technology to the enterprise of modern scientific research, where "IT meets scientists", summarizes Gray (Hey; Tansley; Tolle, 2009, p. xviii). The overwhelming force brought about by this epistemic revolution affects almost all disciplinary domains, which are being transformed in one way or another; taken together, all branches of the sciences - from the exact sciences to the social sciences and humanities, and even culture and the arts - are now included in the promises of eScience (Wouters, 2006).

As a close witness and protagonist of this transition, Jim Gray (Hey; Tansley; Tolle, 2009) has noted that, in the context of eScience, we are observing the evolution of two strands of scientific exploration developing within each discipline, both of which generate and consume massive amounts of data: **computational science** and **data-intensive science** (x-computational and x-computing). Although these strands occur in the same disciplinary domain, they are so different that it is worth distinguishing between them. Computational science is related to the **simulation** of natural and social phenomena, for example, computational neuroscience simulates how the brain works. Data-intensive science, on the other hand, **collects, curates, integrates and analyzes data and information** from many different experiments. For example, while computational ecology simulates the dynamics of ecological systems, ecoinformatics collects, integrates and analyzes the data and information obtained during experiments. Newman (2019) considers that the effect of this scientific computational paradigm is the emergence of "computational thinking" which means applying computational processes to the problem at hand, reformulating the apparently difficult problem - such as human behavior or a systems design - into a problem that we know how to solve, by reduction, incorporation, transformation or simulation, that is, based on concepts fundamental to computer science. Computational thinking includes a spectrum of mental tools that reflect the breadth of computer science, concludes Wing (2006), locating abstraction and representation at the center of the discussion. These two branches of science - computational science and data-intensive science - establish a dynamic interlocution towards the generation of new knowledge through powerful infrastructures, tools and methodologies, which can be summarized by the term "cyberinfrastructure".

A research cyberinfrastructure "is a medium that allows access to and circulation of distributed knowledge, in which different communities and disciplines collaborate and communicate, breaking cultural, geographical and temporal boundaries", explains Pérez-González (2010, p. 3); it is "a new form of scientific culture that is supported by a robust high-level technological infrastructure", he adds. The devices offered by this infrastructure support unprecedented collaboration mechanisms, based on access to an extraordinary amount of data, information resources interpreted and reused by powerful observation, visualization and simulation tools.

These advanced research infrastructures have profoundly affected the long-established process of traditional scientific methodology and have greatly increased the level of production and use of data in research, enabling new types of experiments, observations, measurements, analysis, images and data visualization, according to Gray (Hey; Tansley; Tolle, 2009). The resulting methodological changes imply a shift in the starting point of research, since the integrative dynamics of data flows create patterns, anomalies, relationships and contexts that are the lenses through which a given phenomenon is studied and not an initial hypothesis based on inductive or deductive rationalities: the data is the starting point and not the proof of a hypothesis or theory (Newman, 2019). However, these ruptures are not only limited to research infrastructures and methodological conceptualization; a new generation of e-scientists is emerging to work in these new spaces of meaning. They are creating new ways of working, deeply understanding the possibilities of technologies and carrying out their research, not as an individual human being, but as a node in a network of humans and machines (Wouters, 2006), which redefines the standards of scientific communication based on scientific journals.

In this new scientific context, human beings are unable to synchronize their operational capacity with the scope, scale and speed appropriate to the magnitude of scientific big data and the complexity of eScience. Consequently, this modern-day phenomenon posits the need for humans to increasingly rely on computational agents to perform data and information discovery and integration tasks on their behalf (Wilkinson et al., 2016). "Computers are so essential in the simulation and processing of experimental and observational data that it is often difficult to draw a line between data and analysis (or code) when discussing data curation" (Goodman et al., 2014, p.1), which makes, in the view of Jeannette Wing (2006, p.33), the "interpretation of code as data and of data as code". In this way, we are increasingly living "in the era of agent-based knowledge discovery from data", as Batista et al. (2022, p.1) summarize. At the heart of this phenomenon is the availability of **machine-actionable metadata,** which provides essential contextual information for the interpretation and reuse of data in different spaces of meaning.

The promising scenarios engendered by eScience postulate the essentiality of accessing, sharing and integrating the data that is produced and consumed by its endeavors. For this to be possible, it is essential that data sets are properly **self-described** so that both computer programs and people can understand and analyze them in various contexts other than those in which they were originally created. "In some cases, breakthroughs come from analyzing existing data sources in new ways - "the pentaquark was found in the archives as soon as the theorist told us what to look for", exemplify Gray and Szalay (2004, p.3).

