THE ROLE OF VOCABULARIES IN THE AGE OF DATA: the question of research data

Journal:	*Knowledge Organization*
Manuscript ID	KO-2022-0003.R2
Manuscript Type:	Article

THE ROLE OF VOCABULARIES IN BIG DATA: the quest of research data[1]

Carlos H. Marcondes

PPG-GOC/UFMG - Postgraduate Program in Knowledge Management and Organization, Minas Gerais Federal University, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, Brazil, ch_marcondes@id.uff.br

[1] This is a revised and extended version of the article with the same title presented at IV Seminário de Pesquisas do grupo MHTX, Nov 2021, http://eci.ufmg.br/iv-seminario-do-grupo-de-pesquisa-mhtx/.

THE ROLE OF VOCABULARIES IN BIG DATA: the quest of research data

Abstract

Objective: This paper discusses the role of vocabularies in addressing the issues associated with Big Data.

Methodology: The materials used are definitions of Big Data found in literature, standards, and technologies used in the Semantic Web and Linked Open Data, as well as the use case of a research dataset; we use the conceptual bases of semiotics and ontology to analyze the role of vocabularies in knowledge organization (KO) in assigning subjects to documents as a special, limited, use case that may be expanded within such context.

Results: We develop and expand the conception of data as an artificial, intentional construction that represents a property of an entity within a specific domain and serving as the essential component of the Big Data. We present a comprehensive conceptualization of semantic expressivity and use it to classify the different vocabularies. We suggest and specify features to vocabularies that may be used within the context of the Semantic Web and the Linked Open Data to assign machine-processable semantics to Big Data. We identify computational ontologies as a type of knowledge organization system with a higher degree of semantic expressivity. It is suggested that such themes should be incorporated into professional qualifications in KO.

> The ultimate Big Data challenge lies not in the data, but in the metadata—the machine-readable descriptions that provide data about the data. It is not enough to simply put data online; data are not usable until they can be 'explained' in a manner that both humans and computers can process." Researcher Mark Musen Declaration (FAIR Compliant Biomedical Metadata Templates | CEDAR, 2019).

1. Introduction

How do we discover, access, process, and reuse the huge and growing amount of digital data that are continuously made available by our society, so-called Big Data, a significant part of which is constituted by research data? Research data is an important product of science, along with scientific publications. How can we enable its large-scale reuse? In light of Big Data and the statement by researcher Mark Musen, how can knowledge organization (KO) contribute?

1.1. The Big Data

Big Data, the term for a recent phenomenon describing the amount of data produced in digital format, its explosive growth and the difficulties of storing, processing, and reusing the data, is increasingly present in information technology media. The headlines also call the phenomenon "information deluge," "data deluge," or "tsunami of data" (Hey and Trefethen, 2003). According to these sources, it is impacting business, government, culture, science, and society.

Big Data reminds us from the so-called "information explosion," a fundamental phenomenon in the area. In response, KO created knowledge organization systems (KOS) as auxiliary systems to information retrieval systems (IRS), which are traditionally computerized databases containing representations of scientific documents that control or standardize the natural language used both for indexing the documents entered in the IRS and to standardize natural language keywords in the queries formulated by users in an "information retrieval thesaurus" (Dextre Clarke, 2016, 138).

Although Big Data has been sparking interest in KO, contributions from the area to contextualize it or to propose practical solutions are still few. The conceptualizations tend to

repeat those originating in computer science, emphasizing aspects technological aspects as volume, variety, velocity, heterogeneity and the need of massive computer power to process it.

The best-known product of science, to which the KO has been dedicated since its beginnings, are scientific publications. More recently, science has been giving increasing importance to another of its products, research data. Today, research data, practically entirely digital, is produced in increasing quantities as a result of scientific activity carried out with the support of information technologies. Examples of this huge amount of digital survey data are those generated by the Hubble Space Telescope, https://www.nasa.gov/mission_pages/hubble/main/index.html, the Human Genome research project, https://www.genome.gov/human-genome-project, or the Large Hadron Collider, https://home.cern/science/accelerators/large-hadron-collider, the largest and most powerful particle accelerator in the world. This digital research data is part of the Big Data phenomenon.

As quoted by researcher Mark Musen at the beginning of this work: we cannot address Big Data without using computers to help us. This observation refers to the Semantic Web project (Berners_Lee et al 2001), the proposal for a Web whose resources would be represented in a way that had a precise and formal meaning or semantics and would be intelligible and understandable by both people and machines.

In previous works, we have already discussed how to link digital representations of objects of memory and culture through the Web (Marcondes, 2020), and how one of the products of science, scientific publications, could be intelligible and understandable by both people and machines (Marcondes and Costa, 2016), when represented with the technologies

of Linked Open Data (LOD) and the Semantic Web. Here we are interested in doing the same for science's other great product, digital research data.

1.2. Traditional use of vocabularies to assign subjects to documents

Representing documents and their subjects has been foundational to the practices developed by KO, especially when, unlike today, there was no access to full-text documents in digital format and the descriptive and thematic representation of the documents was a fundamental mechanism in the intermediation, and relevance assessment processes carried out in the retreival of information (Saracevic, 2007). KO methodologies have always represented domains of knowledge when building KOS like controlled/standardized vocabularies, subject headings, and classification schemas. The early KOS, such as thesauri, were intended to enable subject-based retrieval in the context of IRS because their records were representations of objects that had subjects as one of their properties. But not all objects in a domain have subjects as one of their properties, like documents. We see today that this is just a case of representing objects in digital space.

Today, it is not only about retrieving documents (or their representations) but also to create digital representations of anything, such as in the "Internet of Things" (IoT), If the documentation movement (Otlet, 2018) and then information science intended the empowerment of "information" by separating it from books, the Semantic Web proposes to also, in a certain sense, empower "knowledge," which is no longer just inserted into texts to be interpreted by humans, but rather recorded directly in Resource Description Framework (RDF) triples (RDF 1.1 PRIMER, 2014), forming representations/descriptions of "things." The Web thus becomes a large knowledge base that can be consulted about the "things" thus represented (SPARQL 1.1 QUERY LANGUAGE, 2013).

The objective of this work is to discuss how KO can contribute to assigning computational semantics to Big Data, especially to research data, so that computers can process them, allowing their reuse on a large scale. As a methodology, the work discusses the conceptualizations of data and (the few of) Big Data originating in KO in an attempt to make it clearer what would be data, essential elements of the Big Data phenomenon, and in particular, digital research data. It then proceeds to analyze digital research data based on the Case Report Form (CRF), WHO-COVID-CRF/WHO-2019-nCoV-Clinical_CRF-2020.3-eng.pdf at master · FAIRDataTeam/WHO-COVID-CRF · GitHub, proposed by the World Health Organization (WHO) to standardize and unify the registration of cases of patients with COVID-19 worldwide.

The work is organized as follows. After this introduction, section 2 analyzes data definitions, their traditional use in KO, and develops a conceptualization of data that is illustrated by an example of research datasets, relating them to the representation of things in a domain and organized into vocabularies. Section 3 presents a comprehensive view of vocabularies based on Semantic Web and LOD technologies and discusses which functionalities vocabularies must incorporate to integrate with these technologies. Section 4 raises research questions to be developed and presents final considerations.

2. A Semiotic and Ontological view of data

None of the most common Big Data definitions exclude the data component. It seems reasonable, then, that understanding what Big Data is and how to operationalize solutions to the problem begins by elucidating what data is. This section proposes a semiotic and ontological analysis of data, understood as the essential component of Big Data. This analysis begins with the question of elucidating how to assign semantics to the data. Then we discuss what data is, from an ontological point of view. From the elucidation of these

questions, concepts of research data, data concerning domains of human action, and vocabularies as representations of domains, are developed.

2.1. Data as Representations

What is Big Data? What is its relationship with data? What is data and how is it related to metadata? How should semantics be assigned to data? The ISO/IEC 20546/2019 Standard notes, "The big data paradigm is a rapidly changing field with rapidly changing technologies," later suggesting a definition: "extensive datasets (3.1.11) — primarily in the data (3.1.5) characteristics of volume, variety, velocity, and/or variability — that require a scalable technology for efficient storage, manipulation, management, and analysis."

The conceptualizations of Big Data originating from KO are few (Marcondes et al 2021) and replicate those originating in computer science, defining it as a phenomenon that involves large amounts of data, the heterogeneity of that data, a continuous flow of generation and updating, and a need for large processing capacity, so that the data reveal patterns or trends (De Mauro et al 2015). However, the same is not true for the conceptualizations of data originating from KO. Data is mentioned frequently in the literature, along with its relationships with information and knowledge (Buckland, 1991), often called the data, information, knowledge, wisdom (DIKW) hierarchy (Rowley, 2007). In Floridi (2019), information is related to data and semantics.

An important exception is from Hjørland (2018), who proposes a conceptualization of Big Data arising from definitions of data, a phenomenon much better known and conceptualized in the area. Data is the essence of the Big Data phenomenon, it could not exist without data. In this work, Hjørland lists several similar conceptualizations of data and highlights that of Fox and Levitin:

> Within this framework, we define a datum or data item, as a triple <e, a, v>, where e is an entity in a conceptual model, a is an attribute of entity e, and v is a value from the domain of attribute a. A datum asserts that entity and has value v for attribute a. Data are the members of any collection of data items.

This conceptualization is clarified by the following example: "2018." What does 2018 mean? Others would say it's a given. Let us note, however, this statement: "Giovana was born in 2018." In it we can identify the entity we are talking about: a child called "Giovana," an attribute or property of this entity, she is "born," and the value of this attribute or property, her year of birth, "2018."

In the ontological scheme that goes back to Aristotle (2000), reality is constituted of the first substances, the things that have real existence in space and time, and second substances, the conceptualizations we make of the first substances to think, reason, make sense of, and communicate about the things in reality. Second substances are in turn subdivided into essences, concepts that designate things that have existential independence, and accidents, concepts that designate things that are existentially dependent on other substances. Things that have existential independence are commonly recognized in one of the most well-known ontological schemes, the entity-relationships (ER) model (Chen, 1976) as entities, while those that are existentially dependent, as properties. Properties, in turn, are subdivided into attributes of an entity, relationships between an existentially independent entity and the value of one of its properties, and relationships, involving two or more individuals of the same existentially independent entity, or of more than one existentially independent entity (Orilia and Paoletti, 2020).

We are talking about representations. A piece of data, even in the context of Big Data, then, makes no sense without referencing the entity and one of its properties, the

metadata. The three concepts are inseparable and cannot be understood separately. They correspond to a descriptive, representational element of an entity, describing one of its properties. They correspond linguistically to a claim, a basic unit of knowledge to which, according to Aristotle (2000, p. 39), values of truth or falsity can be attributed.

The statements represented by triples constituted by an entity, one of its properties, and the value of this property correspond to the representation of informational resources in the context of LOD, using the RDF (RDF Primer, 2014). RDF is a Semantic Web standard for describing resources. Everything that is available on the Web can be accessed through a link, or a Uniform Resource Identifier (URI).[i] This representational model describes such a resource through triples formed by subject, the resource being described; predicate, a property that describes the resource; and object, the value of this property for this resource. The RDF model assumes a minimum semantics, that is, the subject, the predicate, and the object that form the triple are identified and appear in this order.

2.2. Data and Big Data: the case for research data

Next, we will attempt to demonstrate how the conceptualization above helps address the issues of Big Data, especially research data. A concrete and dramatic example of the importance of research data and the adoption of principles and technologies that allow its wide dissemination and reuse is the form for collecting data from patients infected with COVID-19, the CRF, which was proposed by the WHO. The GO FAIR initiative, https://www.go-fair.org/, proposes the creation of a worldwide network of catalogs that can reference research data collected through the CRF and deposited in repositories and that are available according to the FAIR principles, the "FAIR Data Points." Brazil participates in this initiative through the VODAN-Br Virus Outbreak Data Network initiative (Veiga et al

2021). The form fields must be filled with metadata and data associated with vocabularies to allow their standardization, without which their processing by computers would not be possible, and consequently neither would the ability to extract conclusions and insights.

