DOI: 10.5433/1981-8920. 2023v28n4p1

METADATA AUTHORING MODEL: DESCRIBING INFORMATION ABOUT CONTEXT AND PROVENANCE OF DISCIPLINARY RESEARCH OBJECTS

MODELO DE AUTORIA DE METADADOS: DESCREVENDO INFORMAÇÃO SOBRE CONTEXTO E PROVENIÊNCIA DE OBJETOS DE DADOS DISCIPLINARES

Luís Fernando Sayão[a]
Luana Farias Sales[b]

ABSTRACT
In the field of research object management, there are a large number of standardized metadata schemas available, but in general they do not address the fragmentation and interdisciplinarity of contemporary science. **Problem:** There are rich discipline-oriented metadata schemas in some key areas, but in other most cases they need to be constructed. Therefore, a major challenge for research objects to achieve an adequate level of FAIRification is that they are described by metadata schemas that have functionalities and qualities that support research reproducibility and data reuse. **Objective:** To address this complexity, the goals of this research was to define the functionalities and quality levels of metadata standards required for FAIR research data management. **Methodology:** This is a theoretical and exploratory research based on the concept of epistemic/technical/informational research object, considering four axes: historical, epistemological, standardization and application. **Results:** As a result, a metadata authoring model was proposed that focused on recording the context and origin of research objects. **Conclusion:** In conclusion, the paper reaffirms the urgent need to develop disciplinary metadata schemes that not only meet the specific needs of

[a] Doutor em Ciência da Informação pela Universidade Federal do Rio de Janeiro (UFRJ). Docente do Programa de Pós-Graduação em Ciência da Informação do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT/UFRJ), do Programa de Pós-Graduação em Biblioteconomia da Universidade Federal do Estado do Rio de Janeiro (UNIRIO) e do Programa de Pós-Graduação em Memória e Acervos da Fundação Casa de Rui Barbosa. Atua na Comissão Nacional de Energia Nuclear (CNEN), Rio de Janeiro, Brasil. E-mail: luis.sayao@cnen.gov.br

b Doutora em Ciência da Informação pela Universidade Federal do Rio de Janeiro (UFRJ). Docente do Programa de Pós-Graduação em Ciência da Informação do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) e do Programa de Pós-Graduação em Biblioteconomia da Universidade Federal do Estado do Rio de Janeiro (UNIRIO). Analista em Ciência e Tecnologia do Instituto Brasileiro de Informação em Ciência e Tecnologia do Rio de Janeiro (IBICT), Rio de Janeiro, Brasil. E-mail: luanasales@ibict.br

the domains, but also ensure interdisciplinary integration and efficient data retrieval, promoting more robust, accessible and collaborative science.

Descriptors: Metadata authoring. Research object. Data provenance. Data contextualization.

1 INTRODUÇÃO

In an attempt to identify a historical starting point for the journey of care with the research data, Alyssa Goodman and collaborators (2014) go back to the early 1600's when Galileo Galilei (1564-1642) turned his telescope toward Jupiter. In his logbook, each night, Galileo drew to-scale schematic diagram of Jupiter and some oddly moving points near it, the moons of Jupiter. "His clear and careful record keeping and publication style not only let Galileo understand the Solar System, but it continues to let anyone understand how Galileo did it" (Goodman, 2014, p. 1). To achieve this purpose, Galileo integrated his data, metadata and text in his notes: data correspond to drawings of Jupiter and his moons; key metadata are time of observation, weather, telescope properties; and texts are the description of method of analysis and conclusions. The integrative approach of Galileo´s scientific records – regarding observation and analysis – contributed decisively to the framing of the modern scientific method. As well as the astronomer Tycho Brahe (1546-1601), Galileo was a pioneer in outlining the concept of curated data - which nowadays plays an important role in contemporary Astronomy - establishing ways and tools to describe and record data (Gray *et al.*, 2002).

After many centuries of the odyssey of scientific progress, where paradigm changes take place according to the proposition of the physicist Thomas Kuhn (1962), who stablished that a scientific paradigm shift occurs because the dominant mode of science cannot account for particular phenomena nor to answer key questions, thus demanding the formulation of new ideas (Kitchin, 2014), Jim Gray – a computational scientists - invites us to consider a new view on scientific paradigm shifts that differ 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 founded on advances in forms of data

and the development of new analytical methods. In Gray´s understanding,

> originally, there was just experimental science, and then there was theoretical science, with Keppler Laws, Newton´s Laws of Motion, Maxwell´s equations, and so on. Then, for many problems, the theoretical model´s grew too complicated to solve analytically, and people had to start simulating. These simulations have carried us through much of the last half of the last millennium. At this point, these simulations are generating a whole lot of data, along with a huge increase in data from experimental science (Hey; Tansley; Tolle, 2009, p. xvii).

At this moment science is entering a fourth paradigm - also called eScience - based on growing availability of Big Data and on new practice of knowledge creation. In this scenario, novel analytics methodology and tools come to be possible by the vertiginous advance of information and communication technology, that has become richer, more flexible and much easier to use (De Roure *et al.,* 2003). eScience claims to and be no less than a revolution on how knowledge can be created; it is generally defined as the combination of three different developments: the sharing of computational resources; distributed access to massive datasets, and the use of digital platforms. The overwhelming force brought about by this epistemic revolution affects almost every disciplinary domain, which will be transformed in some way or other; taken together, every branch of sciences – from exact science to social science and humanities – have now been included in the promises of eScience (Wouters, 2006).

In this sense, Jim Gray (Hey; Tansley; Tolle, 2009) verified that we are witnessing the evolution of two branches of each discipline for scientific exploration in the context of the fourth paradigm. They are so different that it is worth distinguishing them: the computational science which has to do with the simulation of natural and social phenomena; and data-intensive science which collects and analyses data and information from many different experiments.

According to Gray, the new research infrastructures have deeply affected the long-established process of scientific methodology and vastly increased the level of data production and use in research, enabling new types of experiments, observations, measurements, analyzes, imaging and data visualization (Hey; Tansley; Tolle, 2009). However, it is not limited to the infrastructure; to work in these new spaces of meaning a new generation of e-scientists is emerging. They are creating new ways of working, they deeply understand the possibilities of

technologies and perform their research, not as an individual human, but as a node in a network of humans and machine (Wouters, 2006, p. 6).

At this novel scientific context, human beings are unable to operate at the scope, scale and speed needed by the magnitude of the contemporary scientific big data and complexity of eScience; so, this phenomenon of our days implies the need for humans increasingly rely on computational agents to undertaken discovery and integration tasks on their behalf (WILKINSON *et al.,* 2016). Hence, we are more and more living "in the era of agent-driven knowledge discovery from data", as qualified by Batista *et al.* (2022, p. 1). At the heart of this phenomenon is the availability of a machine actionable metadata, which provides contextual information essential to interpret and reuse the data in different signification spaces.

In this transitional context, scientific research is becoming more and more dependent on information and communication technologies and on different modes of representation in some digital domain or cyberinfrastructure. Hence, subjacent to all those new challenges is the idea of building an infrastructure based on rich metadata for the resources in the research environment to support their optimal reuse (Mons *et al.,* 2017). This implies the necessity of creation of a more realistic research data information model that considers the complexity of an open, contextualized and networked data-publishing condition. In the center of this model is the "informational research object" that must be interpreted in different space of signification by humans and systems.

In which refers to digital objects, more specifically research objects, the function of metadata goes far beyond the descriptive processes applied to conventional printed documents, synthesized by the loosely phrase "metadata is data about data". More elaborate and comprehensive definitions are based on the conception of metadata as structural data supporting functions associated to a digital object; the scale, diversity and complexity of these functions will depend on the intrinsic nature of the digital object, the environment where it is inserted – for instance, business or scientific research - and its lifecycle idiosyncrasy. Those function can be, for example: preservation, discovery, access restriction, interoperability, legal and ethics conditions etc. These distinct functions of

metadata clearly indicate that to manage informational digital objects we need an extensive spectrum of metadata types.

Digital research objects in particular are characterized by having a complex and long lifecycle, which depend on different disciplinary contexts and (re)use perspectives. This lifecycle starts before the research begins and extends indefinitely beyond the end of the project. Along this journey, various types of metadata are added to the data, assigned by different stakeholders over time, including those automatically generated by scientific instruments and by workflow software tools, in a continuous process of adding value to the datasets and to others research objects. These metadata, ideally, should be understood by humans as well as by computers. Specifically, in the perspective of FAIR (Findable, Accessible, Interoperable, and Reusable) compliant resources, data, metadata and services should meet the requirements of being actionable by machine without human supervision whenever and wherever possible, mainly to achieve the objectives of an Internet FAIR of Data and Services (MONS *et al*., 2017). Therefore, digital research objects require a wide range of metadata - with many functions and properties - that **often** outweigh the data itself in volume and even in importance.

In face of these new conditions, it is essential to provide guidance on how to produce richly metadata elements that meet the community and the FAIR requirements and to underpin the reproducibility and reuse of datasets. This is especially relevant in a specific disciplinary domain where is relevant to describe complex experiments that involve multiple processes that need to be highly contextualized. Considering those complexities, this essay aims at proposing a modular metadata authoring model that incorporates the concept of “reuse of metadata elements”. These enhancing metadata elements come from others standard schema that describe either specific activities, process or procedures, as computational code, laboratory methods, specific phenomenon and son on; or provide general description, as bibliographic description. The authoring model incorporates type, functionality and quality of metadata, roles of stakeholders and the community participation as well. To theoretically support this proposal, studies are carried out on the concept of informational and epistemic digital object.