It seems clear that **representation** is at the heart of the eScience discussion and is the basis on which sociotechnical information systems are founded, the premise of which is to encapsulate the layers that make the content of the digital research object interpretable by automatic service providers and analysis tools. Therefore, it is necessary to know digital objects and their various faces and the roles of metadata (also digital objects), which will be discussed below.

2.1 Digital Research Object as Epistemic Object

"Many objects of science [...] have been created to generate knowledge. They may be instruments of observation or measurement; they may themselves be objects of study, such as samples or specimens; or they may be **representations or models**" (Tybjerg, 2017, p. 269, emphasis added). All these objects are called "**epistemic objects",** in the sense that they have great potential to **generate knowledge**. The concept of "epistemic objects" is based on the work of Hans-Jörg Rheinberger and his concept of "epistemic things" and experimental installations, announced in his remarkable book published in 1977. In the scope of this essay, the epistemic object or object of knowledge, as an abstract representation, is the focus of attention. To this end, we draw on Rheinberger's (1977) theoretical contribution, since he places **representation** at the heart of the scientific enterprise as a system of signifiers, interpretation and (de)contextualization. It is this approach that helps us understand the crucial role of metadata within the scope of eScience.

In experimental domains, representation can be considered equivalent to bringing epistemic things into existence - this is also our route. Upon closer inspection of the experimental system, Rheinberger (1977) distinguishes two fundamental elements: the first he calls "the object of research - the scientific object or the 'epistemic thing'. This comprises material entities or processes - physical structure, chemical reactions, biological functions - which constitute the object of investigation" (Rheinberger, 1977, p. 28). In other words: **the epistemic thing incorporates that which is not yet known** - which constitutes the object of investigation. The second element - called the "**technical object"** - is the set of experimental conditions in which the research objects are inserted. It is through this arrangement "that the objects of investigation become entrenched and articulated with one another in a broader domain of epistemic practice and material culture, including instruments, inscription devices, organismic models, and the floating theorems or boundary concepts attached to them" (Rheinberger, 1977, p.29).

It is through these technical conditions that the **institutional context** shifts to bench work, the author emphasizes. This happens in terms of local measuring facilities, supply of materials, research traditions, laboratory workflow and skills accumulated by technical personnel over long periods. The difference between experimental conditions (technical object) and epistemic thing is therefore **functional** rather than **structural**. "The technical conditions determine the field of **possible representations** of an epistemic thing; and the sufficiently stabilized epistemic thing becomes the technical repertoire of experimental arrangement," as Rheinberger summarizes in his arguments (Rheinberger, 1977, p.29, emphasis added). Therefore, an epistemic object can be considered a question-generating machine, while the technical product is a question-answering machine.

Rheinberger (1977) also clarifies the role that epistemic objects play in the space of representation created in scientific activities, bringing up the idea of decontextualization, which becomes a relevant concept in the context of curating research objects. "What is significant about representation as inscription is that things can be represented outside their original and local context and inserted into other contexts. It is the type of representation that matters," says Rheinberger (1997, p.106). Wouters (2006, p. 11) highlights the special interest of many eScience projects, at the heart of which is the decontextualization of objects and their subsequent direct contextualization in any context. He asks and at the same time answers: "How is this possible?" (Wouters, 2006, p.11). This is possible through **metadata** that must describe the meaning of the research object so that other machines and humans can (re)use these objects in contexts that were unthinkable at the time the object was produced. "Metadata are representations of the original context of epistemic objects that allow new contexts to be created so that these objects can generate new questions," comments Wouters (2006, p.11). For this to happen, it is necessary to add layers of representation in different spaces of meaning for unexpected audiences.

The requirements of today's research information systems presuppose the idea of **content interpretation by humans and systems.** The representation of information supports the interpretative potential of research objects in different and new contexts, or in new spaces of meaning, which can be explained by Rheinberger's (1977) theoretical construct of epistemic object and technical object. However, the pioneering model of digital object architecture, proposed in 1995 by Robert Kahn and Robert Wilensky, was agnostic in the sense that it did not consider the content it carried, but its elements continue to underlie more advanced architectures, such as FDO - FAIR Digital Object Architecture (Santos, 2020). This is what will be seen below.

2.2 Data Object as Informational Object

Standards - including Internet protocols - are common forms of codified knowledge that circulate between communities to ensure uniformity and similarity in processes or products across space and time (Lischer-Katz, 2017). This is the case with OAIS -- an international ISO standard (ISO 14721), which establishes a technical and applicable relationship between **data object**, **information** object and **knowledge**, which we will apply to compose our proposition.