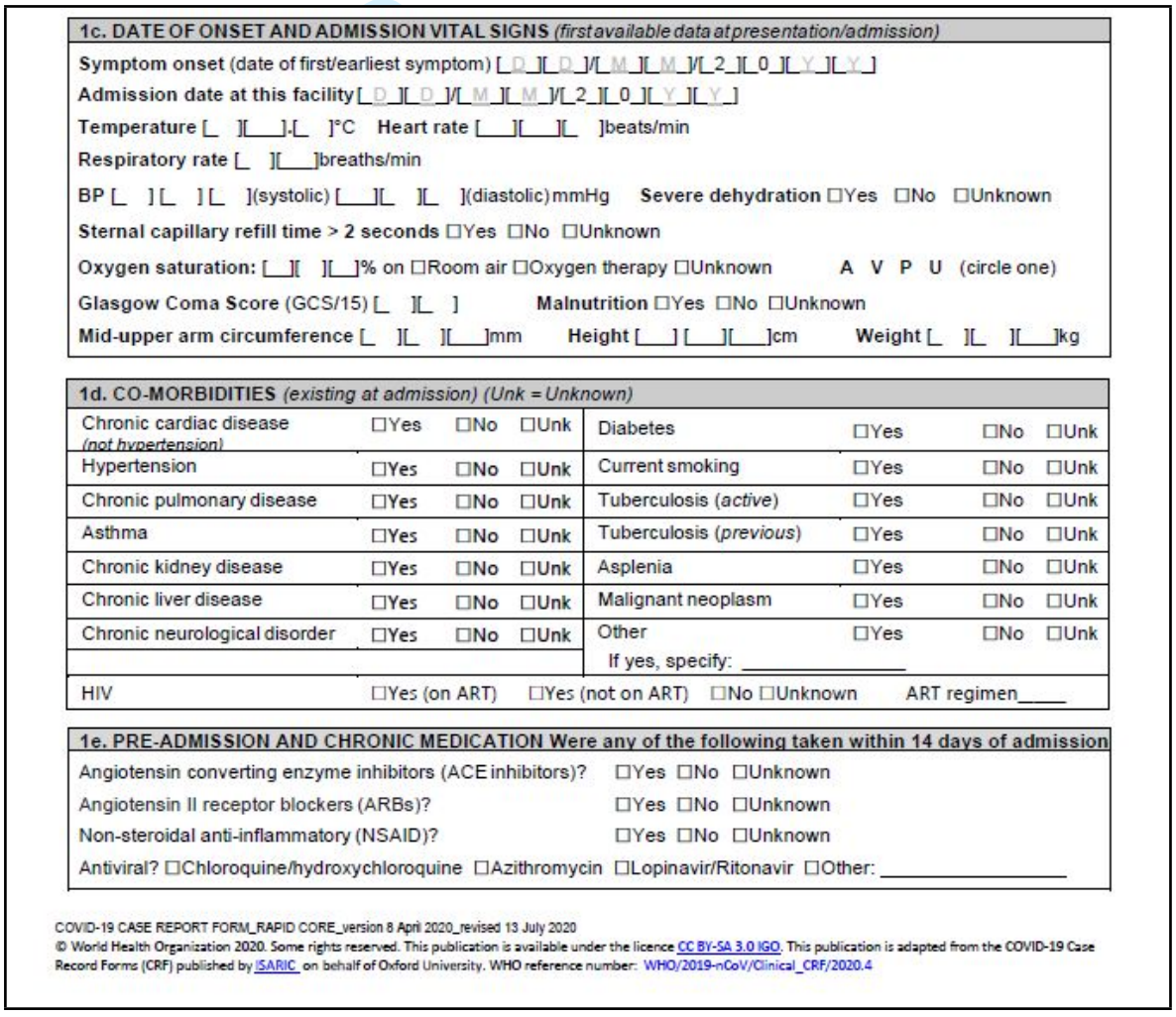
In the RDF model, instead of the subject, predicate, and object of a triple being represented in natural language, which is ambiguous and difficult for programs to process, each can be identified by a URI. These URI identify specific terms, both from metadata vocabularies—descriptive properties of things in a domain—and from data vocabularies—values assumed by these properties for specific instances. This unified characterization as vocabularies, that is, sets of systematized terms that identify both the descriptive properties (metadata) of objects in a domain, as well as the data, as the values assumed by these properties for instance, is due to Marcia Zeng (2019).

Another important feature of using vocabularies with LOD technologies is that different vocabularies can be used simultaneously in the form fields. In Figure 1 we see an excerpt from the CRF. As co-morbidity data, "CO-MORBIDITIES," of a patient (the entity) are recorded, concepts such as chronic cardiac disease (the attribute or metadata) are taken from specific biomedical ontologies or vocabularies: Yes, No, Unk (the value or data). These data have to be processed by programs so that the immense amount of records collected through the CRF around the world can serve as inputs for the planning and control of the pandemic. The question about co-morbidities has several answer options, each of which indicates a type of disease. For it to be processed by machines, each type of co-morbidity expressed in natural language must reference a concept in a vocabulary or ontology, such as SNOMED-CT, https://www.nlm.nih.gov/healthit/snomedct/index.html, for example. Another question on the CRF, such as the one related to "PRE-ADMISSION AND CHRONIC MEDICATION," has as one of its answer options "Angiotensin converting enzyme inhibitors (ACE inhibitors)?", which may be referenced in another

vocabulary, such as MeSH, https://meshb.nlm.nih.gov/search, for example, the identifier: http://id.nlm.nih.gov/mesh/D000806.

In order to have precise meaning, concepts such as those shown in the CRF must refer to specific, standardized ontologies or biomedical vocabularies to enable the processing of these data.

FIGURE 1 - Part of the CRF Form



1c. DATE OF ONSET AND ADMISSION VITAL SIGNS *(first available data at presentation/admission)*

Symptom onset (date of first/earliest symptom) [_D_][_D_]/[_M_][_M_]/[_2_][_0_][_Y_][_Y_]
Admission date at this facility [_D_][_D_]/[_M_][_M_]/[_2_][_0_][_Y_][_Y_]
Temperature [_][___].[_]°C **Heart rate** [___][___][_]beats/min
Respiratory rate [_][___]breaths/min
BP [_] [_] [_](systolic) [___][_][_](diastolic) mmHg **Severe dehydration** ☐Yes ☐No ☐Unknown
Sternal capillary refill time > 2 seconds ☐Yes ☐No ☐Unknown
Oxygen saturation: [__][ ][__]% on ☐Room air ☐Oxygen therapy ☐Unknown **A V P U** (circle one)
Glasgow Coma Score (GCS/15) [_][_] **Malnutrition** ☐Yes ☐No ☐Unknown
Mid-upper arm circumference [_][_][___]mm **Height** [___] [___][___]cm **Weight** [_][_][___]kg

1d. CO-MORBIDITIES *(existing at admission) (Unk = Unknown)*

Chronic cardiac disease *(not hypertension)*	☐Yes	☐No	☐Unk	Diabetes	☐Yes	☐No	☐Unk
Hypertension	☐Yes	☐No	☐Unk	Current smoking	☐Yes	☐No	☐Unk
Chronic pulmonary disease	☐Yes	☐No	☐Unk	Tuberculosis (*active*)	☐Yes	☐No	☐Unk
Asthma	☐Yes	☐No	☐Unk	Tuberculosis (*previous*)	☐Yes	☐No	☐Unk
Chronic kidney disease	☐Yes	☐No	☐Unk	Asplenia	☐Yes	☐No	☐Unk
Chronic liver disease	☐Yes	☐No	☐Unk	Malignant neoplasm	☐Yes	☐No	☐Unk
Chronic neurological disorder	☐Yes	☐No	☐Unk	Other	☐Yes	☐No	☐Unk
				If yes, specify: _____			
HIV	☐Yes (on ART)		☐Yes (not on ART)	☐No ☐Unknown	ART regimen_____		

1e. PRE-ADMISSION AND CHRONIC MEDICATION Were any of the following taken within 14 days of admission

Angiotensin converting enzyme inhibitors (ACE inhibitors)? ☐Yes ☐No ☐Unknown
Angiotensin II receptor blockers (ARBs)? ☐Yes ☐No ☐Unknown
Non-steroidal anti-inflammatory (NSAID)? ☐Yes ☐No ☐Unknown
Antiviral? ☐Chloroquine/hydroxychloroquine ☐Azithromycin ☐Lopinavir/Ritonavir ☐Other: _____

COVID-19 CASE REPORT FORM_RAPID CORE_version 8 April 2020_revised 13 July 2020

For data to be considered open, international recommendations rate it from 1 to 5 stars, https://5stardata.info/en/. The fourth and fifth stars are awarded when data is available

in RDF format, include predicates and objects referring to standardized vocabularies, is widely recognized by the community in a given domain, and linked together to provide rich context. For research data, which has demanded increasing attention and public policies at national and international levels, the international GO FAIR initiative recommends a set of principles for publication so that they have the attributes of FAIR: findability, accessibility, interoperability, and reuse. For research data to achieve these attributes, they must be accessible through a URI, represented in RDF, constituting the Linked Open Vocabularies (LOV).

The idea behind the FAIR principles is to allow research data to be processed by machines. The M4M principle—metadata for machines—states that "There is no FAIR data without machine-actionable metadata. The overall goal of Metadata for Machines workshops (M4M) is to make routine use of machine-actionable metadata in a broad range of fields." An example of the importance of research data and the adoption of principles that allow its wide dissemination and reuse is the CRF form described above. Without standardization, its processing by machines would be impossible.

2.3. Traditional use of vocabularies to assign subjects to documents, generalized use of vocabularies as representations of a domain

Since the onset of the information explosion, thesauri have emerged, complementary systems to the IRS, of which the KOS were one of its components. The development of the IRS drew on sources from other traditions of librarianship, documentation and cataloging. Catalog sheets and bibliographic entries served as models for computational bibliographic formats in projects such as Machine Readable Cataloging (MARC), based on the AACR2 cataloging standard, and the UNISIST Reference Manual (Dierickx, Hopkinson, 1986), based on the ISBD standard. These formats served as a model for library catalog databases

and for indexing and summary services. Significantly, bibliographic formats evolved separately from the also nascent technology of computer databases that, from the 1970s onwards, had the relational model as a paradigm (Codd, 1970). The Text Retrieval Conferences (TREC) conference series illustrates this separate evolution.

Concerning thematic representation, a whole theoretical and methodological foundation were all developed to support the development of KOS, from classification theories, the Faceted Classification Theory (Ranganathan, Gopinath, 1967), the proposals of the Classification Research Group (CRG) (Wilson, 1972), Concept Theory (Dahlberg, 1978), to Terminology (Cabré, 2005). This theoretical and methodological tradition , an area of excellence of KO, meets, with the emergence of the Semantic Web, in subdisciplines such as systems modeling, artificial intelligence, and computational ontologies, áreas originating from computer science. These are understood as one of the foundations of the proposed Semantic Web. Many of these new KOS are developed by computer professionals and scientists from different areas or specialists: biomedicine, statistics, or from curators of digital collections in memory and culture, etc. The words of Hjørland (2008, p. 86) highlight and warn about this approach to other areas: "(LIS) is the central discipline of KO in this narrow sense (although seriously challenged by, among other fields, computer science)." Will KO limit itself to developing traditional KOS and leave this space to computer science specialists as Hjørland warns?

The technical traditions and standards developed by KO to manage the information explosion resulted in the establishment of IRS/KOS assumptions that persist to this day. In most discourse in the area, these assumptions are so implicit that it becomes difficult to make them explicit, consider them, and analyze their consequences. All the theories and methodologies of KO mentioned bring these assumptions implicitly: the IRS represent documents in their computerized databases, MARC and the bibliographic formats that

emerged from the UNISIST Reference Manual are sets of metadata that represent different (descriptive) properties of the documents, while the KOS associated with them are terminological standardization instruments specifically for the subject property, the subject field of the records of the IRS computerized databases. These records represent objects that have, among others, the subject property. They are symbolic objects, documents. It is worth adding that the records themselves, the metadata set, are also symbolic objects, representing document-type objects.

These implicit assumptions account for the division that occurs in the teaching and practice of librarianship and KO between descriptive representation and thematic representation, or of subjects, of a document. To what extent do these assumptions hold up today, and are they sufficient to address the challenges of the Semantic Web era, Big Data, and the Internet of Things?