Underlying this proposal, four axes – historical, pragmatical, standardized and epistemological - are considered, namely: the historical and foundational perspective on research object architecture that has been initiated by Robert Khan and Robert Wilensky (1995); the pragmatical view of the computer scientist by Jim Gray and collaborators (2002, 2005) Gray's view 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 the experiments in eScience environment; the standardized view of OAIS, Open Archive Information System conceptual model (*Consultative Committee For Space Data System*, 2012) on informational digital object and in the representational information that build its meaning; and 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 RESEARCH OBJECT AS AN ARCHITECTURE

Eric Hobsbawm (1995) recognized as the most important historian of the 20th century, considers that human being, in the current world, lives the illusion of an eternal present, a disconnection with the past that affects the perception of how things have continually developed. However, deeper current debates often stimulate research to ask new questions about the past (Burke, 2003). Therefore, we consider that to fully understand the concept of research object, it is necessary to first understand the previous idea of the research data and more deeply the primordial idea of digital object.

The concept of digital object has been introduced by Robert Kahn and Robert Willensky in a classical 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 which supports a large and extensible class of distributed digital information service" (Khan; Willinsky, 2006, p. 1). They also define basic entities to be found in such system, in which information in the form of digital object is

stored, accessed, disseminated and managed; provide naming conventions for identifying and locating digital objects; describe a service for using object names to locate and disseminate objects; and provide an access protocol (Khan; Wilensky, 1995, 2006).

In the historical, conceptual and technical path initiated by Khan and Wilensky, the architectural elements that were outlined by them are currently placed in the context of the FAIR Guiding Principles for findable, accessible, interoperable and reusable data; these principles become a prominent role as a framework for the sustainability of research data. Besides, that approach has always machine actionability, in the sense of the capacity of computational systems to perform services on the data without human intervention (De Smedt; Koureas; Wittenburg, 2020; Schwardemann, 2020). This scenario seems to make possible the realization of an Internet FAIR of Data and Services (IFDS), whose central point is the concept of FAIR Digital Object (FDO), a type of Digital Object that is in the context FAIR Digital Object Framework (FDOF), a framework that defines a model to represent objects in a digital environment and a set of features to provide foundational support for the FAIR principles (Santos, 2021).

A FAIR Digital Object (FDO) is defined formally by Luiz Bonino da Silva Santos (2021) as a sequence of bits that represents a machine-actionable information unit, identified by a globally-unique, persistent, and resolvable identifier with predictable resolution behavior, described by related metadata records – a FDOs themselves -, and classified by the FDOF typing system. From this perspective, an FDO is a stable actionable unit that bundles sufficient information to allow the reliable interpretation and processing of data contained in it.

The pioneering paper of Khan and Wilensky (1995) focuses on the network-based aspects of the infrastructure, “namely those for which knowledge of the content of the digital object is not required. Definition of the content-based aspects of the infrastructure is purposely not addressed […]”, confirm the authors (Khan; Wilensky, 1995, p. 118); in contrast the FDO is a stable actionable unit that bundle sufficient information to allow the reliable interpretation and processing of the data contained in it (De Smedt; Koureas; Wittenburg, 2020).

This characteristic puts the representation in the center of the discussion of the eScience and is the base that underlies the current research information system, which encapsules the layers that become the content of digital object interpretable by automatic services providers and analysis tools.

2.1 DATA RESEARCH AS OBJECT OF E-SCIENCE

Thinking aloud about informatization in knowledge creation Paul Wouters (2006) highlights that, “E-science is a discursive construction at the interface of technical-scientific practices, computer technology design and science policy” (WOUTERS, 2006, p. 7), where very different practices and technologies are being put together. In its domain, scientific instruments and computer simulation are creating vast data stores that require new scientific method to analyze and organize the datasets, made possible by the development of large-scale sociotechnical systems which give rise to promises of new discoveries in almost all areas of science. Those promising scenarios engendered by eScience postulate ease access, sharing and integrating of the data which are produced and consumed by its endeavors. To make this possible, it is essential that the datasets be properly self-described so that computer programs as well as people can understand and analyze them in several other contexts different from those in which they were originally created. “In some cases advances come from analyzing existing data source in new ways – the pentaquark was found in the archives once theoretician told us what to look for”, exemplify Gray and Szalay (2004, p. 3).

From that pragmatic perspective of computer scientists working in the domain of virtual Astronomy, Jim Gray and collaborators (2002) have testified years ago that “data is incomprehensible and hence useless there is a detailed and clear description of how and when it was gathered, and how the derived data was produced” (Gray *et al.*, 2002, p.5). For that, data must be carefully documented and must be published in forms that allow easy access and automatic manipulation, which enables generic automatic tools as well as people to understand and reuse them.

In Gray´s belief, once published, continuous-value scientific dataset

should remain available forever giving support to the varied scale of reproducibility and to new discoveries. However, later on researchers will not implicitly know the details of how the data was gathered and processed. To understand the data those later researchers need to know "(1) how the instruments were designed and built; (2) when, where, and how the data was gathered; and (3) a careful description of the processing steps that led to the derived data products" (Gray *et al.*, 2002, p. 1). The authors highlighted as well that those products are the principal objects of scientific data analysis. Then, to interpret the data, in actual and future setting, researchers need information expressed mainly by metadata. These self-descriptions added to data, accomplished by metadata, are central to all scenarios postulated by eScience.

From a scientifically applicable and pragmatic point of view, metadata may be understood as: "The descriptive information about data that explains the measured attributes, their names, units, precision accuracy, data layout and ideally a great deal more" (Gray *et al.*, 2005, p.3). The authors still emphasize that the most importantly metadata should record the data lineage that describes how the data was measured, acquired or computed. Extending the features of metadata, observes Gray that good metadata become central for interdisciplinary data sharing and for data 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, consolidate Jim Gray and collaborators (2002). For metadata to accomplish those purposes, the astronomers and other scientists who work in data-driven research will likely reinvent many of the concepts already well developed in the library and museum communities, concludes Gray, establishing an important connection to information sciences when he emphasizes the need for the data object to become an informational object.

2.2 DATA OBJECT AS AN INFORMATIONAL OBJECT

The standards - including Internet protocols - are common forms of codified knowledge that circulate across communities to ensure uniformity and sameness within processes or products across space and time (Lischer-Katz,

2017). That is the case of OAIS –-, an international ISO Standard (ISO 14721), that establishes a technical and applicable relation between data object, informational 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” (Consultative Committee for Space Data System, 2012, p. 2-3). This definition requires that the recipient of the signals or patterns (i.e., data) is able to decode them and understand what is communicated; this requires that the recipient of the message has adequate contextual and background knowledge to decode the signals, symbols or patterns and then to understand the message that these represent. So, once the message is received a certain level of knowledge is required to process, interpret, and make sense of it. The OAIS Model use the concept of “knowledge base” to qualify this kind of knowledge. More formally: a person or system – for instance, a computational agent – can be said to have a Knowledge Base, which allows them to understand received information. If a recipient does not already include enough knowledge to understand the information, the data needs to be accompanied by Representation Information – the information that maps a data into more meaningful concepts – in a form that is understandable using recipient’s Knowledge Base; this category of information is included as part of the communication process and can be, in the domain of scientific research for example, a code book, a dictionary, a laboratory or field notebooks, a project, a manual and much more. In this sense, data is interpreted using its Representation Information yields Information (Consultative Committee for Space Data System, 2012) more formally: The Information Object is composed of a Data Object – that is either physical or digital – and the Representation Information that allows for full interpretation of the data.

For instance, the output of a digital scientific instrument is expressed by the bit sequence (the data) it contains that represent, in this example, an ASCII table of numbers; when those bits are combined with Representation Information they are converted into more meaningful information as numbers giving the

coordinates of a location on the Earth measured in degrees latitude and East latitude. To transform bit sequence in meaningful information, the Representation Information must contain two kinds of information: The first one describes the format, or data structure concepts, which are to be applied to the bit sequences and that in turn result in more meaning value such as characters, numbers, graphics, arrays, tables, visualization etc. This kind of information are referred as Structure Information of the Representation Information object. But they are seldom sufficient, in many cases a second kind of information is required: this additional information is referred as Semantic Information. For instance, where the Digital Object is described as a sequence of text characters, the additional information relative as to which language it was being expressed should be provided. That is the case where the text is formatted as .txt and the language is in Portuguese. Then the purpose of the Representation Information object is to convert the bit sequence into more meaningful information

The Khan and Wilensky`model 1995 was agnostic in the sense of being content neutral. However, the requirements of the research information systems presuppose the idea of content interpretation. The representation of information – documentation and metadata - supports the interpretative potential of the research objects in different and in new context or in new space of signification, which can be explained by the Rheinberger's (1977) theoretical concept of epistemic object. That's what we'll see next.

2.3 RESEARCH OBJECT AS AN EPISTEMIC OBJECT

"Many objects of science […] were created in order to generate knowledge. They may be instruments of observation or measurement; they may be the objects of study themselves, such as samples or specimens; or they may be representations or modes" (Tybjerg, 2017, p. 269). All those objects are called "epistemic objects" in the sense that they have a great potential to generate knowledge. The concept of "epistemic objects" draws on the work of Hans-Jörg Rheinberger and from his concept of "epistemic things" and experimental set-ups, announced in a remarkable book published by 1977. In the scope of the present essay, the epistemic object or knowledge object, as abstract

representation, is our focus of attention. For this, we will ask for theoretical support from Rheinberger (1977) who puts representation in the heart of the scientific enterprise as a system of signifiers, interpretation and (des)contextualization, an approach which helps us to understand the role of metadata in the realm of eScience.