In the context of the OAIS Model, **information** is "defined as any type of **knowledge** that can be **exchanged**, and this information is always expressed (i.e. represented) by some type of **data**" (CCSDS, 2012, p. 2-3). This definition requires the recipient of the signs or patterns (i.e. data) to be able to decode them and understand what is communicated; this requires the recipient of the message - whether human or system - to have adequate contextual and tacit knowledge to decode the signs, symbols or patterns, and then understand the message they represent. Thus, once the message has been received, a certain level of knowledge is required to process, interpret and understand it. The OAIS Model uses the concept of "knowledge base" to describe this type of knowledge. More formally, it can be said that a person or system - for example, a computer agent - has a **Knowledge Base,** which allows them to understand the information they receive. If a recipient does not yet have sufficient knowledge to understand the information, the data needs to be accompanied by representation information - that is, information that maps the data into more meaningful concepts and/or engenders a contextualization that gives it meaning - in a way that is understandable using the recipient's knowledge base. This category of information is included in the communication process and can be, in the field of scientific research, for example, a code book, a dictionary, a laboratory or field notebook, a project, a manual, notes and an infinite number of other documents. In this sense, **data** when interpreted using its **Representation Information** produces an **Informational Object** (CCSDS, 2012), more formally: the **Informational Object** is made up of a **Data Object** - which can be physical or digital - and the **Representation Information**, which allows the complete interpretation of the data.

For example, the output of a digital scientific instrument is expressed by the sequence of bits (the data) that represents, in this example, an ASCII table of numbers; when these bits are combined with representation information, they are converted into more meaningful information, such as numbers that provide the coordinates of a location on Earth measured in degrees of latitude and longitude. In order to transform the bit sequence into meaningful information, the Representation Information must contain two types of information: the first describes the format, or data structure concepts, that will be applied to the bit sequences and which in turn result in more meaningful value, such as characters, numbers, graphics, arrays, tables, visualization, to name a few. This type of information is referred to as the **Structural Informatio**n of the representational information object. But they are rarely enough, in many cases a second type of information is required: this additional information is referred to as **Semantic Information**. For example, where the Digital Object is described as a sequence of text characters, additional information must be provided regarding which language it was being expressed in. Thus, the aim of the object representation information is to convert the sequence of bits into more meaningful information.

These ideas have come a long way since they were incorporated into digital object architectures and models. In order to understand the concept of a research object with the necessary breadth, it is first necessary to understand the ideas that preceded it and those that will determine the future of information systems for research. This is what will be seen below.

2.3 Digital Object as Architecture

The concept of a digital object was introduced by Robert Kahn and Robert Wilensky in a classic article published in 1995 - A framework for distributed digital objects service - which was reprinted in 2006. The authors describe the "fundamental aspect of an infrastructure that is open in its architecture and that supports a large and **extensible class of distributed digital information services**" (Kahn; Wilensky, 2006, p.1). They also define the basic entities that must be present in this system, in which information, in the form of digital objects, is stored, accessed, disseminated and managed. The model also establishes naming conventions for identifying and locating digital objects, as well as describing a service for using object names to locate and disseminate them, and an access protocol (Kahn; Wilensky, 1995, 2006).

On the historical, conceptual and technical path initiated by Kahn and Wilensky, the structural elements they outlined are currently placed in the context of the FAIR Guiding Principles, which aim to make data locatable, accessible, interoperable and reusable. These principles assume a prominent role globally as a framework for the sustainability of research data and for the appropriate construction of curation systems. In addition, the FAIR approach always considers the idea of "machine actionability", understood as the ability of computer systems to perform services on the data without human intervention (De Smedt; Koureas; Wittenbuger, 2020; Schwardmann, 2020). This scenario seems to make it possible to realize a FAIR Internet of Data and Services (IFDS), the central point of which is the concept of FAIR Digital Object (FDO) - a type of Digital Object that is part of the FAIR Digital Object Framework (FDOF). The FDOF, as its name implies, is a framework that defines a model for **representing objects** in a digital environment, and a set of resources to provide fundamental support for FAIR principles (Santos, 2020).

A FAIR Digital Object (FDO) is formally defined by Luiz Olavo Bonino da Silva Santos (2020) as a **sequence of bits** that represents a machine-actionable unit of information, **identified** by a globally unique, persistent and resolvable identifier with predictable resolution behavior, described by **metadata** records - which are also FAIR Digital Objects -, and classified by the **FDOF typing system.** From this perspective, an FDO is a stable actionable unit that groups together enough information to allow reliable **interpretation and processing of the data it contains**.