In the 1980s-1990s, as a consequence of the emergence of online bibliographic catalog management systems and databases, the domain of information retrieval in library catalogues, so familiar to us, but also so exclusive, with its diversity of objects, was modeled using a methodology used in computer science to plan database management systems. The conceptual model Functional Requirements for Bibliographic Records (FRBR) appeared in 1998, whose development was promoted by IFLA (1998).

According to Mylopoulus (1992, p. 3) "Conceptual modeling is the activity of formally describing some aspects of the physical and social world around us for purposes of understanding and communication." For Mylopoulus,

> the descriptions that arise from conceptual modeling activities are intended to be used by humans, not machines. . . [and] The adequacy of a conceptual modeling notation rests on its contribution to the construction of models of

reality that promote a common understanding of that reality among their human users.

A conceptual model sets an agreement between users of a system on what kinds of things exist and will be represented in the system, or entities (also called classes) in a given domain of reality, e.g. documents of historical value, the properties of these entities and how they relate (relationships) to each other. Thus, a conceptual model is a representation, in the form of an abstract and generic description, independent of computational implementations (hardware, operating systems, languages, database management systems) of a given domain of reality. To understand this reality, reason about it and establish a common understanding of this reality, a conceptual model answers questions such as: What different things exist in a given domain? How are they distinguished from each other? How do they relate? What are your properties?

A conceptual model as a representation is expressed, communicated, externalized through a language, or more specifically a meta-language or meta-model (Guizzardi, 2007, 23), which is a language to express the languages that express things in specific domains. Examples of these meta-languages are either natural language (through a system requirements document), which functions as the most general of all meta-languages, or a diagrammatic meta-language, such as entity-relationship (meta) Model. The Unified Modeling Language (UML), https://www.uml.org/, class diagram, in which domain-specific ER models or class diagrams are expressed, both an ER model and a class diagram can define a language that designates things in a domain or a specific vocabulary to that domain.

In the descriptive representation, once established and consolidated practical standards of representation such as MARC, UNISIST, AACR2, and ISDB, the KO started to question what things were implicit within them, their conceptual models.

Conceptual models in the area of documentation and information have made these things explicit. They evolved the standards mentioned above for creating automated bibliographic records, starting with the pioneering FRBR (Ifla, 1998). They are cases of representations of a domain, not for indexing documents, but for formalizing, identifying, consensing, and standardizing objects, actors, and processes and their relationships within a domain.

Modeling in documentation and information has its roots in bibliographic classification systems such as the Dewey Decimal Classification (DCC) – and the Universal Decimal Classification (UDC). The DCC and UDC can be viewed as a set of taxonomies, each having as a root a discipline into which the universe of knowledge was classified. The use of taxonomies to organize a domain is typically used today for information management within corporations and to organize the content of websites (Lambe, 2014). Taxonomies only organize the things in a domain in class-subclass relationships. The things being organized in a bibliographic classification are disciplines to use the terms that identify them as a subject to a book.

However, there are more than just things in a domain. A more accurate model of a domain should include not only the things within it but also their relationships and attributes. Things have properties, attributes, and relationships, according to the ER model. The first movement within documentation and information to recognize this was Faceted Classification (Ranganathan, Gopinath, 1967). Facets are the properties of a class of things of interest for information recovery (Giunchiglia et al 2014; Marcondes and Dias, 2020). Besides things, conceptual models embody also properties of things, their attributes, and relationships; recognizing this results in a more accurate representation of a domain. After

the pioneering FRBR model (Ifla, 1998), the Intenational Council of Museums (ICOM) adopted the CIDOC Conceptual Reference Model (CIDOC, 2014), and more recently the International Council of Archives (ICA) adopted the Records in Context Conceptual Model (Ric-CM) (International Council on Archives, 2019).

Conceptual models, when designating things in a domain, define a vocabulary for metadata. They are aligned within different types of KOS (Almeida et al 2011, 196), ordered according to their semantic expressiveness. Semantic expressiveness can be understood, in the context of the previous quote, as the ability of each type of KOS to distinguish and describe, that is, identify the properties and represent the different things that exist in a domain of that reality.

Vocabularies of most types are semantic control devices, formed by systematized sets of semiotic, triadic entities (PEIRCE, 1994), concepts (Dextre Clarke and Zeng, 2012), units of meaning that relate something (a first: object or referents), in some way (through a second: term or code), which generates or induces a third: its meaning.

Vocabularies are constructed to answer the basic ontological question: what exists in a domain? They are representations or models of a domain of reality, taking of what things there are, what their attributes are, their relationships, and how to express them linguistically, the concepts (Dahlberg, 1978), and their units of meaning. Online Public Access Catalogs (OPACs) or databases are used as terminological control tools with the IRS used by institutions, with the function of standardizing the terms used for the description and indexing of scientific publications, memory and cultural objects, and other items included in these systems.

2.4. Domains

Aside from the general library classification systems such as the CDD and the CDU, KOS are developed and used concerning a specific domain. The domain notion commonly used in KO is that of a specialized knowledge area.

Hjørland and Albrechtsen (1995, 400), in the text in which they propose the analysis of domains as the foundation of KO, define domains as: "thought or discourse communities, which are parts of society's division of labor." They also label a domain as a "specialty/discipline/domain/environment" (Hjørland and Albrechtsen, 1995, 401).

Hjørland (2002, 422) conceptualizes domains associated with specialized libraries, questioning what knowledge would be necessary for information professionals to work in "in a specific subject field like medicine, sociology or music?" In Hjørland and Hartel (2003, 239), this view of domains as systems of thought, theories, is reaffirmed.

> Domains are basically of three kinds of theories and concepts: (1) ontological theories and concepts about the objects of human activity; (2) epistemological theories and concepts about knowledge and the ways to obtain knowledge, implying methodological principles about the ways objects are investigated; and (3) sociological concepts about the groups of people concerned with the objects.

The KOS of the early years of KO, such as thesauri, were intended to enable subject-based retrieval in the context of IRS because their records were representations of objects that had subjects as one of their properties, that is, documents. Today, it is not just about retrieving documents (or their representations) but digital representations of anything, as exemplified in the IoT. These representations are no longer just access points for documents, but also information resources themselves, complex descriptions of these objects, sources of knowledge about them, represented in such a way that they can be processed/intelligible by

both machines and humans. Such representations allow machines to make inferences about the knowledge thus represented.

KO today is being called upon to model different domains of knowledge to build new semantic vocabularies. For this, it is necessary to expand the traditional notion of a domain as a discipline or subject. In the area of software development, the notion of a domain has a broader scope: it is 'a sphere of activity or interest: field" [Webster]. In the context of software engineering, it is most often understood as an application area, a field for which software systems are developed (Prieto Diáz, 1990, 50).

If we consider that a KOS is a terminological system that represents the "things" of interest in a domain of action for the community of agents/users of that domain, to create a KOS (an artifact, similar to software) several aspects must be considered. We must first determine what things exist in a domain and which are relevant to this community, what rules exist about these things or are created/approved/agreed on about these things, how this community uses them to act in this domain and, finally, how the conceptualizations (Dahlberg, 1978), generating as one of the by-products of this process as a set of terms, are to be systematized, for example, in a thesaurus.

What things are in a domain? How should they be represented? These are the questions of ontology and semiotics. They must be answered to create a representation, or a conceptual model, of a domain.

2.5. Vocabularies as representations of a domain.

As shown, a vocabulary is a representation of a domain, regardless of its use, either to assign subjects to documents: a) vocabulary for indexing, which identifies what things exist in a domain (e.g. MeSH categories describing the entities within the Healthcare

domain, https://meshb.nlm.nih.gov/treeView), or to b) vocabulary to describe objects in this domain, descriptive metadata standards that, in addition to identifying what things exist in a domain, also describe and list their properties. Among the things within a domain, there are vocabularies of specific facets for special purposes: archival science and records management uses functional classification plans in an organization to assign the organizational provenance or the function that generated or used a record.

3. A Comprehensive view of vocabularies

In this section we compiled and developed a comprehensive view of vocabularies based on previous discussion in section 2 and on contributions by Hjørland (2018) and Zeng (2019).

3.1. Vocabularies, Semantic Web, Linked Open Data, and Big Data

LOD technologies are an integral part of the Semantic Web project. Although this is its best-known name, the project is also known as Web of Data, a name that describes it better, since semantics concerns meanings (Chierchia, 2003), and the ability of the Web of Data to convey meanings is quite limited and different from the sense in our understanding of expressions in natural language.

The project was initially formulated by computer scientist Tim Berners-Lee, the creator of the Web, among others. According to its formulators, the Semantic Web aims to propose "A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities" (Berners-Lee et al 2001). To its authors, "Most of the Web's content today is designed for humans to read, not for computer programs to manipulate meaningfully." The Semantic Web then "will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users."

The Web of Data then refers to content represented in such a way that it can be understood by both machines and people. The current Web is made up of pages, such as http://www.uff.br, formatted in Hypertext Markup Language (HTML), accessible and interconnected with each other through links. Navigating these pages through these links is done by browsers, such as Internet Explorer, Google Chrome, or Mozilla Firefox. HTML is a content markup language; it formats the content through a pre-defined set of markups, which instruct browsers to display them on computer screens for human users. The content of HTML pages is interpreted by browsers to make it readable and visually pleasing to people.

The proposed Web of Data is quite different. The Web will no longer be constituted of pages to be read by people, but of content, called informational resources, digital representations of things: concrete, like me, you, an industrial product, a monument, a geographical accident; abstract, like a musical genre, a scientific discipline; or just has a digital existence, such as a photo in a JPG file or a scientific article in a PDF file. These are the entities in the proposal by Hjørland (2018). Each of these resources is uniquely identified by a link, or a URI. A resource, identified/accessed by its URI, is described in a structured way through triples, each one formed by the URI of the resource, by each of its properties, and by the corresponding values of each of these properties. An example of how this representational model works is the Leonardo Da Vinci resource on Wikidata, https://www.wikidata.org/wiki/Q762.

This model of structuring data through the description of resources formed by one or more linguistic claims made up of triples <Subject> <Predicate> <Object> is RDF (RDF Primer, 2004). From an ontological point of view, subject, predicate, and object can be understood as an entity, a property, and the value of this property.

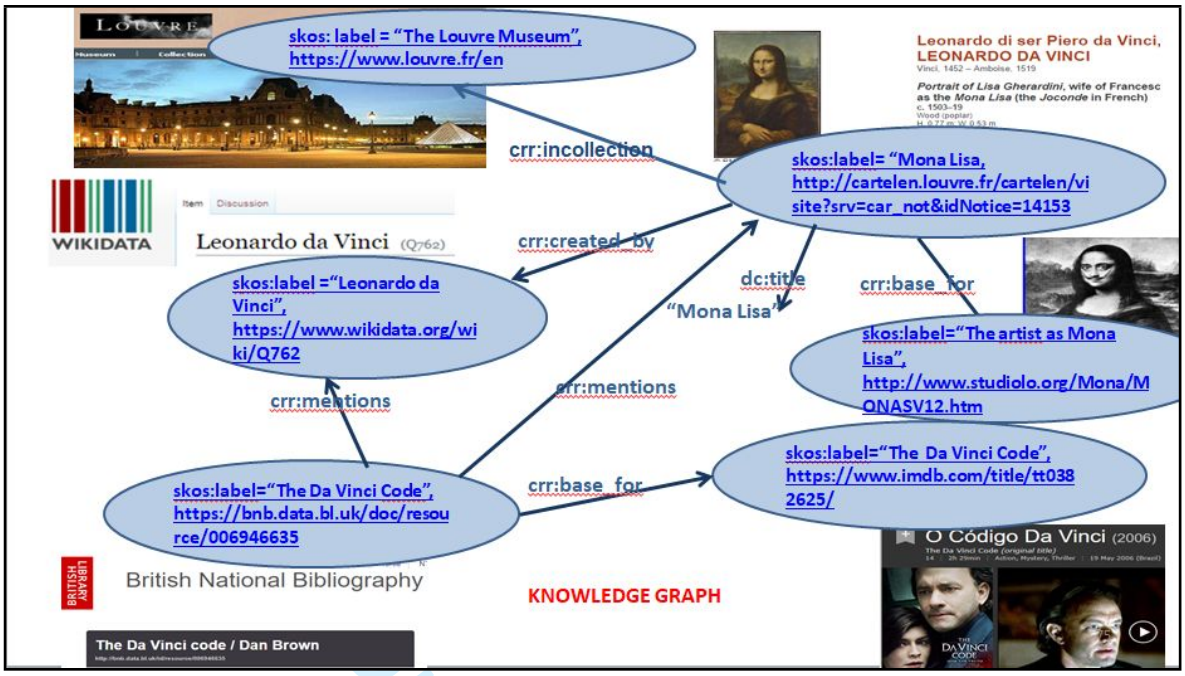
Looking in more detail at structuring a triple; for example,

"The page http://www.uff.br is authored by _____."