In the experimental domains, representation may be taken to be equivalent to bringing epistemic thing into to the existent. In inspecting experimental system more closely Rheinberger (1977) distinguishes two fundamental elements: the first he calls “the research object, the scientific object or the ‘epistemic thing’. This comprehends material entities or processes - physical structure, chemical reactions, biological functions – that constitute the object of inquiry” (Rheinberger, 1977, p. 28). In few words: epistemic thing embodies that which is not yet known. The second element – called the “technical object” - is the set of experimental conditions in which the research objects are embedded. It is through that arrangement “that the objects of investigation become entrenched and articulated themselves in a wider field of epistemic practice and material culture, including instruments, inscriptions devices, model organisms, and the floating theorems or boundary concepts attached to them” (Rheinberger, 1977, p. 29). It is through these technical conditions that the institutional context passes down to bench work, emphasizes the author. This happens in terms of local measuring facilities, supply of materials, research traditions, laboratory workflow and accumulated skills carried on by long-term technical personnel. The difference between experimental conditions (technical object) and epistemic things is, therefore, functional rather than structural. “The technical conditions determine the realm of possible representations of an epistemic thing; and sufficiently stabilized epistemic thing turns into technical repertoire of experimental arrangement”, as Rheinberger summarizes his arguments (Rheinberger, 1977, p. 29). Therefore, an epistemic object is first and foremost a question-generating machine, while the technical product is an answering-machine.

Rheinberger (1977) also clarifies us the role the epistemic objects play in the space of representation created in scientific activities, bringing up the idea of decontextualization that becomes a relevant concept in the context of research

data curation. “What significant about representation qua inscription is that things can be represented outside their original and local context and inserted into other contexts. It is the kind of representation that matters”, states Rheinberger (1997, p. 106). Wouters (2006) highlights the special interest of many eScience projects, whose very core aims at the decontextualization of objects and subsequent contextualization on the fly and in any context. He asks and at same time answers: “How is this being made possible?” (Wouters, 2006, p. 11) This is being possible by metadata that should describe the meaning of the research object so that other machines and humans could make (re) use of those objects in contexts that have been unthinkable at the moment of the production of the object. “Metadata are representations of the original context of epistemic objects in such terms that new contexts can be created for these objects to generate new questions”, remarks Wouters (2006, p. 11).

Finally, in contemporary science the product of research is not stand-alone publication in some prestigious journal or book, but the addition of both technical and epistemic object into a specific space of representation, that needs a system of meaning (Wouters, 2006). In this perspective, the metadata must support the interpretative potential of the epistemic object in different and new context. Therefore, it is crucial that metadata schema be built to represent - with appropriate level of precision and granularity - the context and provenance of the datasets and others research objects in order to generate and answer unprecedented questions and to improve the repeatability and verifiability of the results. In this sense, scientific research is becoming more dependent on information and communication technologies, all of them represented in the same digital domain, that which in science is frequently called search cyberinfrastructure.

3 WHAT IS THE IMPORTANCE OF METADATA IN DATA INTENSIVE SCIENCE?

In the scientific research domain, **metadata** is an umbrella term for structured information and attributes that provide provenance, context, and interpretability potential to research data object. The metadata is being applied to datasets and to the data contained therein, transforming them into informational

objects. In this sense, metadata can be thought of as a form of data affixed to other data that provides description and enhance the meaning of the data in specific (disciplinary) space of significances. Within this approach, metadata can be thought of as a way of adding value to research data increasing its potential to convey information and knowledge across space and time, being essential to accurate data interpretation and (re)use by both humans and machines, completes Jane Greenberg (2017). To fully performs that role, metadata should ideally record everything that would be of interest to the researcher (Gray *et al.,* 2002). Furthermore, records of the complete scientific discovery process will enable peers to review the method of conducting the science as well as the final conclusions. It will also enable greater sharing, re-use and comparison of scientific results (Hunter, 2005).

Further expanding functionality of metadata and connecting the scientific landscape to a data curation perspective, Sarah Higgins (2007 p. 1) considers that “Metadata is the backbone of the digital curation. Without it a digital resource may be irretrievable, unidentifiable or unusable”, and it – we can add - cannot even be preserved and made interoperable and connected with other research objects. According to Hunter (2005), assigning information to a digital research object becomes the main responsibility of digital curation; these additions and associations occur in all points in the curation lifecycle. However, from the researcher's perspective, two related concepts stand out as vital for reproducibility and reuse: **context** and **provenance**. In addition to these guiding concepts, in order to establish the parameters for the construction of a discipline-oriented metadata model, it is also necessary to include the functionality requirements of the metadata elements, their level of quality and the stakeholders involved. This is what we will address in the following sections.

3.1 RESEARCH DATA HAVE NO VALUE OR MEANING WITHOUT CONTEXT

Commonly, data does not speak for itself, this is the main reason it needs additional information, expressed by metadata and documentation – or representation information, in the perspective of OAIS - , that allow researchers, other than those who created it, to understand the context of the data, i.e., is the

environment and conditions in which the contents are created, processed, received, and its relationship with others research object. For instance, a sequence of naked numbers – as 03, 11, 27, 53 - could mean anything, without the layers of context that explain, how the data is collected or generated who stated the data, what type of data is it, when and where it was stated, what else was going on in the world when this data was stated, and so forth. According to this perspective, the types of contextual information that may be relevant are virtually limitless and its richness contribute strongly to add value to the data. Lorentz (2018, p. 2-3) defends "Contextualization is crucial in transforming senseless data into real information – information that can be used as actionable insights that enable intelligent corporate decision-making". In this sense, contextual information is an instance of information itself and can be thought as a meta-information that be expressed as documentation and metadata (LORENTZ, 2018).

About the spaces of representation/signification that could be created by contextualization, Chistine Borgman (2015) an often-cited authority on scholarly communication, argues that "data have no value or meaning in isolation; they exist within a knowledge infrastructure -- an ecology of people, practices, technologies, institutions, material objects, and relationships" (Borgman, 2015, p.4). Thus, contextualization not only helps to comprehend the data in the best possible manner, but it also helps in gathering and storing data in ordered groups and sequences. Furthermore, "the publication of the processing steps of the research offers essential context for interpretating and reusing data", argument Goodman and collaborators (2014, p. 3), enhancing that is necessary to indicate how intermediate and final products are generated. However, for a complete theoretical and practical understanding of contextualization information in research data management scenarios, it is necessary to relate it to the concepts of data provenance and lineage.

3.2 PROVENANCE AND LINEAGE: TRACKING WHERE THE DATA COME FROM

Question about the origins of the data and their processing steps are answered through the tracing of the provenance and lineage of data. Those

concepts are critical to improve the repeatability and the verifiability of the research results (Hunter, 2005). Metadata has a vital importance in this process since the degree of confidence in the data depends on the availability of accurate information about the many variables that establish its provenance and lineage. “For the replication of the data, data provenance is an important facet of metadata”, confirm Schembera and Iglezakis (2020, p. 27), relating provenance to richness of metadata.

The terms data provenance, lineage and traceability are often used interchangeably, because such concepts are about where the data comes from. Depending on the source, their definitions may slightly differ (MAYERNIK *et al.*, 2013); although, in practice, the resolution of data provenance and lineage usually is just one process: tracing how data has evolved - e.g., computational processes done on it - is inseparable from understanding where data comes from. This is because various records of the inputs, entities, systems, and processes executed on the data are crucial to both trace the evolution of the data and to understand the origins of the intangible, perfectly replicable asset that is data (CHOUDHURY *et al.*, 2018). Those records are not only captured by manual processes or by scientific instruments, in the daily life of laboratories the workflow systems play an important role in in recording the steps that testify the traceability of the datasets.

In the realm of eScience, digital technologies, experimental techniques and scientific instrumentation have changed the way the scientist works. In the laboratory activities, workflow technologies represent an increasingly important component of the scientific process since they capture the chain or pipeline of processing steps used to generate research data and secondary products. The workflow systems also assist researchers to describe and conduct their experiments in a repeatable, verifiable, and distributed way; those computational programs enable also the scientist to track the sources of errors, anomalies or faulty processing during the experimental flow (HUNTER, 2005). “One of the major aims of […] web-service based workflow systems, is to relieve the effort required by scientists to capture the precise provenance metadata demanded by scientists in order to validate scientific results and enable their duplication”,

summarizes Jane Hunter (2005, p. 4). In this sense, workflow is a system that permits a prospective vision, defining plans and execution steps for a desired processing (Bose; Frew, 2005) while data provenance and lineage – considering that both concepts are about where the data come from - are retrospective records like an audit trail; they may be defined as a data life cycle that includes data origins and where it moves over time; they describe what happens to data as it goes through diverse processes and help provide visibility into the analytics pipeline and simplifies tracing errors back to their sources; more simply, Bose and Frew (2005) highlight that those records describe the relationships between data products and data transformation after processing has occurred, providing a historic vision of the data and its origins. Beyond the source of observation or information, as highlighted by Clark and Clark (1995), cited by Hunter (2005), in a more practical perspective, the lineage includes data acquisition and compilation method, conversions, transformations and analyses, along with the assumption and criteria applied at any stage of the data life cycle.

Certifying the importance of the data track in building the data trust, W3C has formalized a standard to capture provenance information in a structured way. The PROV standard defines a data model, serializations, and definitions to support the interchange of provenance information on the Web. In this context provenance one has the "information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness" (W3C Working Group, 2013).