The pioneering digital object architecture model proposed in 1995 by Robert Kahn and Robert Wilensky - which was intended to underpin a network of digital libraries of computer reports (Sayão, 2009) - was agnostic in the sense that it did not consider the content it carried. In the article, Kahn and Wilensky (1995) focus on the network aspects of the infrastructure, "i.e. those for which knowledge of the **content of the digital object is not required.** The definition of the content-based aspects of the infrastructure is deliberately not addressed [...]," the authors confirm (Kahn; Wilensky, 1995, p.118). In contrast, the requirements of today's research information systems presuppose the idea of **content interpretation**; in this sense, the FDO is a stable actionable unit that groups together enough information to allow reliable interpretation and processing of the data it contains (De Smedt; Koureas; Wittenbuger, 2020). This characteristic places **representation** at the center of the eScience discussion and is the basis on which techno-social information systems are founded, the premise of which is to encapsulate the layers that make the content of the digital object interpretable by automatic service providers and analysis tools. Once the idea of a Digital Object is understood, there are enough elements to understand what a Digital Research Object is in the eScience sphere.

3 DIGITAL RESEARCH OBJECTS IN ESCIENCE: THE BIG METADATA

In these digitally advanced scientific environments, a large part of the actions on research objects, such as analysis and visualization, are rarely performed on the objects themselves, but on abstractions and representations appropriate to the research objectives, such as models and graphic representations, or on representations in the form of metadata, supported by semantic tools such as controlled vocabularies and ontologies. Contemporary research is therefore linked to different ways of abstracting and representing research objects and their contextual relationships. This highlights the need for research infrastructures to include models for representing research objects based on rich, semantically well-structured metadata that can be interpreted by humans and machines. This condition creates an essential connection between big data science and the **phenomenon of big metadata,**

As discussed earlier, data science endeavors are motivated by the exceptional availability of digital data and new computational skills that enable data-based solutions. These ideas are also central to the realization of the fourth scientific paradigm, as advocated by Jin Gray (Hey; Tansley; Tolle, 2009) to explain the unprecedented opportunities for data-driven science. In his view, **metadata is a vital component** for the realization of eScience, although the meaning of metadata is often neglected or interpreted in a limited way, insofar as only its merely descriptive or "just data about data" face is considered, leaving all its representational, semantic and functional complexity hidden. Nevertheless, "In this new information ecology, metadata can attract a new look from research if it is understood as **big metadata"**, emphasizes Jane Greenberg (2017, p.25).

As well as an association with the diversity and size of big data, big metadata reflects the wide range of data lifecycle activities found across projects, configurations and systems. The conception of metadata as structural data that supports functions associated with a digital object, and the scale, diversity and complexity of these functions will depend on the intrinsic nature of the digital object, the environment in which it is embedded - for example, business, government or scientific research - and the idiosyncrasies of its lifecycle. However, it is at the meta level of the data lifecycle that the metadata lifecycle lies, which generates big metadata (Grenberg, 2017).

Contextualizing the phenomenon of big metadata from the example of a big data platform, such as GenBank, which is enabled by an advanced research cyberinfrastructure, whose processes produce huge amounts of data and associated with this data, also generate a large amount of metadata, "since each dataset [produced] includes its own metadata, we now have not only big scientific data but also big metadata" (Bratt et al., 2017, p.1). Just as research data is essential to contemporary science, big metadata added to these research objects is also essential for the discovery of other data, sharing, reuse, author credit and reproducibility of research and, more recently, with the emergence of data science methodologies and tools, big metadata is being used as an object of analytics and offers important insights and directions within scientific endeavors.

As part of the recent history of data science, the term "big metadata" began to appear in the literature at the beginning of the 2010s as a challenge within the scope of big data. In this context, Smith and collaborators (2014, p.1) already warned that the big data ecosystems of the time lacked an approach that considered the principles of metadata management: "In most cases, big data ecosystems have emerged without any kind of support for **metadata management** that is widely recognized as essential in traditional business systems". This became an obstacle for large organizations to share data and data preparation and analysis codes, to integrate data and ensure that analytical codes made assumptions compatible with the data that was used. This problem is also present in scientific big data, revealing the contours of **scientific big metadata,** where metadata management must become an essential part of the life cycle of data curation and fairification.

Therefore, big metadata in the scientific domain can be identified as the massive availability of metadata of various categories and types that are added to data and other research objects and that perform various functionalities that are relevant to the various aspects of the data lifecycle, such as retrieval, interoperability, reuse and various types of analysis. This complexity, intensified by eScience, implies the need for methodologies and actions aimed at organizing, classifying and schematizing metadata elements, collectively known as **metadata management.**

Chart 1 below is an interpretation by these authors of the classic 5Vs scheme, initially proposed by Marr (2014) for big data, applied to the phenomenon of big metadata that is incorporated into the challenges of data-driven science.