We have then a claim that consists of three elements: the subject, "http://www.uff.br," the predicate, "has as author" and the object, "_____"

The RDF model presupposes a minimum semantics, derived from its corresponding linguistic claim. That is, they are identified and appear in this order: the subject, the predicate and the object of the claim that form the triple (Resource Description Framework (RDF) Model and Syntax Specification, 1998). A triple describes a specific piece of data from the resource description (what Hjørland calls a "datum:" a unit of data). Sets of triples with the same subject describe the same resource. Sets of linked interlinked triples describing a resource form a graph. Next, we see the graphical representation of an RDF graph.

FIGURE 2. Graphical representation of an RDF graph



Every computerized system rests on the data processing model. A computer system has as basic components data, processed by programs. While the RDF model describes data, the counterpart in terms of processing are programs that perform inferences on RDF graphs. The minimal semantics of the RDF model allows these programs to navigate through the graphs formed by the triples and infer one or two of the subject(s), predicate(s), or object(s) when they are unknown, such as:

- Who is the author of the page http://www.uff.br?

  - < http://www.uff.br > < authored > < ???>.

- What role does _____ have in relation to the page http://www.uff.br?

  - < http://www.uff.br > < ??? > < _____>.

- What are all the claims about the page http://www.uff.br?

- < http://www.uff.br > < ??? > < ??? >.

(SPARQL 1.1 QUERY LANGUAGE, 2013) is the query language that allows users to query sets of RDF triples, navigating through the graphs formed by them and performing inferences. It is the materialization of the Web of Data proposal for a Web that can be consulted as if it were a database.

RDF can be represented ("serialized" in computing technical language) in several formats, such as RDF/XML, N Triples, JSON, or TURTLE (RDF Primer, 2004). Of course, RDF triples represented in these formats are not as human-friendly or as clearly readable as HTML pages when viewed by browsers. But they contain elements that allow browsers to understand these formats and display them in a human-friendly manner, if applicable. The main objective of the resources described in RDF is that they can be processed by machines (including their user-friend visualization), thus helping to organize, retrieve, and make these resources accessible.

Naturally, given a triple like < http://www.uff.br > < is authored > < _____ >, a machine cannot do much more than identify the subject, the predicate or the object of the triple. In this example, predicate and object are names, strings of characters understandable only by people, holders of a set of contextual and cultural information, accumulated throughout their life histories. RDF Semantics is limited to its model of structuring triples as subject, predicate, object.

The way to extend these semantics beyond the limits of the RDF model is also to make predicates and/or objects into URI and that these URI refer to concepts of vocabularies with specific semantics. According to RDF Semantics (2004) "There are several aspects of meaning in RDF which are ignored by these semantics; in particular, it treats URI references

as simple names, ignoring aspects of meaning encoded in particular URI forms." A URI in the RDF model is just a name, an identifier. The advantage of a URI over a natural language identifier as the linguistic term "author," is its uniqueness (other properties can be identified as the natural language term "author," synonyms as creator, for example), its validity, as a URI, throughout the whole webspace, and its persistence, that is, the commitment of whoever assigns a URI to never change it (Berners-Lee, 1998).

Extending the previous example, we have:

<http://www.uff.br> <http://purl.org/dc/elements/1.1/creator> <https://orcid.org/0000-0003-0929-8475>

In this example, the original predicate "author" is replaced by the URI referenced by the "creator" element of the well-known Dublin Core (DC) metadata standard. In its context, dc:creator has a specific semantics. It is defined as "An entity responsible for making the resource." The triple's object, the value or content of dc:creator, has been replaced by the Open Researcher and Contributor ID (ORCID), https://orcid.org, of the page's author.

It is with the semantics in specific vocabularies that the limited semantic expressiveness of the RDF model can be expanded, as seen in the example of the CRF. Once specified in elements of a vocabulary, the semantics can be processed by programs. While the features provided in the Web of Data, represented in markup languages such as XML, RDF, HTML, etc. are contents, programs are procedures or algorithms according to the data processing model. Programs only know how to process content. For this, they need to be clearly instructed (programmed) on what to do with certain content in a certain situation. LOV used to assign semantics to LOD (Zeng, 2019) must clearly define, restrict, and specify the semantics of their concepts. For example, the DC metadata vocabulary clearly defines

the semantics of each of its concepts (called elements in the DC initiative), dc:creator, such as the creator/author or responsible for a resource, e.g., a digital scientific paper. Furthermore, the dc:creator element has itself, a unique persistent identifier, a link, a URI: http://purl.org/dc/elements/1.1/creator. This persistent identifier, unique throughout the Webspace, works as a guarantee of the semantics, allowing a developer to create a specific program to process this element of the DC vocabulary unambiguously, from the semantics specified and standardized in the DC vocabulary, specifically in the dc:creator element.

Here is another example of what was just explained. Let the following RDF triples be:

TABLE 1. Two triples with the same precicates

<libro0237>	<title>	<Don Quixote>.
<http://catalogo.bne.es/libro0237>	<http://purl.org/dc/elements/1.1/title>	<Don Quixote>. And
<emp0027>	<title>	<President>.
<http://www.company.com/0027>	<http://www.w3c.org/2006/vcard/ns/title>	<President> .

The predicates of both triples are apparently identical as “title.” They only differ by the “link” to the vocabulary. In the first example, it is http://purl.org/dc/elements/1.1/title, and in the second it is http://www.w3c.org/2006/vcard/ns/title. These links to different vocabularies, also called namespaces—a kind of delimitation of a scope where those identifiers, with those specific meanings, are valid—allow programs that process the triples to uniquely identify the different concepts in the different vocabularies that serve as predicates for the two triples and even process the two triples simultaneously without confusing their semantics. It is because they are not restricted to the eventual informal

meaning of "title" but to this meaning within the scope ("namespaces") of the DC and Vcard, https://devguide.calconnect.org/vCard/vcard-4/, vocabularies.

This allows programs to do more than just process inferences about a graph, a set of RDF triples. These are programs oriented by ontology or models, such as Application Program Interfaces (APIs) from the Europeana Library, https://pro.europeana.eu/page/apis.

3.2. Functionalities for vocabularies to be used to assign semantics to data within the context of the Semantic Web and LOD

Through unique and persistent identifiers, metadata and data vocabularies can be used to assign machine-understandable semantics to predicates and objects in triples RDF. Many old vocabularies are being restructured to be compatible with LOD technologies (Soergel, 2004; Dos Santos Maculan, 2015), such as the UNESCO Thesaurus, http://vocabularies.unesco.org/browser/thesaurus/en/, the FAO Thesaurus, http://aims.fao.org/aos/agrovoc/c_8003.html, the AGROVOC Thesaurus, https://agrovoc.fao.org/browse/agrovoc/en/, the Paul Getty Foundation Vocabularies, https://www.getty.edu/research/tools/vocabularies/lod/, the Art and Architecture Thesaurus, the Union List of Artists Names, the Cultural Objects Name Authority, the Getty Thesaurus of Geographic Names, the DeCS/MeSH, Health Science Descriptors, https://decs.bvsalud.org/ths/, the Library of Congress Subject Headings (LCSH), https://id.loc.gov/authorities/subjects.html, in addition to many others.

Vocabularies used with LOD need to meet requirements such as having their concepts persistently and univocally identified through valid URIs on the internet, being represented in machine-readable formats such as RDF, containing precise definitions of the

semantics of their concepts, and generally, being multilingual. Many of these vocabularies that meet the principles of LOD can be found in the aforementioned LOV vocabulary registry service. By meeting the requirements for use with LOD as described above, vocabularies, an area of study, research, and practical use of KO, can contribute to addressing the issues brought about by Big Data.

Elements of data or metadata vocabularies referenced by URI account for the semantics of an individual datum, an element of a triple, the "datum" according to Hjørland (2018). An example is the fields of the CRF form. These vocabularies use different approaches to semantics, as pointed out in Almeida et al (2011, p. 195), ranging from a semantics for humans, that is implicit, informal or formal, to semantics for machines, that is informal, formal or even "powerful semantics." In any case, used in the context of the RDF model these vocabularies already allow the processing of RDF triples by machines.

3.3. Semantics beyond the data.

The concept of "powerful semantics," originally devised by Shet, Ramakrishnan, and Thomas (2005) and developed in Shet (2020, slide 42), is defined as "statistical analysis [that] allows the exploration of relationships that are not stated." Semantics is obtained from statistical patterns, not from individual datum referenced by metadata describing an entity, but rather from data sets, or Big Data. Naturally, to identify this semantics, Big Data, whether structured or unstructured, has to be processed by programs. This is so-called data science (Dhar, 2013).

Other vocabularies also have emerged, not to describe or provide standardized values for each piece of data, but to provide descriptive and value metadata of the datasets as a whole. Digital curation is an emerging field of activity for KO professionals.see https://www.dcc.ac.uk/. For the curation of these datasets, metadata standards such as Data Catalog Vocabulary (DCAT) https://www.w3.org/TR/vocab-dcat-2/, or the Provenance Ontology (PROV-O) https://www.w3.org/TR/prov-o/, have been adopted to describe the provenance of the dataset. As datasets have been made available as informational resources on the Web, information on their provenance and the record of the processing carried out on them, the extract, transform, load (ETL) processes (See https://en.wikipedia.org/wiki/Extract,_transform,_load) are essential elements for the data to have credibility and to be able to be reused (See https://www.go-fair.org/fair-principles/).

These datasets, in addition to the metadata that describe their fields, which describe the entity represented by the dataset, have additional metadata provided by vocabularies such as DCAT and PROV-O for the dataset as a whole. For example, they contain metadata such as its format, its quantity, its update date, licenses to use this dataset, etc. (from DCAT), or metadata such as the entity (in this case, the dataset for which the provenance is to be registered), the agent that created the dataset, and the process that generated it (from PROV-O). Standards such as these have been used in several research data repositories to index the files deposited there, an increasingly common application by KO professionals.

3.4. Ontologies as domain models, definitions, specifications

The language specification OWL – Ontology Web Language Overview (2004) states that:

> OWL can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms. This representation of terms and their interrelationships is called an ontology. OWL has more facilities for expressing meaning and semantics than XML, RDF, and RDF-S, and thus OWL goes beyond these languages in its ability to represent machine interpretable content on the Web.

OWL then is a standard language (meta-language in the aforementioned sense) of the W3C for representing ontologies, that is, vocabularies that specify the things existing in a domain and their interrelationships. Further on, the same specification compares the semantic expressiveness of OWL with that of other languages to represent machine-interpretable content such as XML, XML Schema, RDF, and RDFS (ONTOLOGY WEB LANGUAGE OVERVIEW, 2004). It can thus be concluded that, with current technologies, a computational ontology developed in OWL is the most expressive type of KOS, because the "facilities" provided by OWL allow restricting, specifying, and expressing the intended meaning – "intended meaning –" (Guarino, 1994, 560) of the conceptual model of a domain obtained by the modeling process.

Among these facilities is the possibility of specifying data properties (attributes, in Chen's ER model), object properties (relationships in Chen's ER model), domain and scope of the two types of properties, cardinality constraints of each class involved in an object property, transitivity and reflexivity of properties, the disjunction between individuals of different classes, axioms for restricting the inclusion of instances in a class (ONTOLOGY WEB LANGUAGE OVERVIEW, 2004), etc. These facilities can make the conceptual models implicit in a computational ontology in OWL more faithful to reality.