Recording data traceability as part of data management services is a complex task where metadata is the main performer. “Among the key services that institutional data management infrastructures must provide are provenance and lineage tracking and the ability to associate data with contextual information needed for understanding and use”, emphasize Mayernik *et al.* (2013, p. 1). Besides, for the replication of the data, data provenance and lineage are an essential facet of the metadata. In fact, information about every processing step has to be assigned and metadata related to the technical dependence as well. Synthetizing the complexity of register the information encompassed by tracking

the data origin, Jane Hunter (2005, p. 5) observes: “Capturing precise lineage data can be a very complex process, particularly if the metadata captured at each stage during the workflow is inadequate or ambiguous.”

3.3 The functionality of metadata in the context of research data

In the data curation context, the metadata set must preferably include much more information that goes far beyond the general descriptive conception mostly presented by more traditional metadata schemes, such as Dublin Core or even DataCite. A relevant data description and representation needs to comprise features that allow some relevant **functions** to be activated and achieved, such as persistent identification, retrieval, access, preservation and interoperability; besides that, significant information must be read and interpreted by human beings or by computational programs, as usage information, licenses, permissions, rights associated to the research object, context. Thus, a set of rich metadata should provide an information environment that allows the data to be findable, understandable, replicable and reusable in a discipline-specific context (Schembera; Iglezakis, 2020).

Generalizing the value of the metadata in the context of information system, Jane Greenberg (2017, p. 22) observes that “Beyond labelling and categorization, metadata can more universally be thought of as **value-added language** that serves as an integrative layer in an information system”. This abstraction connects the research object to a desired activity, as information retrieval or preservation, or to levels of contextualization necessary to data management. Thus, the activity related to building, improvement, maintenance and application of metadata schema becomes a relevant task of the research data information system sometimes called **metadata management** and may be included as part of the data curation.

The **diversity of** research cycles highlights the essentiality of the precise levels of descriptions. No matter where research is conducted - on a computer, in the field, in the library or in the laboratory – there will be some similarities in what a researcher might want to know about the data used or created by some experiment or observation originally performed by another researcher group. He

would clearly need to know the description of data formats and structure, labels or descriptors for the dataset, and any relevant unit of measurement; however, sometimes it is very relevant as well to know about the conditions of access, reuse, rights associated to the dataset and ethical issues involved. Certain type of research demands specific information, for instance: in the case of **observational research**, the metadata set could further include information about the instrument used or the technique used for observation, calibration of the instrument, and the local and data the observation is taken from, the weather conditions and so on; for **experimental research** it is critical to represent by disciplinary-specific metadata the processing steps and its workflow; for **computational research,** or even research that involve code or software, the metadata set could also include information about the configurations, set parameters, operation system, computational model, details about the computer and necessary specific device, and mainly the scripts and codes used to generated the data (CHIU *et al.*, 2019).

On the other hand, for accurate findings (retrieval) the metadata schema must be able to support the formulation of discipline-specific search criteria. In this sense, the metadata elements need to present a necessary degree of granularity, understood as the capacity to formulate precise queries, as this allows the specification of processes, parameters, variables, methods and instruments; these issues are relevant to the research field being familiar to its research community and contribute to retrieve the precise segment of the dataset the research is looking for. However, what is observed is that the existing metadata currently poorly reflect information data needs and are the biggest obstacle in retrieving relevant data. Therefore, locating and finding proper data for synthesis is a key challenge in a daily research practice; even more critical is to realize that large data repositories tend to use metadata standards with domain-independent metadata elements that cover search interests only to some extent (Löffler *et al.*, 2021). Besides domain-specific metadata standards, the usage of controlled vocabularies is a relevant building block for successful dataset retrieval.

It is important to remark the core idea of the long-term value as well of the datasets in the context of (meta)data curation: data are often reused by researchers (and other stakeholders), different from those who collected and managed them, dispersed in space and time. Next generations of data users – humans and systems – must be able to understand and access the information that the preserved data represent. “Individual scientists and research teams, as well managers and communication specialists need to be aware of this and document their work accordingly”, observe Ciambrella, Mcmahon and Fekete (2017, p. 1), extending the spectra of players involved in (meta) data curation. It is also necessary to keep in mind that metadata is also vital to unique and global identification of the research object, to management concerning legal and ethical issues, and rights associated to the objects. Metadata is also used to record information (meta-metadata) about the metadata itself, such as identification, namespace, data of creation, versions etc (Higgins, 2007). Additionally, the set of metadata elements - a metadata schema - comprises formally description of the syntax and structure of metadata elements in the domain of a specific scientific community that adopts it as standard, defining rules of application.

Ross Harvey (2010) aligns the information that the data curator and several other players - including scientific instruments and software, as workflow system - add to research objects in the form of metadata and documentation; those information permit that the dataset can be effectively managed, identified, founded, accessed, and reused now and in the future. Based on some items listed by Harvey (2010), we introduce below a **classification for the metadata functions** for research objects.

Representation functions

- **Represent** (or codify) and contribute to **organize** the **knowledge** of a disciplinary domain in a relevant way that are relevant to the research field and are familiar to its research community.
- **Transform** data object into **information object** (CCSDS, 2012) or **epistemic object** (RHEINBERGER, 1977).
- Describe **what is needed to re-present** to the research object at the standard required by the users (HARVEY, 2010)

Description functions

- **Identify** research objects uniquely, globally and persistently (considering persistent identifiers as part of metadata schema elements)
- Clearly **describe** what the digital object are.
- Record the **bibliographic information** of the research object, allowing the research object to be cited in a standardized way.
- Clearly identify the **technical properties and structure** of the data comprising the research digital object.

Management functions

- Support the **management** of the complete **life cycle of the research object**.
- Inform **usage information** as rights, licenses, and permissions.
- Describe **what can be done** to the research object.
- Identify who is **responsible** for the **management** and **preservation** of the research object.
- Describe the **provenance of metadata** schema (informed by meta-metadata).
- Support the **evaluation** the research object collection.
- Underlie disciplinary **research information system.**
- Support the long term **preservation** of research object (PREMIS, 2015)

Retrieval functions

- Support the **formulation of queries** with the appropriate level of granularity and precision that takes into account the orientation by discipline.
- Support the **localization** and **retrieval** of research object.
- Support the **selection** and **evaluation** of retrieved datasets.

Relation functions

- Maintain **reliable links** to the research object.
- Provide **link** from the research object **to others related objects** (journal articles, models, software, datasets etc.) to make visible the ecosystem where the object is located and its relationship with others research object.

Interpretability functions

- Enhance the interpretability of the research data in different space of signification for humans and computational agents now and in the future.
- facilitate the reuse of research objects by their creators and by other researchers.
- Support the machine actionability of the research object.

Provenance and contextualization functions

- Record the history of the research objects (provenance, traceability and lineage)
- Allows the specification of processes. parameters, variables, methods and instruments that are relevant to collect/generate, process and analyze the research field and are familiar to its research community.

Preservation functions

- Support long-term preservation strategies.
- Describe the technical dependence of the research object.
- Provide semantic and structural representation information that permit interoperability in the future.

Trustworthiness function

- document the authenticity, integrity and trustworthiness of the digital object.

Interoperability function

- Support the integration of data with other data and with applications or workflow for analysis, storage and processing (Mons *et al.*, 2017); the interoperability may request frameworks and data model semantically richer as RDF and OWL.
- Beyond the set of metadata elements, a metadata schema comprises as well as a formally description of the syntax, semantic and structure of metadata elements in the domain of a specific scientific community that adopts it as standard defining rules of application. This understanding reflects, for instance, how much the structure schema meets the information needs of a specific scientific community, what can be evaluated by meeting the functionalities required by the area.
- In order to fulfill those functionalities in a certain domain, it is necessary to select a set of metadata quality attributes that make it possible to carry out them within the scope of a research data management platform.

3.4 METADATA QUALITY

A relevant feature for metadata schemas to fulfill their functions is related to the conditions that give them quality and provenance (which is recorded by the meta-metadata). Despite the wide consensus on the need to construct high quality metadata formats, specially to informational research objects, there is less agreement on what high quality means and even less on how it should be measured. The model proposed by Ochoa and Duval in 2009 considers quality as the measure of fitness for some important purposes as the functionality and information needs of a specific domain. Thus, in our scope metadata quality is understood as how much the syntactic and the semantic structure of metadata elements reflect the information needs of a specific scientific community that can be evaluated by meeting the functionalities required by the disciplinary domain. The quality requirements most commonly identified by authors dedicated to the subject were examined and those most relevant and useful for the composition of the model were selected (Bruce; Hillmann, 2004; Niso, 2007; Ochoa; Duval, 2009; Király, 2017). For the purposes of the present study, we align the following quality criteria (Table 1):

Table 1 - Description of Metadata Qualities

METADATA QUALITY	DESCRIPTIONS
Support of functional requirements	each data schema is created for supporting a set of functionalities, such as searching, identifying, managing or describing research object. The metadata elements support one or more of these

	functionalities, and their existence, content and quality have an impact of these functionality.
Completeness	metadata should be complete in two senses: first, the elements set used should describe the target system with all the needed information; second, the elements set should be applied to target object population as complete as possible.
Granularity	metadata should describe the research object data at the level of detail required by processes, workflows, instruments, etc. that are intrinsic to the domain and necessary for the correct interpretation of the data.
Accuracy	metadata should be accurate in the way it describes objects, meaning that it should be unambiguous and the information should be correct.
Provenance	provenance information of metadata is a useful base for quality evaluation and should be available for the metadata itself, for instance, which institution has created the schema.
FAIR compliance	as a digital object itself the metadata schema must adhere to FAIR principles, especially the "I" of **Interoperability**; that means that the metadata information should be technically interoperable and understandable without knowing the context.
Adaptability	ability of a metadata schema to dynamically adjust to changes and demands from the disciplinary community.
Machine actionability	ability of metadata provides information to a computational agent on the behalf of a human user or a system, permitting a computational system to use services on data without human intervention. In order to enable machine processing, metadata should be manifested in structured formats such as XML or JASON, but there is an important difference between "readable" and "actionable" as point out Batista *et al.* (2022, p. 2): "the latter form indicates a shift in maturity status, which allows a software agent to exploit the formal representation and understand its content rather than just obtaining a string without any context, as it occurs in a read action".