Quadro 1. 5 Vs applied to big scientific metadata

V	BIG SCIENTIFIC METADATA
VOLUME	The large **volume** of metadata used to describe and record scientific processes, as well as the numerous functionalities they perform during the research lifecycle, confirm the existence of big metadata. Sometimes, the volume of metadata is less than or equal to the extent of the data it describes; other times, due to the complexity of data lifecycle activities, the metadata exceeds the size of the data being described or tracked.
SPEED	In the context of scientific experiments, metadata is generated by automatic processes using instruments, remote sensors, codes or **high-speed** workflow tools. Automated platforms for the collection/generation, organization and analysis of data and metadata are increasingly present on laboratory benches. There are, however, metadata elements created by intellectual means by researchers, computer scientists and information professionals.
VARIETY	The **variety** of metadata reflects the wide diversity of formats, cycles, models, structures and types of metadata present in the scientific world. There is a clear but natural inequality in the ecosystem of metadata that records the various procedures, systems and processes of laboratories. For example: descriptive, administrative, technical, structural, preservation and processing metadata, which go beyond description and perform various functions such as retrieval, access and interoperability. The demand for specific metadata applied to disciplinary domains intensifies this variety, which is amplified by the extensive and diverse types of data and metadata life cycles, which even using the same metadata standard/schema have different implementation practices.
VERACITY	The **veracity** of big scientific metadata is established by good practices for marking metadata as a way of adding value to data and other research objects; strongly contributing to veracity are the provenance of metadata schemes, their level of standardization, the connection with scientific processes, with the practices and idiosyncrasies of disciplinary communities, the completeness of metadata elements and the precision of terminological tools - vocabularies, taxonomies, ontologies.
VALUE	The ultimate **value** of big metadata lies in its support for the proper interpretation of data and its reuse in different spaces of meaning, over time and space, by human beings and computer agents. This also creates value for the data. The value of big metadata is also associated with its role as the primary raw material for analysis methods, i.e. metadata is considered not as a representation of content, but as the very object of analysis. The value of metadata is also associated with its quality, measured by parameters such as: granularity, timeliness, accuracy, completeness and provenance (meta-metadata).

Fonte: Prepared by the authors, inspired by the structure proposed by Marr (2014).

There are several relevant dimensions that externalize the interface between eScience and CI. In this section, we have been able to highlight some of the connections most directly related to our study: the phenomenon of big scientific metadata and its management, and the extended functionalities of metadata necessary to carry out data science actions on digital research objects, such as program action. However, other connections can also be made if looked at from other perspectives.

4 THE SYNERGY BETWEEN eSCIENCE AND INFORMATION SCIENCE: THE PROTAGONISM OF DATA

In addition to the phenomenon of Big Metadata and its management, there is also a need to include new socialization network models and concepts of scientific communication vehicles as essential for building the pillar of eScience cooperation. In this environment, "some scientific communities are already experimenting with new forms of knowledge representation, such as **nano-publications,** which are basically statements in some form of semantic language such as RDF, augmented by sufficient metadata", confirm De Smedt, Koureas and Wittenbuger (2020, p.14), emphasizing metadata as the link between eScience and CI. It should be noted, however, that this perspective has been consolidated since the early days of eScience, as introduced by Jim Gray.

From the pragmatic perspective provided by his work as a computer scientist working in the field of Virtual Astronomy, Jim Gray and collaborators (2002) noted decades ago what today seems to be essential for curation: "data is incomprehensible and therefore useless unless there is a detailed and clear description of how and when it was collected and how the derived data was produced" (Gray et al., 2002, p.5). To this end, the data must be carefully documented and published in a way that allows easy access and **automatic processing**, thus opening up the possibility for generic computer tools and people to understand and reuse it. Therefore, adding information to a digital research object becomes the main responsibility of digital curation. These additions and associations occur at all points in the curation lifecycle, according to Hunter (2006).

In Gray's view, a set of scientific data of continuous value, once published, should remain available **forever**, supporting a varied scale of reproducibility and verifiability and new discoveries. However, it must be understood that the researchers who will examine this data later will not explicitly know the details of how the data was collected and processed. To understand the data, these researchers will need to know "(1) how the instruments were designed and constructed; (2) when, where and how the data were collected; and (3) a careful description of the processing steps that led to the final derived products" (Gray et al., 2002, p.1). The authors also point out that these derived products are the main objects of **investigation and scientific analysis** of the **data.** Therefore, in order to interpret data, in the current and future scenario, researchers need information expressed mainly through **metadata**. These self-descriptions added to the data are central to all the scenarios postulated by eScience.