As seen earlier, the Web of Data project, the large-scale reuse of Big Data available in increasing amounts on the Web, depends on the one hand on the most expressive

vocabularies that describe them, and on the other hand on programs capable of make inferences, or at least algorithmic processing, on these representations. In this context, specific domain models, intelligible by machines and represented with the maximum possible semantic expressiveness, gain importance, which, in the current stage of technology, are computational ontologies.

Another important aspect related to this issue; Bergman (2011) discusses ODapps: The Ontology-Driven Application Approach, an automatic program development methodology based heavily on ontologies, a set of them, from high-level ontologies, task ontologies, domain ontologies, to specific application ontologies (Guarino, 1997, 145). In the context of ODApps, domain computational ontologies, with a high degree of semantic expressiveness, are an essential component for developing generic application programs, capable of processing, making inferences, discovering, and reusing the knowledge contained in the domain representation.

It is, therefore, necessary for KO to advance in the creation of computational ontologies of specific domains that are increasingly semantically expressive to equip programs capable of processing these representations to make inferences about them and extract and reuse the knowledge contained therein. The research on patterns of definitions for concepts in ontologies (Campos, 2010) plays a fundamental role in the specification of machine-intelligible semantics, developing the proposals of Dahllberg (1978) of a typology of definitions; just as issues of interoperability between concepts of different ontologies (Barbosa and Campos, 2017), as suggested in Standard 25964-2 (2013), in the SKOS standard (2012) and Zeng (2019).

4. Final considerations

Issues involving information technologies are obscured by the metaphorical denominations often adopted that, didactically and scientifically, make it difficult to understand and operate them, such as Big Data and the Semantic Web. For an accurate understanding of current information technologies, the semantic capacity of computers has to be analyzed, understood, and the real potential identified.

This article sought to demonstrate that data, which have a semiotic and ontological character and are artificial and intentional representations, cannot be understood apart from the entity to which they refer and from the metadata—the properties of this entity—that describe it. Unspecified data is also an imprecise concept. It is necessary to distinguish one piece of datum as referred to by Hjørland (2018), which is a unit that represents the value of one (of the) properties of an entity, from a record, a set of several datum describing various properties of an entity, from datasets, rerpesenting the various entities and their properties, and from databases, bringing together different datasets representing different interrelated entities. The datum as a unit has its own semantics, but records, datasets, and databases already have other levels of semantics in the computational environment.

The Web of Data technologies bring a significant advance by incorporating more semantic expressiveness and program independence to data published on the Web according to the RDF model. In this model, vocabularies can play a significant role, as has been suggested. There are however, several levels of semantics in the variety and heterogeneity of data published on the Web: the “powerful” semantics of the different datasets (the semantic expressiveness of the aggregated datasets of other data), the semantic expressiveness embedded in textual Big Data which needs to be processed for the identification of entity names, named-entity recognition (NER) (Freitas et al 2010), for aggregating annotations and making this data structured, the semantic expressiveness given by programs according

to the data processing model (for data being processed one way and not another), etc. A systematization of these issues should be included in the KO research agenda.

References

Ameida, Mauricio; Souza, Renato and Fonseca, Fred. 2011. "Semantics in the Semantic Web: A Critical Evaluation". *Knowledge Organization,* 38(3):187-203. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1041.7976&rep=rep1&type=pdf, accessed 25 Mar 2021.

Aristóteles. Categorias. Porto: Porto Editora Ltda, 1995.

Barbosa, Nilson. T. and ; CAMPOS, Maria. L. de Almeida. 2017. "A questão da interoperabilidade em repositórios institucionais e sistemas de informação de pesquisas correntes (cris): uma abordagem preliminary". In *Encontro Nacional de Pesquisa em Ciência da Informação, n. XVIII ENANCIB, 2017*. http://hdl.handle.net/20.500.11959/brapci/104600, accessed 25 Dez. 2021.

Bergman, Mike. 2011. "Ontology-Driven Apps Using Generic Applications". *AI3 blog*. https://www.mkbergman.com/948/ontology-driven-apps-using-generic-applications/.

Berners-Lee, Tim. 1998. "Cool URIs don't change". https://www.w3.org/Provider/Style/URI.

Cabré, María Teresa. 2005. A Terminologia, uma disciplina em evolução: passado, presente e alguns elementos de futuro. *Debate Terminológico*. ISSN: 1813-1867, v1. https://www.seer.ufrgs.br/riterm/article/download/21286/15349, accessed 21 Set. 2020.

Campos, Maria Luiza de Almeida. 2010. "O papel das definições na pesquisa em ontologia". *Perspectivas em Ciência da Informação*, 15: 220-238 https://www.scielo.br/j/pci/a/tJr4GnX9Xp7pj5pf44gK4yD/?lang=pt&format=html.

Chierchia, Gennaro. 2003. *Semântica*. São Paulo: Ed. UNICAMP.

CIDOC Conceptual Reference Model Version 5.1.12. 2014. ICOM/CIDOC. http://www.cidoc-crm.org/Version/version-5.1.2, accessed May 3, 2015.

Codd, Eugene. F. 1970. "A relational model of data for large shared databanks". *Communications of The ACM*, 13(6): 377-387. https://dl.acm.org/doi/pdf/10.1145/362384.362685?casa_token=uOdxFTaktMAAAAAA:i_ewo3eO7rDNRE7VYvlBGeHn452O1VQGi69Jn13MciziUeGNMPy827WA6guuZzLkgq4DGl79ocfO4A.

Dahlberg, Ingetraut. 1978. "A referent-oriented, analytical concept theory for INTERCONCEPT". *KO KNOWLEDGE ORGANIZATION*, 5(3): 142-151 https://www.ergon-verlag.de/isko_ko/downloads/ic_5_1978_3.pdf#page=20.

Dhar, Vasant. 2013. "Data science and prediction". *Communications of the ACM*, 56(12):. 64-73. https://dl.acm.org/doi/pdf/10.1145/2500499.

Dextre Clarke, Stella G. 2019. "The Information Retrieval Thesaurus". *KNOWLEDGE ORGANIZATION*, 46(6): 439-459. https://www.ergon-verlag.de/isko_ko/downloads/ko_46_2019_6_c.pdf.

Dextre Clarke, Stella G. and Zeng, Marcia Lei. 2012. "From ISO 2788 to ISO 25964: The evolution of thesaurus standards towards interoperability and data modelling". *Information Standards Quarterly (ISQ)*, 24(1). http://eprints.rclis.org/16818/1/SP_clarke_zeng_isqv24no1.pdf.

Dierickx, Harold and Hopkinson, Alan. 1986. *Reference manual for machine-readable bibliographic descriptions*. http://biblio.cerist.dz/hrbdonf5214/ouvrages/00000000000000594806000000_2.pdf.

FAIR Compliant Biomedical Metadata Templates. 2019. CEDAR, Center for Expanded Annotation and Retrieval, University of Stanford, Department of Medicine. https://medicine.stanford.edu/2019-report/cedar-to-the-rescue.html.

Floridi, Luciano. 2019. "Semantic Conceptions of Information". In *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/win2019/entries/information-semantic/.

Freitas, C.; Carvalho, P.; Oliveira, H. G.; Mota, C. and Santos, D. 2010. "Second HAREM: advancing the state of the art of named entity recognition in Portuguese". In Nicoletta Calzolari et al. (eds.), Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010). European Language Resources Association, pp. 3630-3637. Valletta, 2010.

Giunchiglia, Fausto; Dutta, Biswanath and Maltese, Vincenzo. 2014. "From knowledge organization to knowledge representation". *KNOWLEDGE ORGANIZATION*, 41(1): 44-56, 2014. http://eprints.biblio.unitn.it/4186/1/techRep027.pdf.

Guarino, Nicola. 1997. "Semantic matching: Formal ontological distinctions for information organization, extraction, and integration". In *International Summer School on Information Extraction*. Springer, Berlin, Heidelberg, 1997. 139-170. https://kask.eti.pg.gda.pl/redmine/projects/sova/repository/revisions/5378040326bc499e118636a1d25ad667285e005c/entry/Praca_dyplomowa/materialy/10.1.1.53.939.pdf.

Guarino, Nicola; Carrara, Massimiliano an Giaretta, Pierdaniele. 1994. "Formalizing ontological commitment". In *AAAI*. 1994. p. 560-567. https://www.aaai.org/Papers/AAAI/1994/AAAI94-085.pdf.

Hey, Tony; Trefethen, Anne. 2003. "The data deluge: An e-science perspective". In *Grid computing: Making the global infrastructure a reality*, p. 809-824. https://eprints.soton.ac.uk/257648/1/The_Data_Deluge.pdf.

Hjørland, Birger. (2018). "Data (with big data and database semantics)". *Knowledge Organization*, 45(8): 685-708.

Hjørland, Birger. (2002). "Domain analysis in information science: eleven approaches–traditional as well as innovative". *Journal of Documentation*, 58(4), 422-462.

Hjørland, Birger, and Albrechtsen, Hanne. (1995). "Toward a new horizon in information science: Domain-analysis". *Journal of the American society for information science*, *46*(6), 400-425.

Hjørland, Birger and Hartel, Jenna. 2003. "Introduction to a special issue of Knowledge Organization". *Knowledge Organization*, *30*(3/4), 125-7.

International Council on Archives. Experts Group on Archival Description. 2019. Records in Context: A Conceptual Model for Archival Description (Consultation Draft v0.1). ICA. https://www.ica.org/sites/default/files/ric-cm-0.2_preview.pdf, accessed December 12, 2018.

International Federation of Library Associations and Institutions (IFLA). 1998. *Study Group on Functional Requirements for Bibliographic Records: Final Report.* UBCIM Publications New Series. München: K. G. Saur.

ISO/IEC 20546:2019(en). Information technology — Big data — Overview and vocabulary. ISO, 2019.

ISO 25964-2 - Information and documentation — Thesauri and interoperability with other vocabularies — Part 2: Interoperability with other vocabularies. ISO, 2013.

Lambe, Patrick. 2014. *Organising knowledge: taxonomies, knowledge and organisational effectiveness*. Elsevier.

Marcondes, Carlos H. and Costa, Leonardo C. da. 2016. "A Model to Represent and Process Scientific Knowledge in Biomedical Articles with Semantic Web Technologies". *Knowledge Organization,* 43(2): 122-137. https://www.ergon-verlag.de/isko_ko/downloads/ko_43_2016_2_b.pdf, accessed Apr. 12, 2017.

Marcondes, Carlos H. and Dias, Celia. 2020. "Representing facet classification in SKOS". In International ISKO Conference, Aalborg, Denmark, 16th, *Proceedings… 1. Edition.* Würzburg: Ergon Verlag. *ISBN print: 978-3-95650-775-5, ISBN online: 978-3-*

95650-776-2, Series: Advances in knowledge organization 9. Würzburg: Ergon Verlag, 254–263. *https://doi.org/10.5771/9783956507762*, acessed Fev. 15, 2021.

Marcondes, Carlo. H.; Martins, Sergio. C. and Ramos Junior, Mauricio. C. 2021. The role of vocabularies for the access and reuse of Big Data. *Informação & Informação*, 26(4): 146-174. https://www.uel.br/revistas/uel/index.php/informacao/article/view/44653/pdf. Access 5 Jan. 2022.

De Mauro, Andrea; Greco, Marco and GrimaldiI, Michele. 2015. "What is big data? A consensual definition and a review of key research topics". In *AIP conference proceedings*. American Institute of Physics, 2015. p. 97-104. http://big-data-fr.com/wp-content/uploads/2015/02/aip-scitation-what-is-bigdata.pdf.

Mylopoulos, John. 1992. "Conceptual modelling and Telos". In *Conceptual modelling, databases, and CASE: An integrated view of information system development*, p. 49-68. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.3647&rep=rep1&type=pdf, accessed Dec. 13, 2020.

ONTOLOGY WEB LANGUAGE OVERVIEW. 2004. W3C. https://www.w3.org/TR/owl-features/, accessed 7 Jan. 2022.

Orilia, Francesco and Paoletti, Michele Paolini. 2020. "Properties", *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/win2020/entries/properties/, accessed 20 Sept. 2021.

Otlet, Paul. (2018). Tratado de Documentação: o livro sobre o livro, teoria e prática. Brasília: Briquet de Lemos Livros.