Source: The Authors, 2023

3.5 CONTEXT AND PROVENANCE: A PROPOSITION OF A SET OF DISCIPLINARY DATA ELEMENTS

As seen previously, several layers of description, documentation and contextualization can be applied to datasets and to others research objects, from granular, discipline specific metadata to metadata for discovery systems and aggregations to descriptive documents accompanying the data set. In short: the application of metadata provides information and context about the data that are not always apparent for the data alone (Hunter, 2005). To address this complexity, we propose the following set of metadata elements that focus on the representation of the contextualization and provenance of the disciplinary research objects (Table 2).

Table 2 - New Metadata Categories

TYPE	DESCRIPTION
Descriptive metadata	Although metadata models and standards exist for data in general like Data Cite (Schembera; Iglezakis, 2020), those schemes do not address the idiosyncrasies of the deep segmentation of contemporary science. The descriptive nature of those models serves other important purposes, as citation or persistent identification, for that they are included as part of the disciplinary metadata schemes.
Bibliographic metadata	Provide information about the dataset from a descriptive or bibliographic point of view: persistent and global identification, authors and other contributors and its standard identifications, title, versions, keywords, description, abstracts, significant dates, funders, funding/award number, corporate authors, bibliographic references, subject area of the research, how to cite etc. In some cases, general metadata, such as Dublin Core providing guideline for basic citation elements (hunter, 2005). Assigned mainly by data librarian or data steward.
Automatic metadata	Metadata assigned at the time of data capture, generally by automatic processes such as those provided by instruments, code or workflow software that generate the data. File format, georeference, time data stamp, are some examples of metadata generated by automatic process. A simple example of that is a scene captured by a photographic camera that, together with the image, automatically adds data such as image description (format, resolution, color depth etc.), coordinates, date, and time.
Contextual Metadata	It is very difficult to interpret and hence reuse data without context describing what the data are and how they were obtained or generated. The concept of provenance, as seen in section 6 is of key importance in materializing the contextualization of data, since it can be thought as the "information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness" (W3C Working Group, 2013). In other words, it is the record of information of all steps, processes, peoples (institutions or agents), documents and others research object (datasets, journal articles, projects etc) as well, that were involved in generating a piece of data. "The higher the quality of provenance information, the higher the chance of enabling data reuse" summarized Alyssa Goodman and collaborators (2014, p. 3). So, contextual metadata, are blocks of information that describe the various facets of the disciplinary environment where research activities are carried out, more specifically in which data are collected or generated and their semantic relationships with relevant entities. Contextual metadata are essential to interpretability of data in different spaces of signification and hence increase the reuse, the potential of replication of the data, the reproducibility of the scientific experiments and improve the verifiability, traceability of the results. They are assigned by various players, including those generated by workflow software.
Inherent metadata	These are metadata that record processes intrinsic to very specific domains. Each discipline's data have nuances that may require specific considerations; those data may have context or complexities that require discipline-specific knowledge to correctly describe them (Choudhury, 2018). In this sense, inherent metadata comprise metadata that describe data characteristics specific to a particular discipline, problem, or research situation, such as information about variables, parameters, methods, techniques, procedures, manipulations, and information about the instruments commonly used in the research domain as calibration, setting, fabricator or even details of design and construction. These features are essential, for instance, for accurate retrieval of specific segments from large datasets; but for that, additionally, is necessary the development and/or application of discipline-specific vocabularies, taxonomies, or ontologies.

TYPE	DESCRIPTION
	Those elements of metadata should be assigned by the researchers involved with the experiment.
Process metadata	Research data are rarely used as soon as they are collected or generated by an instrument, they usually go through several stages of processing that make them more suitable for the purposes for which they are proposed. Legitimacy and trust in data depend, in some degree, on knowing where the data comes from and understand how it is processed. Therefore, it is essential to describe the methodology and process to collect the data. The various records of the inputs, entities systems, and process executed on the data are crucial to both trace the evolution of data and understand the origins of the intangible, perfectly replicable asset that is data. The combination of methods, processing, and data analysis in the context of an experiment is denoted by the term **workflow**, which minimally indicates how intermediate data and outputs, and final results are generated. In many cases, researchers use workflow software packages to run experiments and record what has been done. The information used and captured by the workflow is part of the **provenance** of the data, as well as the workflow software, its version and the configurations applied. “Moreover, the software used for the processing step builds its own entity [metadata element], since it is critical for understanding the conducted research”, conclude Schembera and Iglezakis (2020, p. 28).
Technical metadata	The different steps of data process are commonly carried out within a computational environment or cyberinfrastructure which should be described in proper details to support reuse and reproducibility. In addition to workflow software, other software, codes, instruments and relevant computational resources play several important roles in scientific research activities. That technical framework should be described via a specific type of information called technical metadata. To document the provenance of data involving software, technical metadata should record information on the script or code used to collect or generate those data; metadata could also include information on configurations, parameters or setting, properties of the computers and others hardware used, and as well as operating system (Chiu *et al*., 2019). Goodman (2014, p. 5) emphasizes that “publishing the source code and its version history is crucial to enhance transparency and reproducibility” of the datasets. For code already published it is important to inform, when available, the permanent identifier. It is also important to document how the data products are organized – names, folders and files in a database, column header, coding book – as well as the file size and formats, when not provided by instruments used (Sayão; Sales, 2015). Those elements are, typically, registered by the computational staff.
Relational Metadata	Information about the purpose and methods for collecting the data and results of analysis, products acquired and generated in the course of the research could be found in relevant publication as the project document that gives origin to the dataset, articles, data papers and other research object as different source data used and its versions, software, algorithms, models etc.; the record of those related publications by means of **relational metadata** creates an ecology of data and information around data that enhances its understanding and context. Those connections could be extended to persons, labs, funders, scientific programs, resources of interest, and so on. Hence, connecting research objects to each other opens a world of new expressions, resulting tools and opportunities for discovery; and becomes research objects more easily founded and more readily evaluated. “The impact of science is more readily determined when objects and their connections can be visualized and assessed” reminds us Mike Conlon (2011, p. 57). Those elements are generally assigned by the **data librarian**.

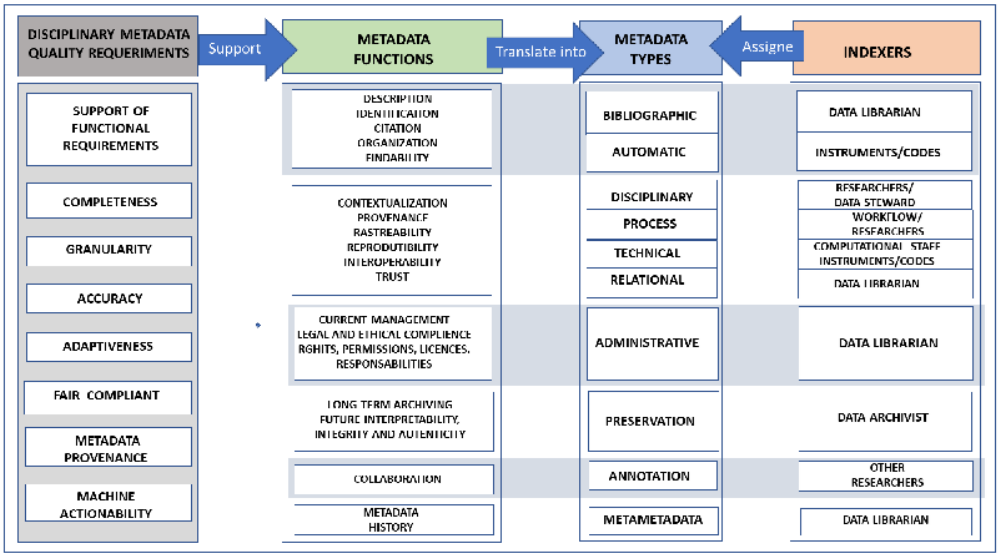
TYPE	DESCRIPTION
Administrative Metadata	In addition to describing the data and its context, it is crucial to describe who as well can access and use the data, how the data could be used, and privacy and intellectual property issues associated to them (Choudhury, 2018). A good metadata schema includes a clear statement of the conditions and terms of use for research digital object defined by licenses; includes, as well, the copyright status of the object – whether it is in the public domain or it is copyright protected; may also include right holder requirements for use the resource, and user attribute required for authorized use of a resource. Administrative metadata may also document information concerning user access, user tracking and reuse, also known as "**use metadata**".
Preservation Metadata	Continuous-value research data, once published, should remain available forever giving support to reproducibility and the science self-correction principles, to new discovery and to unexpected reuses (Gray *et al,* 2002). Furthermore, a researcher in the future needs to rely in data collected or generated by other researchers to continue his research, so this data must preserve its qualities of trustiness, integrity, and authenticity. Considering that the NISO (2007) report preconizes that the metadata schemes should support the long-term management, curation, and preservation of digital objects in collections, preservation metadata are design to support the long-term retention and stability of the digital objects, mainly its archival qualities. It is important to highlight that a significant part of preservation metadata belongs also to other groups as technical and administrative metadata, but in the temporal dimension, these metadata assume the function of supporting the stability of the digital objects over time. It may include detailed technical metadata, as well as information related to the objects context and relationships, custody, processing, archiving. In this sense, the PREMIS Data Dictionary for Preservation Metadata (2015) offers a dictionary which presents a detailed set of semantic units necessary to implement preservation metadata elements to enrich disciplinary schema. Those elements should be registered by the **data archivist**.
Annotation/collaboration metadata	Annotation refers to adding information mainly by other researchers to data to elaborate on them. Besides, annotation of existing data provides an important form of communication and collaboration among researchers. "Annotation can be applied to all kinds of digital data – to describe, correct, interpret, extend and classify those data" (Harvey, 2010, p. 208). In the context of our model, annotation enriching digital content with personal keywords without modifying the data record.
Metametadados	As digital objects themselves, a good metadata records should have the qualities of good objects including provenance, persistence, interoperability and unique identification; some of them are synthetized in accordance with the FAIR Principles. **Metametadata** documents information concerning the creation, alteration and version control of the metadata itself. Therefore, the institution that created the schema should provide sufficient information to allow the user to assess its veracity, including how it was created and what standards and vocabularies were used, in order to consider that, some metadata schemas include within them sets of metadata elements for describing the metadata records themselves like EAD, LOM, and MODS (NISO, 2007).