From the practical and scientifically contextualized point of view of Jim Gray and his collaborators, metadata can be understood as: "the descriptive information about the data that explains the measured attributes, their names, units, precision, accuracy, data layout and, ideally, much more" (Gray et al., 2005, p.3). The authors also emphasize that the most important metadata should record the **lineage of the data** it describes, how the data was measured, acquired or computed. Extending the characteristics of metadata, Gray notes that quality metadata becomes central to interdisciplinary data sharing and analysis and visualization tools. Metadata should ideally record everything that should be of interest to the researcher, including data models, special equipment, instrumentation specification, data lineage and much more, Jim Gray and colleagues (2002) state.

Beyond the more technological issues, metadata plays an important role in sensitive guidelines for the scientific communication cycle, such as peer review and the reproducibility of experiments. This is because eScience materials have been subject to scrutiny in relation to the main issues that address the process of scientific integrity and ethics. For example, the reproducibility of many experiments may not be feasible for various reasons, including the difficulties of reconstructing experimental apparatus and environments. Therefore, close scrutiny of datasets, documented by rich, published metadata, may be required by academic journals to support the presumption of reproducibility. In addition, the value of providing datasets accompanied by clear and transparent methodological information used in experiments is essential for maintaining professional ethics and the integrity of research and its conclusions (Bohle, 2013).

In order for metadata to play these new roles in the context of contemporary science, which go far beyond bibliographic description - which remains essential - a wide variety of metadata types, properties and functions need to be organized to support the construction of schemes that meet disciplinary requirements. This is what will be seen below.

5 METADATA FUNCTIONALITY: FAR BEYOND BIBLIOGRAPHIC DESCRIPTION

In the context of scientific research, metadata is an umbrella term that designates the structured information and attributes that, when added to data, give it provenance, context and the potential for understanding and interpretability, elements that are critical to expanding the possibilities for data reuse. Seen in this way, metadata can be understood as a means of adding value to research data, expanding its potential to convey information and knowledge in space and time. "Beyond labelling and categorization, metadata can be thought of more universally as a value-added language that serves as an integrating layer in an information system," adds Jane Greenberg (2017, p.22). This abstraction connects the research object to a set of important functionalities that support data management in the context of an information system - such as a digital repository - such as identification, retrieval, preservation, levels of contextualization and provenance, or even reuse permissions.

For their part, digital research objects are characterized by having a complex and long life cycle, which depends on different disciplinary contexts and perspectives of (re)use in different domains. This lifecycle starts before the research begins and extends indefinitely beyond the end of the project, when the data needs to be archived for the long term in reliable systems. Throughout this journey, various types of metadata are added to objects, assigned by different stakeholders, including those generated automatically by scientific instruments (Wittenburg et al., 2018) and by laboratory workflow tools, in a continuous process of adding value to datasets and other research objects. Ideally, this metadata should be understood by both humans and computers. Specifically, as far as FAIR-compatible resources are concerned, data, metadata and services should meet the requirements of being machine actionable without human supervision wherever possible, especially to achieve the goals of a FAIR Internet of Data and Services (Mons et al., 2017). Consequently, these digital research objects require a wide range of metadata - with many functions and properties - which often **surpass** the data itself in volume and even importance, in some situations amounting to big metadata. It is clear, therefore, that for digital research objects, the function of metadata goes far beyond the descriptive processes applied to conventional printed documents, loosely summarized by the phrase "metadata is data about data". The functions associated with the metadata of a digital research object and its scale, diversity and complexity depend on the nature of the digital object, the environment in which it is inserted, its structural and semantic components and the disciplinary peculiarities that determine its life cycle.

In view of the boundary conditions required by eScience, it is essential to establish guidelines for the construction, management and application of rich metadata that supports the reproducibility and reuse of research objects. This is especially relevant in specific, vertical disciplinary domains where it is critical to describe complex experiments involving multiple processes that need to be highly contextualized. This takes on a whole new perspective in the context of open science, where not only the end results - data and publications - must be described, but also the entire apparatus for achieving them, such as models, computer codes, algorithms, laboratory methods, equipment, workflow, etc. which, in the end, have a very disciplinary character.

Ross Harvey (2010) outlines some of the information that the data curator and various other agents - including scientific instruments and software, such as laboratory workflow systems - add to the data during its life cycle in the form of metadata and documentation. This information allows the data to be effectively managed, accessed and reused, now and in the future. The result of this essay is a systematization of metadata functionalities for digital research objects (Chart 2).