Prieto-Díaz, Ruben. 1990. "Domain analysis: An introduction". *ACM SIGSOFT Software Engineering Notes*, 15(2): 47-54.

RDF semantics. W3C, 2004. http://www.w3.org/TR/rdf-mt/, acessed Mar, 10, 2010.

Ranganathan, S. R. and Gopinath, M. A. *Prolegomena to Library Classification*. 3 ed. Bombay: Asia Publishing House, 1967.

RDF 1.1. PRIMER. 2014. W3C. https://www.w3.org/TR/rdf11-primer/, accessed 12 Dez. 2019.

Resource Description Framework (RDF) Model and Syntax Specification. W3C, 1998. https://www.w3.org/1998/10/WD-rdf-syntax-19981008/. Acessed May 5, 2011.

Saracevic, Tefko. 2007. "Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance". *Journal of the american society for information science and technology*, 58(13): 1915-1933.

Shet, Amith. 2020. "Knowledge Graphs and their central role in big data processing: Past, Present, and Future". In 7th ACM India Joint Conference on Data Science & management of Data (COD-COMAD), Indian School of Business, Hyderabad Campus, 5-7 January 2020. https://www.slideshare.net/apsheth/knowledge-graphs-and-their-central-role-in-big-data-processing-past-present-and-future, accessed Jun. 5, 2021.

Shet, Amith; Ramakrishnan, Cartic and Thomas, Christopher. 2005. "Semantics for the semantic web: The implicit, the formal and the powerful". *International Journal on Semantic Web and Information Systems (IJSWIS)*, 1(1): 1-18.

http://www.ebusinessforum.gr/old/content/downloads/JSWIS.pdf#page=19, accessed Jul 14, 2010.

SKOS – Simple Knowledge Organization System Namespace Document. W3C, 2012. https://www.w3.org/2009/08/skos-reference/skos.html#, accessed Aug 10, 1013.

SPARQL 1.1 QUERY LANGUAGE, 2013. W3C. https://www.w3.org/TR/sparql11-query/, accessed 12 Fev. 2010.

Veiga, Viviane Santo de Oliveira; Campos, Maria Luiza; Silva, Carlos Roberto Lyra; Henning, Patricia and Moreira, João. 2021. "Vodan br: a gestão de dados no enfrentamento da pandemia coronavirus". *Páginas A&B, Arquivos e Bibliotecas (Portugal),* n. Especial: 51-58. http://hdl.handle.net/20.500.11959/brapci/157353, accessed Out 7, 2021.

Wilson, Thomas. D. 1972. "The work of the British Classification Research Group". In Wellish, H. (ed). Subject retrieval in the seventies. Westport: Greenword Publishing Co., 62-71.

Zeng, Marcia Lei. 2019. "Interoperability". In Hjørland, Birger and Gnoli, Claudio eds. ISKO Encyclopedia of Knowledge Organization. ISKO. http://www.isko.org/cyclo/interoperability, accessed Jun 4, 2020.

[i] Today URI evolved towards IRI, the Internationalized Resource Identifier, which strings incorporate characters from alphabets others than the Latin Alphabet

KO-2022-0003 – Answers to the Reviwers comments

07-Aug-2022

Dear Reviewers

Thank you for your valuable comments to our text. We tried to apply them to improve the revised version the text.

The paper was rewritten with focus on digital research data within the context of big data. Text excluded was highlighted in yellow, text added was highlighted red.

Reviewer(s)' Comments to Author:

Reviewer: 1

Comments to the Author

This is a very unique paper, with great connections between big data and knowledge organization. Good that the authors have solid background of Linked Data and semantic technologies. I have a few comments and suggestions and hope the authors will be able to accommodate, turning into a better one that is well-connected to the main trends in big data.

1. The title, “THE ROLE OF VOCABULARIES IN BIG DATA: the quest of research data,” should be reconsidered. Instead of a general term ‘VOCABULARIES’, it should be more specific, especially addressing that these are knowledge organization systems (KOS), not normal dictionaries, thesauri (those not used as a controlled vocabulary), or vocabularies. Answer: The text is discussing vocabularies, i.e. value, data (for exemple, subject) and metadata, descriptive vocabularies (Zeng 2019).
This term is also unclearly used in the text of the paper, e.g., sometime it is clear you are talking about KOS, but you used this term again (1.2. Traditional use of vocabularies to assign subjects to documents).

2. When talking about “Next, we will attempt to demonstrate how the conceptualization above helps address the issues of Big Data, especially research data”, it is not clear that if you already considered research data as big data. Please notice the 5Vs. Big data has been characterized by multiple “V”s, with the number of “V”’s still increasing. Volume (data quantity), Velocity (data speed), Variety (data types and nature), Variability (data consistency), and Veracity data quality) (Kobielus, 2016; Zeng 1017). Any data that have been processed cannot be considered as part of ‘big data’. Even OCLC would not consider its huge data as big data. So, do not just think the ‘volume’ (or amount’ to be the feature of big data (refer to your statement at the beginning of 1.1. The Big Data: “Big Data, the term for a recent phenomenon describing the amount of data produced in digital format.”) Answer: We consider, like Shiri (2013, 18), that Big Data is made up of research data, open data, linked data and semantic. Today’s research data has also the characteristics of Big Data (Fillinger et al. 2019). The same is also stressed in the National Institutes of Health Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA) report (Shiri 2013, 17). Consider the different

information resources containing data of interest to research about the COVID-19 outbreak: scientific articles, healthcare patient records, genomic experiments, posts on social media, etc.

Fillinger, Sven et al. 2019. "Challenges of big data integration in the life sciences." *Analytical and bioanalytical chemistry* 411 no. 26: 6791-6800. doi:10.1007/s00216-019-02074-9

3. The features of the big data need to be enhanced. Big data's five V's are fundamental but the Value should be the one that connecting your research with structured data (metadata) and KOSs. These related to the understanding of the concept of smart data. The following are the references you may consider to read and incorporate into your paper.

• Big Data can bring big Value, if used appropriately, because it is now possible to find the hidden patterns, the unexpected correlations, and the surprising connections within large datasets through effective processing (Gardner, 2012).
• The realization of the last "V", Value, is dependent on "Smart Data," the "ability to achieve big insights from trusted, contextualized, relevant, cognitive, predictive, and consumable data at any scale, great or small" (Kobielus, 2016, p. 8).
• Simply speaking, Smart Data makes sense out of Big Data. It provides value from harnessing the challenges posed by volume, velocity, variety and veracity of big data, in-turn providing actionable information and improving decision making (Sheth, 2014).
• Smart Data "is the way in which different data sources (including Big Data) are brought together, correlated, analyzed, etc., to be able to feed decision-making and action processes" (Iafrate, 2015, p. 13).
• Christof Schöch's paper (2013) is one of the earliest to bring the concept of smart data into humanities. In this article, he wrote, for the Journal for Digital Humanities in 2013, the title is very interesting: Big? Smart? Clean? Messy? Data in the humanities.
o Data has to be cleaned, transformed, and analyzed to unlock its hidden potential.
o Once tamed through organizing and integrating processes, large volumes of unstructured, semi-structured, and structured data are turned into "smart data" that reflect the research priorities of a particular discipline or field.
o Smart data inquiries can then be used to provide comprehensive analyses and generate new products and services.
• In short, the relationship between Big Data and Smart Data can be characterized as "what it is" and "what it is for" (Iafrate, 2015).
Answer: Thank you for your generous contributions. We tried to apply them to the analysis of COVID-19 data within the scope of VODAN Project, section 4.1

4. The abstract can be better written, like majority parts in the paper. May consider not use so many 'we' there. Answer: All "we" are excluded.

Once the authors read these references and consider a better way to connect big data and KOS, I'd be more than happy to review it again.

Reference recommendations:

Iafrate, F. (2015). From Big Data to Smart Data. London: ISTE Ltd., and Hoboken, NJ: John Wiley & Sons, Inc.

IEEE Smart Data conferences. CFPs, etc.

Kobielus, J. (2016, June). The evolution of big data to smart data [PowerPoint slides]. Keynote at Smart Data Online 2016. Video available at https://www.dataversity.net/big-data-smart-data-big-drivers-smart-decision-making/

Schöch, C. (2013). Big? Smart? Clean? Messy? Data in the humanities. Journal of Digital Humanities, 2(3), 13.

Sheth, A. (2014, March). Transforming Big Data into Smart Data: Deriving Value via harnessing
Volume, Variety and Velocity using semantics and Semantic Web [Keynote address]. 30th IEEE
International Conference on Data Engineering, Chicago, IL, United States. DOI:10.1109/ICDE.2014.6816634

Zeng, M. L. (2017). Smart data for digital humanities. Journal of Data and Information Science. 2(1), 1-12. DOI: 10.1515/jdis-2017-0001

Reviewer: 2

Comments to the Author
The topic of the paper is relevant for the journal. Unfortunately, as it stands now, it does not seem to meet necessary quality criteria. CLARITY/LANGUAGE -- The text of the paper is difficult to follow due to the wrong usage / choice of words and other language issues. Occasionally concepts are linked to actions that they cannot perform or are given properties that they cannot have. Therefore the meaning is not entirely clear or sentences do not make sense. An example is in the subtitle itself “the quest of research data” [who is doing the quest here?] COMPOSITION: -- The authors occasionally digress into explaining common knowledge or technical or historical details which are not relevant for the topic of this paper and do not contribute to the argument they are trying to make. RESEARCH BACKGROUND AND CONTEXT -- The proposal outlined in this paper are not novel - therefore it needs to be put into the context of the research already published on this very topic rather than providing history prior to the semantic web and big data phenomena.
Answer: We are interested in showing in detail how value and metadata vocabularies work within LOD technologies to assign meaning to data, section 3.1. The history prior to the semantic web and big data phenomena is necessary to review previous KO initiatives concerning how to represent things within a domain, section 2. There is a lack of references to the literature discussing relationship between Big Data and KO (e.g. Ibekwe-SanJuan & Geoffrey, Kwak, Hajibayova & Salaba, Shiri, Bauer, DeMauro) or KOS/LOD semantic web technology (specifically Zeng, Zeng & Mayr, Busch, Stellato, Isaac, Mendez & Greenberg, etc.) Answer: Now most of such references are cited

FURTHER SPECIFIC ISSUES NEED ADDRESSING:
1- ABSTRACT is poorly structured, incomprehensible and would need to be rewritten. The section ‘Methodology’ does not explain methodology. It contains statements that make no sense. Examples: “materials used are definitions of Big data found in… technologies used in the Semantic Web and Linked Open Data” [technologies do not contain definitions] ---

How does the citation at the bottom of the abstract relate to the abstract or to this research? The paragraph is not properly cited/referenced, it requires a proper reference and page from which the citation is taken, if this is attributed to a person then it would work better if incorporated somehow in the text). This citation should be placed on Page 4 (see L 37). Answer: The abstract has been rewritten.

2- "We present a comprehensive conceptualization of the concept of semantic expressivity and use it to classify the different vocabularies." --- [What does this mean, how and why one uses conceptualization of semantic expressivity to classify different vocabularies? Which vocabularies? What this has to do with Big Data?]. Answer: We excluded such claim. We used the concept of semantic expressivity by Almeida, Souza and Fonseca (2011) to compare different vocabularies and their capacity to represent details of a domain with greater accuracy --- "We identify computational ontologies as a type of knowledge organization system with a higher degree of semantic expressivity. It is suggested that such themes should be incorporated into professional qualifications in KO." --- [How does the statement in the first sentence, which is common knowledge, become a 'theme' and what is meant by "incorporating this into professional qualification"]?

3- PAGE 6 - The authors state (Page 6) that their objective is to show : "how KO can contribute to assigning computational semantics to Big Data, especially to research data, so that computers can process them, allowing their reuse on a large scale". Answer: We changed the focus and objective to "The objective of this work is to discuss how vocabularies, in the sense used within LOD Technologies i.e., value and metadata vocabularies (Zeng 2019), can contribute to assigning computational semantics digital research data within the context of Big Data, so that computers can process them, allowing their reuse on large scale". After describing and explaining well-known characteristics of semantic technology and linked data and associated ontology standards the authors conclusion is that KOS should be made available in the machine-understandable formats i.e. as ontologies using OWL (or some other similar formal ontology language). This has been widely accepted and considered a norm in the KO domain for the past two decades, need not illustration and cannot be represented as a finding. Obviously, once available for machine processing KOS can be utilized for any kind of automatic data processing and linking not withstanding Big Data. Another assertion that the authors wanted to demonstrate is that data cannot be understood without the context i.e. semantics provided through metadata – which is again stating the obvious (data is not information!). Answer: We are not only talking about metadata but also about the entity described by the metadata and data. The authors should consider explaining the context of their research and its focus starting from the above mentioned well known facts.