Source: The authors, 2023

Figure 1 seeks to gather all the variables – quality requirements, metadata functions and metadata types - demanded by the disciplinary community; the

figure includes also the role of indexers – human or machine - who assign metadata records to the object.

Figure 1 – Metadata quality, functions, types and indexers



Source: Elaborated by the authors (2023).

After established the main categories of metadata that will compose the model, it might be interesting to move towards a general model for implementing a disciplinary metadata schema. In that direction we will go next.

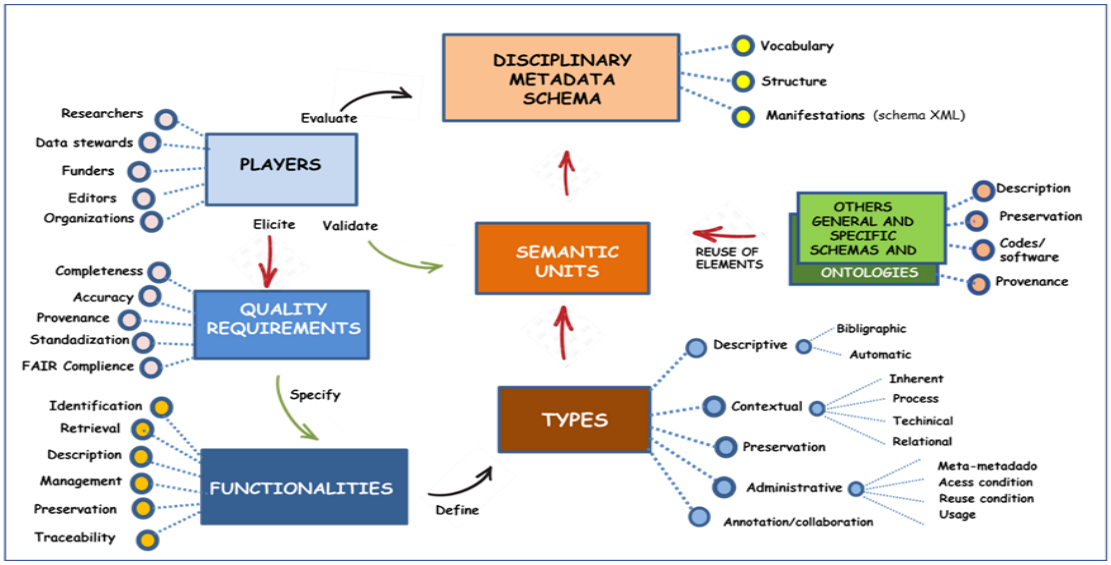
4 A PROPOSITION OF AN IMPLEMENTATION MODEL FOR AUTHORING A METADATA DISCIPLINARY SCHEMA

Even though there are some disciplines with elaborated and accepted metadata standards like DDI for social science and Darwin Core for biodiversity, most of the existing metadata schemas are generalist models applicable to large spectra of datasets, even in a disciplinary domain, although disciplinary interests are better covered by metadata elements of domain-specific standard (LÖFFLER *et al.,* 2021), as confirmed by Schembera and Iglezakis (2020 p. 3): “What missing in all these schemes are the discipline-specific description of the observed system and parameters of the observation itself as well as a possibility to track the provenance of the data with all relevant methods, utilities and parameters”. The growing disciplinary specialization of the contemporary science requires more and more suitable metadata schema for the novel emerging

discipline occurring at an accelerated pace in the eScience environment. The postulates of the contemporary science demand agile methods for building appropriate metadata schemas, rich enough to describe its process, variables, parameters, methods precisely and with desired granularity, that have consensus from the involved community, and make datasets compliant with FAIR principles (Wilkison *et al*., 2016).

Figure 2 describes a possible workflow for building disciplinary metadata schemas for research datasets; this workflow considers an extensive community participation, an elicitation of qualities and functionalities that permit functions such as retrieval and preservation and processes description – which defines the types of metadata and its translation into semantic units that are enriched by incorporation and reuse of general schema (as DC, DataCite), specific one (as Premis, CodeMeta) and ontologies and taxonomies (as PROV). In addition, the workflow considers schemas, ontologies and other specific and general terminological tools already established, such as Dublin Core, Premis and ontologies like W3C PROV for provenance.

Figure 2- Workflow for building disciplinary metadata schemas for research data



Source: Elaborated by the authors (2023).

- **PLAYERS** - This step considers the different interests of the various players in the metadata schema involved in data management. The emphasis of this mapping falls on the requirements established by the researcher community and data stewards. In addition to those players, funders, research organization and scientific editors are also regarded.
- **QUALITY REQUERIMENTS -** Establishes the quality requirements necessary for implementation of the level of representation and functionality demanded by the

disciplinary community, for instance: completeness, granularity, accuracy, and FAIR compliance.

- **FUNCTIONALITY** - Determines the functionalities that must be fulfilled by the metadata scheme, for example: identification, bibliographic description, accurate retrieval, traceability, information about context and provenance, **machine actionability**, support for dataset management and preservation.
- **TYPES OF METADATA ELEMENTS** - Outline the types of metadata elements needed to describe, manage, and represent disciplinary datasets: descriptive (intrinsic, bibliographic), contextual (disciplinary, technical, process), preservation, administrative, metametadata.
- **REUSE OF OTHER METADATA SCHEMA, ONTOLOGIES AND TAXONOMIES** – The disciplinary metadata schema, sometimes, requires reuse of large numbers of metadata elements from other disciplines to enrich its capacity of representation. In addition to the metadata elements for description of particular research processes, the workflow reinterprets the notion of "reuse" applying it in the construction of disciplinary metadata schema. This is carried out through incorporation of existing metadata standards for general description (dc and DataCite) and for more specific technical information as the description of codes/software (example CodeMeta) and information that support long-term preservation (PREMIS); other terminological instruments such ontologies and taxonomies may contribute with the definition of relevant semantic units as, for instance, ontology provenance PROV, provided by W3C.
- **SEMANTIC UNITS** - definition of the semantic units that will be translated into elements of the disciplinary metadata schema to be built; in this phase specific elements from other metadata schemes and other representation tools, such as ontologies and taxonomies might be incorporated to the new schema. For instance: provenance ontology and metadata for computational code, preservation metadata and descriptive metadata (DC, DataCite Schema etc.)**.**
- **METADATA ELEMENTS** – as a result, the list of elements that will compose the **disciplinary metadata schema** are obtained in the form of a structured vocabulary that can be manifested in different formats, for example, structure formats as XML or JSON, more suitable to automatic processing or in RDF or OWL semantically more expressive. The metadata elements should provide an appropriate description of a dataset and its research process with rich information.

5 FINAL REMARKS

The contemporary science becomes increasingly dependent on technology that is essential to support its data intensive knowledge generation, integrative analysis and collaborative and distributed characteristics. In this scenario globally distributed research groups work together capturing and generating, curating, sharing and analyzing, by means of innovative methodologies, massive data sets looking for resolution of complex and interdisciplinary problems in science, humanities, business and in art and culture as well. This current condition places the construction of forms of representation such as metadata schemes, taxonomies and ontologies at the heart of eScience. The fragmentation of science and the necessary interdisciplinarity require metadata schemes that are both specific, whose depth goes beyond

disciplinarity, and modular, in the sense of receiving contributions from other more specific and general schemes as well.

Nowadays there are a large number of metadata standards available, but they do not face the fragmentation of knowledge and the desirable interdisciplinarity of the contemporary science (De Smedt; Koureas; Wittenburg, 2020). Standardized discipline-specific metadata schema may exist in some privileged areas, while most need to be built. In general, we can say that the existing metadata standards poorly reflect information needs of the e-scientists, and the current metadata in research object repositories match scholarly search interests only to some extent. This problem demands the effort of many interested actors to define the actual functionality and level of quality and representation granularity required by the communities involved.