Chart 2. Functionality of metadata

CATEGORY	FUNCTION
REPRESENTATION	It **codifies** and contributes to organizing the knowledge of a **disciplinary domain** in a way that is relevant to the research field and familiar to the research community.
	It **transforms** data objects into **information object**s (CCSDS, 2012) or **epistemic objects** (Rheinberger, 1977).
	It shows what needs to be **represented in the research object,** to the standard required by users (HARVEY, 2010).
DESCRIPTION	It **identifies** the research object uniquely, globally and persistently (considering the persistent identifier as part of the elements of the metadata schema).
	Locates the object of research.
	Clarifies what the object of research is.
	It records **bibliographic information** about the research object, allowing it to be **referenced and cited** according to the relevant standards; it gives **credit to the various authors of** the objects.
	It points out the **properties and technical structures** of the elements that make up the digital research object.
MANAGEMENT	Supports the complete **management** of the research object's **life cycle.**
	Records the **provenance of the metadata schema** (informed by the meta metadata) and the degree of management applied to the current schema.
RECOVERY	It supports the **formulation of queries** with an appropriate level of granularity and precision, considering disciplinary characteristics.
	Supports **findability** and **access** to digital objects.
	Supports the **selection** and **evaluation** of retrieved objects.
	Maintains **reliable links** to the object.

CATEGORY	FUNCTION
RELATIONSHIP	Provides **links** from the research object to other related objects (journal articles, software, datasets, etc.) to **make visible the ecosystem where the object is located and its relationship with other research objects.**
INTERPRETABILITY	It expands the **level of interpretability** of research objects in different spaces of meaning for humans and computer agents, now and in the future.
	It facilitates the **reuse** of research objects by their creators and by other researchers.
	It supports **interlocution with other collections and with different systems** by automated means.
PROVENANCE AND CONTEXT	Records the **history of the research object** (provenance, traceability and lineage).
	It reports on **the processes, parameters, variables, methodologies, codes and instruments** that were relevant to collecting/generating, processing and analyzing the research object.
QUALITY	Indicates **the quality assurance and control actions** applied to the objects (e.g. flags for missing or discrepant data, etc.).
INTEROPERABILITY	Enables **machine actionability** of search objects.
PEER REVIEW	**It informs reviewers of conventional articles and data articles about the processes of obtaining**, processing, analysis, quality levels and the potential for reproducibility of research objects.
PRESERVATION	It supports long-term **preservation strategies.**
	Inform the **technical dependencies** of the object of research.
	It provides semantic and structural **representation information** that allows interoperability with the future. Reconfiguration of the object in the future.
	It is part of the **archiving information packages** (AIP/OAIS).
TRUST	It supports the presumption of **authenticity and reliability** of research objects.
	Records the actions applied to guarantee the **integrity** of objects, such as hashing and chercksun.
PERMISSIONS	Announces the **licenses** associated with the object.
	Tells you **how sensitive** the object is.
	Informs about **access permissions.**
	It points out the **actions that can be carried out** on the object of research.
COPYRIGHT	Informs about the **intellectual property** rights associated with the object.

CATEGORY	FUNCTION
ANNOTATION	It includes collaborative comments **from researchers other than** the authors.
ADMINISTRATION	Identifies the **people/teams who operate** the data lifecycle: who collects, who indexes, who ensures physical security, etc.
	Identify the ***stakeholders* involved,** for example: funders, partner institutions.
	Identifies who is **responsible for the management and preservation** of research objects.

Fonte: Prepared by the authors (2023)

It is believed that this structure can contribute to the management of metadata that turns digital research objects into informational or epistemic objects. However, in order to perform their functions, metadata elements still need another layer of representation which corresponds to the codification of content by means of standardized terminological instruments/schemes, such as controlled vocabularies, thesauruses, ontologies, taxonomies and other classificatory structures. These instances are outside the context of this article, but they also present an opportunity for study in the field of Information Science, especially in the context of Knowledge Organization, once again bringing this area closer to eScience.

6 BY WAY OF CONCLUSION

The fourth scientific paradigm, realized by eScience, is, in a nutshell, a methodological approach that leads these authors to insights and new discoveries based on advanced processes of integrating and analyzing the abundance of data. As Newman (2019) notes, this is not exactly new, insofar as increasingly complex and large amounts of data can be considered part of the impetus for empiricism in 19th and early 20th century science. "What is new - and not yet clear - is how this new paradigm will transform the fundamental ways of researching and acting in twenty-first century science and technology" (Newman, 2019, p.525). Gim Gray (Hey; Tansley; Tolle, 2009) summarizes this transformation, which is still embodied pragmatically, by stating that eScience is where computing meets scientists. Thus, a unifying dimension of eScience, essential for the realization of its dreams and promises, is the information and communication technology infrastructure that surrounds it (Wouters, 2006). It seems that the transformations in science will continue at the dizzying pace of the evolution of computing.

At the end of this essay, the conclusion is that there is a fundamental role for Information Science and Librarianship in the revolution of computational discovery that underlies eScience.