4- BIBLIOGRAPHIC DATA - The way the authors deal with the topic of metadata principles and bibliographic control, KO and KOS shows some lack of knowledge and it would be better to avoid elaborating these topics in great length. E.g. in section 3, there is an irrelevant historical overview which only demonstrates the lack of understanding of the typology and nature of bibliographic standards. For instance, UNISIST Reference Model is placed in the same category MARC (metadata format schema) Answer: We rewrote such comparison, comparing the MARC format with the UNISIST reference manual for machine-readable bibliographic descriptions (Dierickx and Hopkinson 1985) and ISBD and ACCR ISBD, ACCR, MARC and UNISIST are collectively called "standards" which are metadata/cataloguing description standards. There is, for instance, an entire paragraph on MARC data format development on page 12 – in continuation of the authors' explanation

of the role of KOS in IR systems. The authors somehow relate KOS to MARC and this to relational databases and TREC - none of which making any sense.

While allocating significant space to descriptive metadata – the authors fail to introduce properly subject metadata or explain or connect subject metadata to KOS. For some reason the authors' interest in bibliographic standards including, now obsolete, conceptual model FRBR, does not go beyond the 1990s. While writing about semantic technologies and Big Data both of which have become relevant over the previous 20 years they do not mention BIBFRAME which has in these two decades replaced FRBR - and RDA which has since replaced AACR and other descriptive cataloguing standards. This omission is unacceptable given that these standards were created specifically for a linked data environment and linked data seem to be the utmost focus of this paper. Not to mention, while discussing descriptive metadata standards used in bibliographic domain there is no mention of Dublin Core which is one of the most widely spread descriptive metadata standards in bibliographic domain and beyond. Dublin Core get mentioned in passing only on Page 25. When it comes to KOS one can observe that the authors are not really familiar with this topic. Examples: Page 3 KOS used in IR are explained as 'information retrieval thesaurus', Page 5 ' the early KOS, such as thesauri', Page 12 ' Since the onset of the information explosion, thesauri have emerged, complementary systems to the IRS, of which the KOS were one of its componentssubsystems', Page 16 'Modeling in documentation and information has its roots in bibliographic classification systems such as the Dewey Decimal Classification (DCC) – and the Universal Decimal Classification (UDC). The DCC and UDC can be viewed as a set of taxonomies … The use of taxonomies to organize a domain is typically used today for information management within corporations and to organize the content of websites (Lambe, 2014). Taxonomies only organize the things in a domain…', etc. (SIC!)

> Dierickx, H. (1985). The UNISIST reference manual for machine-readable bibliographic descriptions within the context of international exchange formats. http://nopr.niscair.res.in/bitstream/123456789/27938/1/ALIS%2032(1-2)%207-14.pdf

BALANCE - There seem to be lack of balance in the level of details – some important general principles and aspects of information presentation and information retrieval (e.g. metadata-based retrieval vs text retrieval, metadata typology, metadata architecture including authorities now usually published as linked data), the role of automation, advanced information retrieval techniques are only assumed – then some excessive details are provided for a selected technology/infrastructure, some of which are obsolete (MARC, FRBR), some of which are common knowledge (relational databases / ERM) and some of which are explained with no good reason to a great level of technical detail (RDF, LOD). Some of which are irrelevant in this context (ISBD, AACR, UNISIST, MARC).

PAGE 3 - the authors refer to Big Data as something produced or being made available "by our society", research data being one examples provided. Big Data phenomenon, however, is more frequently associated with data continuously generated by digital technology, computer networks and instruments in e.g. medicine, meteorology, navigation, transport, commerce, industry, satellites, digital cameras/imaging, media. **It would be helpful that at the very beginning the Big Data phenomenon is defined more accurately and comprehensively.** If the authors want to focus on the area of scientific information (as was later done on page 4) – this has to be put into a proper context.

PAGE 3 “Although Big Data has been sparkling interest in KO, contribution from the area to contextualise it or to propose practical solutions are still few” --- [Unlikely - needs to be supported by reference**. Interest by who? Is there any evidence that Big Data has been sparkling interest in KO** when it comes to the domain of computer technology which is the one predominantly dealing with the Big Data issue. The authors addressing this topic come primarily form the KO domain] Answer: This text has been rewritten.

PAGE 4 "… we cannot address Big Data without using computers to help us. This observation refers to the Semantic Web --- [Stating the obvious - Big Data is produced by computers, hence the role of computers is self-evident. How does this observation relate to Semantic Web?] Answer: This text is deleted.

PAGE 5 - "KO methodologies have always represented domains of knowledge when building KOS like controlled /standardized vocabularies, subject headings and classification schemas. The early KOS, such as thesauri --- [why KOS-like, SH and classification are KOS? Not all KO-associated methodologies are dealing with domain knowledge. Thesauri are not “early KOS”] Answer: This text is deleted.

PAGE 5 L48: “empower “knowledge,” which is no longer just inserted into texts to be interpreted by humans, but rather recorded directly in Resource Description Framework” --- [knowledge is not inserted in to texts. One does not ‘record’ in RDF - RDF is a data representation model.] Answer: The text has been replaced by “knowledge (Soergel 2015). It is no longer just inserted into texts to be interpreted by humans, but rather serialized in Resource Description Framework (RDF) triples”.

PAGE 13: " Significantly, bibliographic formats evolved separately from the also nascent technology of computer databases that, from the 1970s onwards, had the relational model as a paradigm (Codd, 1970). The Text Retrieval Conferences (TREC) conference series illustrates this separate evolution. -> Answer: This text is deleted. --- [Irrelevant as well as illogical/wrong series of statements. What do relational databases or any type of database technology have to do with data schemas and formats and their evolution? Although irrelevant for this argument MARC format which is a data element schema / format is used in bibliographic databases that are predominantly relational - but not obligatory. The only thing that is relevant to state in this context is that information retrieval in bibliographic systems is metadata based (irrespective of the type of databases technology used an irrespective of data exchange formats). The text retrieval approach, on the other hand, is not metadata based and depends on availability of digital texts and (advanced) computer based text processing/IR methods. TREC conferences and MARC format do not help in explaining this fact.]

Data Integration for Research and Innovation Policy An Ontologybased Data Management Approach

THE ROLE OF VOCABULARIES IN THE AGE OF DATA: the question of research data

Abstract

Objective: The objective of this work is to discuss how vocabularies, can contribute to assigning computational semantics to digital research data within the context of Big Data, so that computers can process them, allowing their reuse on large scale. Methodology: A conceptualization of data is developed in an attempt to make it clearer what would be data, as an essential element of the Big Data phenomenon, and in particular, digital research data. It then proceeds to analyse digital research data uses and cases and their relation to semantics and vocabularies. Results: Data is conceptualized as an artificial, intentional construction that represents a property of an entity within a specific domain and serves as the essential component of Big Data. The concept of semantic expressivity and use it to classify the different vocabularies and within such classification ontologies are shown to be the type of knowledge organization system with a higher degree of semantic expressivity. Features of vocabularies that may be used within the context of the Semantic Web and the Linked Open Data to assign machine-processable semantics to Big Data are suggested. It is shown that semantics may be assigned at different data aggregation levels.

> The ultimate Big Data challenge lies not in the data, but in the metadata—the machine-readable descriptions that provide data about the data. It is not enough to simply put data online; data are not usable until they can be 'explained' in a manner that both humans and computers can process." Researcher Mark Musen Declaration (FAIR Compliant Biomedical Metadata Templates | CEDAR, 2019).

1. Introduction

How do we discover, access, process, and reuse the huge and growing amount of digital data that is continuously made available by our society, so-called Big Data, a significant part of which is constituted by research data. Big Data has been called the phenomenon describing the huge amount of digital data that is being created at enormous velocity, great heterogeneity as the result of social, economic, scientific and cultural activities centred on the web. Today's research data has also the characteristics of Big Data (Fillinger et al. 2019). Data is created in huge quantities and velocity directly from monitoring devices and projects, like the Hubble Space Telescope, the Human Genome research project, the Large Hadron Collider. Besides the data created directly by scientific activities, Big Data in itself is of interest for scientific research. Shiri (2013, 18) claims that Big Data is made up of research data, open data, linked data and semantic. In today's Web landscape such themes are interwived. Research data is an important product of science, along with scientific publications. How to deal with the "V"s of Big Data, Volume, Velocity, Variety, Variability, Veracity, in research data to enhance its "V"alue and achieve insights of such data (Iafrate 2015, 3)? How can its large-scale reuse be facilitated? Within such a context In light of Big Data and considering the statement by the researcher Mark Musen, what can be the contribution of vocabularies, an important research area in Knowledge Organization (KO) contribute? (Reviewer 1)

1.1. The Big Data

Big Data, the term for a recent phenomenon describing the amount of data produced in digital format, its explosive growth, and the difficulties of storing, processing, and reusing the data, is increasingly present in information technology media. The headlines also call the

phenomenon "information deluge," "data deluge," or "tsunami of data" (Hey and Trefethen 2003). According to these sources, it is impacting business, government, culture, science, and society.

Big Data reminds us of the so-called "information explosion," a fundamental phenomenon connected to the rise of Information Science and KOin the area. In response, KO created knowledge organization systems (KOS) that work in conjunction as auxiliary systems with information retrieval systems (IRS), which are traditionally computerized databases containing representations of scientific documents. Such KOS, the "information retrieval thesaurus" (Dextre Clarke 2016, 138), that control and standardize the natural language used both for indexing the documents entered in the IRS and standardizing natural language the keywords used in the user's queries formulated by users. in an "information retrieval thesaurus" (Dextre Clarke, 2016, 138).

Most The conceptualizations of Big Data tend to repeat those originating in computer science, emphasizeing technological aspects such as volume, variety, velocity, heterogeneity, and the need for massive computer power to process it (Gandomi and Haider 2015). Although Big Data has also been sparking interest in KO (Ibekwe-SanJuan and Bowker 2017, 192), raising questions like its impact in KO epistemology and methodologies (Hajibayova and Salaba, 2018), (Frické 2015) (Reviewer 2). However contributions from the area to contextualize it and propose practical solutions are still few; Hjørland (2013, 179) stressed: "But such progress is brought to us from the outside; it is not something the field of KO has provided". The availability today of huge datasets recording user interactions with different systems, their interests, and preferences, gave rise to the development of data-driven methodologies to guide the interactions between users and such systems, including IRS, an area of application of KO. Nonetheless, methodologies and tools created on their bases have been developed by private enterprises such as Google, Amazon, Netflix.

Hajibayova and Salaba (2018, 147) stress the "opacity of the algorithms behind the platforms and systems".

The best-known product of science, to which the KO has been dedicated since its beginnings, is scientific publications. More recently, science has been giving increasing importance to another of its products, research data. Today, research data, practically entirely digital, is produced in increasing quantities as a result of scientific activity carried out with the support of information technologies. Examples of this huge amount of digital survey data are those generated by the Hubble Space Telescope, https://www.nasa.gov/mission_pages/hubble/main/index.html, the Human Genome research project, https://www.genome.gov/human-genome-project, or the Large Hadron Collider, https://home.cern/science/accelerators/large-hadron-collider, the largest and most powerful particle accelerator in the world. Research data is part of the Big Data phenomenon. A large amount of digital research data now available has even raised debates concerning scientific methodology (Gray 2009), (Leonelli 2012), (Frické 2015).

Research data is defined as "factual records (numerical scores, textual records, images, and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings." (OECD, 2007, 13). Share and reuse of research data presupposes its openness but not only that. BIG DATA X SMALL DATA. As quoted by researcher the Mark Musen at the beginning of this work: "the metadata— the machine-readable descriptions that provide data about the data", has been gaining increasing importance. we cannot address Big Data without using computers to help us. Vocabularies, i.e., data or metadata vocabularies (Zeng 2019), is an important research area in KO. Musen's This observation refers to the Semantic Web project (Berners_Lee et al 2001), the proposal for a Web whose resources would be represented in a

way that had a precise and formal meaning or semantics and would be intelligible and understandable by both people and machines.