Do the existing metadata standards reflect the disciplinary demands of the e-scientists regarding context and provenance? Are the metadata standards utilized by repository truly useful for specific research object recovery? Is the metadata management part of the research object? And finally: In what extent the current metadata in data repositories match scholarly search interest?

Those are some questions raised in this study that highlight the necessity to develop strategies and methodologies for the development of disciplinary metadata schemas that stand out in three main aspects: 1) a **social and administrative** aspect that involves consultations with the communities involved, organizational and political alignments, disciplinary needs, roles, and metadata quality; 2) **technical** aspect involving selected standards, data modeling techniques, appropriate schemas for modulation and reuse, software tools, etc.; 3) and a **metadata management** system that supports the construction, implementation and maintenance of disciplinary metadata schemes, and that take part of the data curation processes.

Considering those complexities, the underlying objective of the proposed study was synthesizing in a model – as an abstract form to concentrate a knowledge - some principles for capturing, reusing and generating disciplinary metadata involving the diverse community and players, that reveal the context and provenience of the data. The proposed authorship model is configured as a

way to help data stewards in the construction of metadata standards for areas that don`t have their own standards, in addition, It can help to complete the metadata set already given aims to turn the research object in a FAIR digital object.

AGRADECIMENTOS

To the funding agencies CNPq and FAPERJ, for the financial support for the development of this research; to Professor Hagar Espanha Gomes, for providing us with valuable suggestions; and to our friend Teodora Marly Gama, for her precious normative review.

REFERENCES

BATISTA, Dominique; GONZALEZ-BELTRAN, Alejandra; SANSONE, Susanna-Assunta; ROCCA-SERRA, Philippe. Machine actionable metadata model. **Scientific Data**, [*S. l.*], v. 9, n. 1, 2022. Disponível em: https://www.nature.com/articles/s41597-022-01707-6. Acesso em: 20 fev. 2023.

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

BOSE, Rajendra; FREW, James. Lineage Retrieval for Scientific Data Processing. **ACM Computing Survey**, [*S. l.*], v. 37, n. 1, 2005.

BRUCE, Thomas; HILLMANN, Diane I. The continuum of metadata quality: defining, expressing, exploitation. *In*: HILLMANN, Diane I.; WESTBROOKS, Elaine L. (ed.). **Metadata in Practice**. Chicago: ALA Editions, 2004. Disponível em: https://ecommons.cornell.edu/handle/1813/7895. Acesso em: 16 fev. 2023.

BURKE, Peter. **Uma história social do conhecimento**: de Gutenberg a Diderot. Rio de Janeiro: Zahar, 2003.

CHIU, Chen; BLAKE, Mara; BOEHM, Reid; FEARON Dave. **Metadata for effective research data management**. Center of Open Science. 2019. Disponível em: https://osf.io/7q4cu. Acesso em: 06 dez. 2022.

CHOUDHURY, Sayeed; COWLES, Esme; CROFT, Holly; ESTLUND, Karen; FARY, Michael; FAUSTINO, Gracy; HAUSER, Thomas; LINTON, Anne; LYNCH, Clifford; MENARD, Karen; MINOR, David; MONACO, Gregory E.; NOONAN, Daniel; SHREEVES, Sarah; ULATE, David; WATERS, Natalie. **Research Data Curation**: A framework for an Institution Approach. Louisville,

CO: ECAR, 2018. Disponível em: https://www.researchgate.net/publication/325093683_Research_Data_Curation_A_Framework_for_an_Institution-Wide_Services_Approach. Acesso em: 16 fev. 2023.

CIAMBRELLA, Massimo; MCMAHON, Kevin; FEKETE, József I. **Metadata in geological disposal of radioactive waste**: The RepMet Initiative. Albuquerque, NM: Sandia National Lab., 2017. Disponível em: https://www.osti.gov/servlets/purl/1479240. Acesso em: 16 fev. 2023.

CONLON, Mike. The objects of science and their representation in eScience. *In*: **Changing the Conduct of Science in the Information Age.** 2011 p. 57-58. Disponível em: https://www.nsf.gov/pubs/2011/oise11003/oise11003_10.pdf. Acesso em: 20 fev. 2023.

CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEM - CCSDS. **Reference Model for an Open Archival Information System (OAIS)**. Washington, DC: CCSDS, 2012. Magenta book (CCSDS 650.0-M-2). Disponível em: https://public.ccsds.org/pubs/650x0m2.pdf. Acesso em: 20 fev. 2023.

DE ROURE, David; JENNINGS, *Nicholas R.; SHADBOLT, Nigel R.* The Semantic Grid: a future e-Science infrastructure. *In*: BERMAN, Fran; FOX, Geoffrey; HEY, Anthony J. G. (ed.). **Grid Computing. Making the Global Infrastructure a Reality**. Chichester, West-Sussex, UK: John Wiley & Sons, 2003. p. 437-470.

DE SMEDT, Koenraad; KOUREAS, Dimitris; WITTENBURG, Peter. FAIR Digital Objects for Science: From data pieces to actionable knowledge units. **Publications**, [*S. l.*], v. 8, n. 2, 2020. Disponível em: https://www.mdpi.com/2304-6775/8/2/21?type=check_update&version=2. Acesso em: 06 jan. 2023.

GOODMAN, Alyssa; PEPE, Alberto; BLOCKER, Alexander W; BORGMAN, Christine L.; CRANMER, Kyle; CROSAS, Merce; STEFANO, Rosanne Di; GIL, Yolanda; GROTH, Paul; HEDSTROM, Margaret; HOGG, David W.; KASHYAP, Vinay; MAHABAL, Ashish; SIEMIGINOWSKA, Aneta; SLAVKOVIC, Aleksandra. Ten simple rules for the care and feeding of scientific data. **PLoS Computer Biology**, [*S. l.*], v. 10, n. 4, 2014. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3998871/. Acesso em: 29 jul. 2022.

GRAY, Jim; SZALAY, Alexander S.; THAKAR, Ani R.; STOUGHTON, Christopher; VANDENBERG, Jan. **Online scientific data curation, publication, and archiving**. Redmont, WA: Microsoft Corporation, July 2002. Disponível em: https://arxiv.org/ftp/cs/papers/0208/0208012.pdf. Acesso em: 29 jul. 2022.

GRAY, Jim; LIU, David T.; NIETO-SANTISTEBAN, Maria; SZALAY, Alexander S.; DEWITT, David; HEBER, Gerd. **Scientific data management in the coming decade**. Redmont, WA: Microsoft Corporation, Jan. 2005. Disponível em: https://arxiv.org/ftp/cs/papers/0502/0502008.pdf. Acesso em: 19 jul. 2022.

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

GREENBERG, Jane. Big Metadata, Smart Metadata, and Metadata Capital: Toward Greater Synergy Between Data Science and Metadata. **Journal of Data and Information Science**, [*S. l.*], v. 2, n. 3, p. 19-36, 2017. Disponível em: https://sciendo.com/pdf/10.1515/jdis-2017-0012. Acesso em: 06 dez. 2003.

HARVEY, Ross. **Digital Curation**: A How-To-Do-It Manual. New York, NY: Neal-Schuman Publishers, 2010.

HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin. **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.

HIGGINS, Sarah. **What are metadata standards?** Edinburgh: Digital Curation Centre, 2007. Disponível em: https://www.dcc.ac.uk/guidance/briefing-papers/standards-watch-papers/what-are-metadata-standards. Acesso em: 06 dez. 2022.

HOBSBAWM, Eric. **The Age of Extremes:** The Short Twentieth Century, 1914-1991. London: Abacus, 1995.

HUNTER, Jane. **Scientific Models**: A user-oriented approach to the integration of scientific data and digital libraries. 2005. Disponível em: https://core.ac.uk/download/pdf/14984655.pdf. Acesso em: 20 mar. 2023.

KHAN, Robert, WILENSKY, Robert. **A framework for distributed digital objects service**. 1995. Disponível em: http://www.cnri.reston.va.us/home/cstr/arch/k-w.html. Acesso em: 06 dez. 2022.

KHAN, Robert, WILENSKY, Robert. A framework for distributed digital objects service. **International Journal on Digital Libraries**, [*S. l.*], v. 6, n. 2, p. 115-123, 2006.

KIRÁLY, Péter. Towards an extensible measurement of metadata quality. *In:* INTERNATIONAL CONFERENCE ON DIGITAL ACCESS TO TEXTUAL CULTURAL HERITAGE - DATeCH2017, 2., 2017, Göttingen, Germany. **Proceedings** […]. New York: ACM Digital Library, 2017. p. 111-115.

KITCHIN, Rob. Big data, new epistemologies and paradigm shifts. **Big Data & Society**, [*S. l.*], v. 1, n. 1, 2014. Disponível em: https://journals.sagepub.com/doi/epub/10.1177/2053951714528481. Acesso em: 29 jul. 2022.

KUHN, Thomas S. **The Structure of Scientific Revolutions**. Chicago: University of Chicago Press, 1962.

LISCHER-KATZ, Zack. Studying the materiality of media archives in the age of digitization: Forensics, infrastructures and ecologies. **First Monday**, [*S. l.*], v. 22, n. 1-2, 2017. Disponível em: https://firstmonday.org/ojs/index.php/fm/article/view/7263/5769. Acesso em: 06 dez. 2022.

LÖFFLER, Felicitas; WESP, Valentin; KÖNING-RIES, Birgitta; KLAN, Friederike. Dataset search in biodiversity research: Do metadata in data repositories reflect scholarly information needs? **PloS one,** [*S. l.*], v. 16, n. 3, 2021. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0246099. Acesso em: 06 dez. 2022.