As we have seen throughout this work, what connects eScience to Information Science are the various relevant dimensions of representation: epistemological, technological and sociological - referring to the idea of computational thinking about abstraction and actions that are performed by the various functions that extend the long-established concept of metadata.

This path can be measured by the ability of computer agents to interpret and act on metadata when providing advanced information services, such as integrative analysis and visualization. Automatic interpretation is the key to the Internet of the future: the FAIR Internet of Data and Services, whose main element is FAIR Digital Objects, endowed with their own architecture whose key is rich metadata. In this sense, it is worth remembering that FAIR is first and foremost about metadata, and a quick reading of the principles shows that the concept of metadata is expressed in almost all of the 15 principles.

In the course of this study, it was possible to observe the confluence of ideas around the relevance of representation and abstraction for the realization of eScience undertakings. Simulations, modelling, visualization, virtualization of physical research objects (such as samples, herbaria, etc.), algorithms, as well as metadata are at the heart of the epistemic abstractionism of eScience discovery processes. On the CI side, issues relating to representation have always accompanied its broad spectrum of study and application, proving fundamental also for improving communication between people, institutions and, more recently, between machines. In the specific context of Library Science, metadata has always been a key element for cataloging and indexing books and other documents contained in libraries. In the context of Documentation, from the concept of document advocated by Suzanne Briet (1951), in her treatise Qu`est-ce que la documentation? the importance of representation was put on the agenda, so that any object could become a document. Jin Gray, in his founding vision, was able to foresee the importance of the connection between eScience and CI, which today is further intensified by big metadata: "Astronomers will probably reinvent many of the **concepts** already developed in the **library and museum communities**. Librarians have thought deeply about these issues and we would do well to learn from their experience" (Hey; Tansley; Tolle, 2009).

The practical results of this research broaden the perspective of representation studies as it reveals its importance within a new information context: that of digital research objects and their accumulation of metadata, known as big metadata, which, as seen in table 1, establishes a dialogue with the 5Vs of big data, when they themselves become sources for new analysis methodologies. The essay also tries to show that metadata also becomes a management object within research cyberinfrastructures, consolidating the idea of "metadata management".

In this sense, as Information Science and Librarianship converge on the eScience areas, the role of metadata becomes increasingly present, generating new challenges that should be configured in future studies, especially with regard to the construction of FAIR disciplinary metadata schemes that meet the representation and functionality needs of the x-computational and x-informatics sciences, located in the context of eScience. Thus, this research continues as new studies are being developed on "metadata authoring", FAIR vocabularies and semantic compatibility, which should be published soon.

REFERENCES

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

BOHLE, S. **What is E-science and how should it be managed?** 2013. Available at: https://bit.ly/3Je1raz. Access on: 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. Available at: https://bit.ly/43Ul9Qu. Access on: 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). Available at: https://public.ccsds.org/pubs/650x0m2.pdf. Access on: 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. Available at: https://bit.ly/3CrUnU7. Access on: 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. Available at: https://ury1.com/lB3fB. Access on: 29 jul. 2022.

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

GRAY, J. *et al*. **Scientific data management in the coming decade**. Redmont, WA: Microsoft Corporation, 2005. Available at: https://ury1.com/rAhgO. Access on: 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. Available at: https://arxiv.org/abs/cs/0502011. Access on: 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. Available at: https://urx1.com/qcMN3. Access on: 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. Available at: bit.ly/3Cv2f7e. Access on: 29 jul. 2022.

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

KAHN, R.; WILENSKY, R. **A framework for distributed digital objects service**. 1995. Available at: https://urx1.com/78YyW. Access on: 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. Available at: https://ury1.com/yxnq5. Access on: 06 dez. 2022.

KITCHIN, R. Big data, new epistemologies and paradigm shifts. **Big Data & Society**, v. 1, n. 12, 2014. Available at: https://l1nq.com/4DYs5. Access on: 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. Available at: https://urx1.com/tDjon. Access on: 06 dez. 2022.

MARR, B. **Big data**: The 5 Vs everyone must know. 2014. Available at: https://bit.ly/42M05dT. Access on: 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. Available at: https://l1nq.com/AeVRd. Access on: 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. Available at: http://eprints.rclis.org/8555/1/Perez.pdf. Access on: 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. Available at: https://fairdigitalobjectframework.org/. Access on: 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. Available at: https://l1nq.com/nHRen. Access on: 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. Available at: https://ury1.com/Xg5zY. Access on: 22 mar. 2023.

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

WITTENBURG, P. *et al*. **Digital objects as drivers towards convergence in data infrastructures**. 2018. Available at: https://urx1.com/XXJZD. Access on: 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. Available at: https://urx1.com/J0QjK. Access on: 06 jan. 2023.