LORENTZ, Alissa. With Big Data, Context is a Big Issue. **Wired**. 2018. Disponível em: https://www.wired.com/insights/2013/04/with-big-data-context-is-a-big-issue/. Acesso em: 06 dez. 2022.

MAYERNIK, Matthew S.; DILAURO, Tim; DUERR, Ruth; METSGER, Elliot; THESSEN, Anne E.; CHOUDHURY, G. Sayeed. Data conservancy provenance, context, and lineage services: Key components for data preservation and curation. **Data Science Journal**, [*S. l.*], v. 12, p. 158-171, 2013. Disponível em: https://www.jstage.jst.go.jp/article/dsj/12/0/12_12-039/_article/-char/ja/. Acesso em: 28 fev. 2023.

MONS, Barend; NEYLON, Cameron; VELTEROP, Jan; DUMONTIER, Michel; SANTOS, Luiz Olavo Bonino da Silva; WILKINSON, Mark D. Cloudy, increasingly FAIR: revisiting the FAIR Data guiding principle for the European Open Science. **Information Service & Use**, [*S. l.*], v. 37, n. 1, p. 49-56, 2017. Disponível em: https://content.iospress.com/articles/information-services-and-use/isu824. Acesso em: 22 mar. 2023.

NISO. **A Framework of Guidance for Building Good Digital Collections**. 3. ed. Bethesda, MD: NISO Press, 2007. Disponível em: https://www.niso.org/sites/default/files/2017-08/framework3.pdf. Acesso em: 16 fev. 2023.

OCHOA, Xavier; DUVAL, Erik. Automatic evaluation of metadata quality in digital repositories. **International Journal on Digital Libraries**, [*S. l.*], v. 10, n. 2-3, 2009. Disponível em: https://www.academia.edu/2808304/Automatic_evaluation_of_metadata_quality_in_digital_repositories. Acesso em: 16 fev. 2023.

PREMIS Data Dictionary for Preservation Metadata. Version 3.0. June 2015. Disponível em: http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf. Acesso em: 15 fev. 2023.

RHEINBERGER, Hans-Jörg. **Toward a History of Epistemic Things**: Synthesizing proteins in the test tube. California: Stanford University Press, 1977.

SANTOS, Luiz Olavo Bonino da Silva (ed.). **FAIR digital object framework documentation**. 2021. Disponível em: https://fairdigitalobjectframework.org/. Acesso em: 06 jan. 2023.

SAYÃO, Luis Fernando; SALES, Luana Farias. **Guia de Gestão de dados de pesquisa para bibliotecários e pesquisadores**. Rio de Janeiro: CNEN, 2015. Disponível em: https://oasisbr.ibict.br/vufind/Record/IEN_b6a823ef451ba363fe2d3f83088db887. Acesso em: 20 mar. 2023.

SCHEMBERA, Björn; IGLEZAKIS, Dorothea. Metadata for Computational Engineering. **International Journal of Metadata, Semantics and Ontologies**, [*S. l.*], v. 14, n. 1, p. 26-38, 2020. Disponível em: https://www.inderscienceonline.com/doi/abs/10.1504/IJMSO.2020.107792. Acesso em: 06 dez. 2022.

SCHWARDEMANN, Ulrich. Digital objects – FAIR Digital Objects: Which services are required? **Data Science Journal**, [*S. l.*], v. 19, n. 1, 2020. Disponível em: https://datascience.codata.org/articles/10.5334/dsj-2020-015/. Acesso em: 20 mar. 2023.

TYBJERG, Karin. Exhibiting Epistemic Objects. **Museum & Society**, [*S. l.*], v. 15, n. 3, p. 269-286, 2017. Disponível em: https://pdfs.semanticscholar.org/a152/1b29555cb7982794a6c80a1e3504ba2d8782.pdf. Acesso em: 06 mar. 2023.

W3C WORKING GROUP. **PROV-Overview:** An Overview of the PROV Family of Documents. Apr. 2013. Note 30. Disponível em: https://www.w3.org/TR/prov-overview/.Acesso em: 16 fev. 2023.

WILKINSON, Mark D. DUMONTIER, Michel; AALBERSBERG, IJsbrand Jan; APPLETON, Gabrielle; AXTON, Myles; BAAK, Arie; BLOMBERG, Niklas; BOITEN, Jan-Willem; SANTOS, Luiz Bonino da Silva; BOURNE, Philip E.; BOUWMAN, Jildau; BROOKES, Anthony J.; CLARK, Tim; CROSAS, Mercè; DILLO, Ingrid; DUMON, Olivier; EDMUNDS, Scott; EVELO, Chris T.; FINKERS, Richard; GONZALEZ-BELTRAN, Alejandra; GRAY, Alasdair J. G.; GROTH, Paul; GOBLE, Carole; GRETHE, Jeffrey S.; HERINGA, Jaap; HOEN, Peter A.C 't; HOOFT, Rob; KUHN, Tobias; KOK, Ruben; KOK, Joost; LUSHER, Scott J.; MARTONE, Maryann E.; MONS, Albert; PACKER, Abel L.; PERSSON, Bengt; ROCCA-SERRA, Philippe; ROOS, Marco; VAN SCHAIK, Rene; SANSONE, Susanna-Assunta; SCHULTES, Erik;

SENGSTAG, Thierry; SLATER, Ted; STRAWN, George; SWERTZ, Morris A.; THOMPSON, Mark; VAN DER LEI, Johan; VAN MULLIGEN, Erik; VELTEROP, Jan; WAAGMEESTER, Andra; WITTENBURG, Peter; WOLSTENCROFT, Katherine; ZHAO, Jun; MONS, Barend.The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, [*S. l.*], v. 3, n. 1, 2016. Disponível em: https://www.nature.com/articles/sdata201618. Acesso em: 22 mar. 2023.

MODELO DE AUTORIA DE METADADOS: DESCREVENDO INFORMAÇÃO SOBRE CONTEXTO E PROVENIÊNCIA DE OBJETOS DE DADOS DISCIPLINARES

RESUMO
No campo da gestão de objetos de pesquisa, há um grande número de esquemas de metadados padronizados disponíveis, mas em geral eles não abordam a fragmentação e a interdisciplinaridade da ciência contemporânea. **Problema:** Existem esquemas de metadados ricos e orientados a disciplinas em algumas áreas-chave, mas em outros casos eles precisam ser construídos. Portanto, um grande desafio para que os objetos de pesquisa atinjam um nível adequado de FAIRificação é que eles sejam descritos por esquemas de metadados que tenham funcionalidades e qualidades que suportem a reprodutibilidade da pesquisa e a reutilização de dados. **Objetivo:** Para abordar essa complexidade, o objetivo desta pesquisa foi definir as funcionalidades e os níveis de qualidade dos padrões de metadados necessários para a gestão de dados de pesquisa FAIR. **Metodologia:** Esta é uma pesquisa teórica e exploratória baseada no conceito de objeto de pesquisa epistêmico/técnico/informacional, considerando quatro eixos: histórico, epistemológico, padronização e aplicação. **Resultado:** Como resultado, foi proposto um modelo de autoria de metadados que se concentrou no registro do contexto e da origem dos objetos de pesquisa. **Conclusão:** Concluindo, o artigo reafirma a necessidade urgente de desenvolver esquemas de metadados disciplinares que não apenas atendam às necessidades específicas dos domínios, mas também garantam a integração interdisciplinar e a recuperação eficiente de dados, promovendo uma ciência mais robusta, acessível e colaborativa.

Descritores: Autoria de metadados. Objeto digital de pesquisa. Linhagem dos dados. Contextualização dos dados.

MODELO DE AUTORÍA DE METADATOS: INFORMACIÓN SOBRE CONTEXTO Y PROCEDENCIA DE OBJETOS DE INVESTIGACIÓN DISCIPLINARIOS

RESUMEN
En el campo de la gestión de objetos de investigación, existe una gran cantidad de esquemas de metadatos estandarizados disponibles, pero en general no abordan la fragmentación y la interdisciplinariedad de la ciencia contemporánea. **Problema**: En

algunas áreas clave existen esquemas de metadatos ricos y orientados a disciplinas, pero en otros casos es necesario crearlos. Por lo tanto, un desafío importante para que los objetos de investigación alcancen un nivel adecuado de FAIRificación es que estén descritos mediante esquemas de metadatos que tengan funcionalidades y cualidades que respalden la reproducibilidad de la investigación y la reutilización de datos. **Objetivo:** Para abordar esta complejidad, el objetivo de esta investigación fue definir las funcionalidades y los niveles de calidad de los estándares de metadatos necesarios para la gestión de datos de investigación FAIR. **Metodología**: Se trata de una investigación teórica y exploratoria basada en el concepto de objeto de investigación epistémico/técnico/informativo, considerando cuatro ejes: histórico, epistemológico, de normalización y de aplicación. **Resultado:** Como resultado, se propuso un modelo de creación de metadatos que se centró en registrar el contexto y el origen de los objetos de investigación. **Conclusión**: En conclusión, el artículo reafirma la urgente necesidad de desarrollar esquemas de metadatos disciplinarios que no solo satisfagan las necesidades específicas del dominio, sino que también garanticen la integración interdisciplinaria y la recuperación eficiente de datos, promoviendo una ciencia más sólida, accesible y colaborativa.

Descriptores: Creación de metadatos. Objeto de investigación digital. Procedencia de los datos. Contextualización de datos.

Recebido em: 25.09.2023
Aceito em: 28.04.2024