In previous works, we have already discussed how to link digital representations of objects of memory and culture through the Web (_____, 2020), and how one of the products of science, scientific publications, could be intelligible and understandable by both people and machines (_____ and Costa 2016), when represented with the technologies of Linked Open Data (LOD) and the Semantic Web. Here we are interested in doing the same for science's other great product, digital research data.

1.2. Traditional use of vocabularies to assign subjects to documents The document centered vision of vocabularies

The technical traditions and standards developed by KO to manage the information explosion rest on resulted in the establishment of IRS/KOS assumptions that persist to this day. In most discourses in the area, these assumptions are so implicit that it becomes difficult to make them explicit, consider them, and analyse their consequences. All the theories and methodologies of KO mentioned bringing these assumptions implicitly: the IRS represent documents in their computerized databases; MARC and the bibliographic formats that emerged from the UNISIST Reference Manual for machine-readable bibliographic descriptions (Dierickx and Hopkinson, 1986) (Reviewer 2) are metadata sets that represent different descriptive properties of the documents.

KOS associated with IRS confirm such assumptions; they "have been designed to support the organization of knowledge and information to make their management and retrieval easier" (Mazzocchi 2018). They are terminological control instruments for documents' descriptive properties, mainly subjects, among a few others (as authorities and geographical

names). used to standardize the records' subject and authorities fields in IRS computerized databases, so useful for users' subject-based (Foskett 1996) retrieval.

These records represent objects that have, among others, the property of having subjects. They are symbolic objects, documents. It is worth adding that the records themselves, the metadata set, are also symbolic objects, representing document-type objects.

These implicit assumptions account for the division that occurs in the teaching and practice of librarianship and KO between descriptive representation and thematic representation, or of subjects, of a document.

Representing documents and their subjects has been foundational to the practices developed by KO, Representing documents and their subjects is a practice with a long tradition in KO. In the past such documents surrogates were especially when, unlike today, there was no access to full-text documents in digital format and the descriptive and thematic representation of the documents was a fundamental mechanism to provide access to information and enable processes of relevance assessment in the intermediation, and relevance assessment processes carried out by libraries and IRS in the retrieval of information (Saracevic 2007). KO methodologies have always represented domains of knowledge when building KOS like controlled/standardized vocabularies, subject headings, and taxonomies classification schemas. The early KOS, such as thesauri, were intended to enable subject-based retrieval in the context of IRS because their records were representations of objects that have as one of their properties subjects. But not all objects in a domain have subjects as one of their properties like documents. We see now that this is just one among many cases of representing different objects in digital space.

To what extent do these assumptions hold up today, and are they sufficient to address the challenges of the Semantic Web era, Big Data, research data, and the Internet of Things?

Today, it is not only the case of about retrieving documents (or their representations) but also to create digital representations of anything, as demanded by such as in the "Internet of Things" (IoT) (Gershenfeld, Krikorian, and Cohen 2004). If the documentation movement (Otlet 2018) and then Information Science intended the empowered rment of information by separating it from books, the Semantic Web proposal and Big Data did the same with the knowledge to also, (Soergel 2015). which It is no longer just inserted into texts to be interpreted by humans, but rather serialized recorded directly in Resource Description Framework (RDF) triples (Reviewer 2) (RDF 1.1 PRIMER 2014), forming representations/descriptions of "things" The Web thus becomes a large knowledge base that can be consulted about the "things" thus represented (SPARQL 1.1 QUERY LANGUAGE, 2013).

The objective of this work is to discuss how KO vocabularies, in the sense used within LOD Technologies i.e., value and metadata vocabularies (Zeng 2019), can contribute to assigning computational semantics digital research data within the context of Big Data to Big Data, especially to research data, so that computers can process them, allowing their reuse on large scale. (Reviewer 2)

As a methodology, the work develops a conceptualization of data and (the few of) Big Data originating in KO in an attempt to make it clearer what would be data, as an essential element of the Big Data phenomenon, and in particular, digital research data. It then proceeds to analyse digital research data uses and cases and their relation to semantics and vocabularies its different levels of aggregation, based on the Case Report Form (CRF), WHO-COVID-CRF/WHO-2019-nCoV-Clinical_CRF-2020.3-eng.pdf at master · FAIRDataTeam/WHO-COVID-CRF · GitHub, proposed by the World Health Organization (WHO) to standardize and unify the registration of cases of patients with COVID-19 worldwide.

The work is organized as follows. After this introduction, section 2 analyses data from a semiotic and ontological point of view. Section 3 presents a comprehensive view of vocabularies within the context of Semantic Web and LOD. Within such a context Section 4 definitions, their traditional use in KO, and develops a conceptualization of data that is illustrated by examples of research data, research datasets, and related initiatives, and shows how research data at different levels of aggregation yields semantics., relating them to the representation of things in a domain and organized into vocabularies. Section 3 presents a comprehensive view of vocabularies based on Semantic Web and LOD technologies and discusses which functionalities vocabularies must incorporate to integrate with these technologies. Section 5 draws conclusions, raises research questions to be developed and presents final considerations.

2. Semiotic and ontological view of data

None of the most common Big Data definitions exclude the data component. It seems reasonable, then, that to understand what Big Data is and how to operationalize solutions to the problem begins by elucidating what is data. After presenting the traditional use of vocabularies to represent and assign subjects to documents this section proposes a semiotic and ontological analysis of data, understood as the essential component of Big Data and research data. This analysis begins with the question of conceptual models and domains and goes on to analyse how conceptual models of domains are expressed linguistically as vocabularies. elucidating how semantics arises from data. Then data is discussed from a semiotic and ontological point of view. From the elucidation of these questions, concepts of research data, data concerning domains of human action, and vocabularies as representations of domains, are developed.

2.3. Traditional use of vocabularies to assign subjects to documents, - generalized use of vocabularies as representations of a domain

Since the onset of the information explosion, thesauri have emerged, complementary systems to the IRS, of which the KOS were one of its components. The development of the IRS drew on sources from other traditions of librarianship, documentation, and cataloging. Catalog sheets and bibliographic entries served as models for computational bibliographic formats in projects such as Machine Readable Cataloging (MARC), based on the AACR2 cataloging standard, and the UNISIST reference manual for machine-readable bibliographic descriptions (Reviewer 2) Reference Manual (Dierickx, Hopkinson 1986), based on the ISBD standard. These formats served as a model for library catalog databases and for indexing and summary services. Significantly, bibliographic formats evolved separately from the also nascent technology of computer databases that, from the 1970s onwards, had the relational model as a paradigm (Codd, 1970). The Text Retrieval Conferences (TREC) conference series illustrates this separate evolution. (Reviewer 2)

Concerning thematic representation, a whole theoretical and methodological foundation were all developed to support the development of KOS, from classification theories, the Faceted Classification Theory (Ranganathan and Gopinath 1967), the proposals of the Classification Research Group (CRG) (Wilson 1972), Concept Theory (Dahlberg 1978), to Terminology (Cabré 2005). This theoretical and methodological tradition, an area of excellence of KO, meets now, with the emergence of the Semantic Web, in subdisciplines such as systems modeling, artificial intelligence, and computational ontologies, areas originating from computer science. These are understood as one of the foundations of the proposed Semantic Web proposal. Many of these new KOS vocabularies are developed by computer professionals and scientists from different areas or specialities: biomedicine, statistics, or from curators of digital collections in memory and culture, etc. The words of

Hjørland (2008, 86) highlight and warn about this approach to other areas: "(LIS) is the central discipline of KO in this narrow sense (although seriously challenged by, among other fields, computer science)." Will KO limit itself to developing traditional KOS and leave this space to computer science specialists as Hjørland warns?

The technical traditions and standards developed by KO to manage the information explosion resulted in the establishment of IRS/KOS assumptions that persist to this day. In most discourse in the area, these assumptions are so implicit that it becomes difficult to make them explicit, consider them, and analyze their consequences. All the theories and methodologies of KO mentioned bring these assumptions implicitly: the IRS represent documents in their computerized databases, MARC and the bibliographic formats that emerged from the UNISIST Reference Manual for machine-readable bibliographic descriptions (Dierickx and Hopkinson, 1986) are sets of metadata that represent different (descriptive) properties of the documents, while the KOS associated with them are terminological standardization instruments specifically for the subject property, the subject field of the records of the IRS computerized databases. These records represent objects that have, among others, the property of having subjects. They are symbolic objects, documents. It is worth adding that the records themselves, the metadata set, are also symbolic objects, representing document-type objects.

These implicit assumptions account for the division that occurs in the teaching and practice of librarianship and KO between descriptive representation and thematic representation, or of subjects, of a document. To what extent do these assumptions hold up today, and are they sufficient to address the challenges of the Semantic Web era, Big Data, research data and the Internet of Things?

2.1. Vocabularies as representations of domains

In the 1980s-1990s, as a consequence of the emergence of online bibliographic catalog management systems and databases, the domain of information retrieval in library catalogues, so familiar to us but also so exclusive, with its diversity of objects, was first modelled using a methodology used in computer science to plan database management systems. The Functional Requirements for Bibliographic Records conceptual model (FRBR) based on Chen (1976) Entity-Relationship (E-R) model, appeared in 1998, whose development was promoted by IFLA (1998).

According to Mylopoulus (1992, 3) "Conceptual modeling is the activity of formally describing some aspects of the physical and social world around us for purposes of understanding and communication." For Mylopoulus,

> the descriptions that arise from conceptual modeling activities are intended to be used by humans, not machines. . . [and] The adequacy of a conceptual modeling notation rests on its contribution to the construction of models of reality that promote a common understanding of that reality among their human users.

A conceptual model sets an agreement between users of a system on what kinds of things exist and will be represented in the system, or entities (also called classes) in a given domain of reality, e.g. documents of historical value, the properties of these entities and how they relate to each other (relationships). Thus, a conceptual model is a representation, in the form of an abstract and generic description, independent of computational implementations (hardware, operating systems, languages, database management systems) of a given domain of reality. It aims at understand this reality, reason about it, and establish a common view of this reality; a conceptual model answers questions such as: What different things exist in a

given domain? How are they distinguished from each other? How do they relate? What are their properties?

As a representation, a conceptual model is expressed, communicated, and externalized through a language, or more specifically a meta-language or meta-model (Guizzardi 2007, 23), which is a language to express the vocabulary (concepts, terms) that express things in specific domains. Examples of these meta-languages are either natural language (through a system requirements document), which functions as the most general of all meta-languages, or a diagrammatic meta-language, such as entity-relationship (meta) Model or the Unified Modelling Language (UML), https://www.uml.org/, class diagram, in which domain-specific ER models or class diagrams are expressed. Both an ER model and a class diagram can define a language that designates things in a domain or a specific vocabulary to that domain.

Within descriptive representation, once established and consolidated practical standards such as MARC, UNISIST, AACR2 and ISDB (Reviewer 2), the question of what are the "things" represented, were implicit in them, their conceptual models, is raised, a view with a higher level of abstraction of a domain.

Conceptual models in the area of documentation and information have made things like documents, authors, and subjects explicit. They evolved from the previously mentioned standards for creating automated bibliographic records, starting with the pioneering FRBR (Ifla 1998). FRBR, as a conceptual model of the bibliographic domain, is not intended for describing or indexing documents, but for formalizing, identifying, agreeing, and standardizing objects, actors, and processes and their relationships within such domain.

Modeling in documentation and information has roots in Universal bibliographic classification systems such as the Dewey Decimal Classification (DCC) – and the Universal

Decimal Classification (UDC). The DCC and UDC can be viewed are used for thematic representation, for assigning subjects – as discipline names - to books. They model the universe of knowledge as a set of taxonomies, each having as a root a discipline into which the universe of knowledge was classified. The use of taxonomies to organize a domain is typically used today for information management within corporations and to organize the content of websites (Lambe 2014). Taxonomies only organize the things in a domain in class-subclass relationships. The things being organized in a universal bibliographic classification are discipline names to be used as subjects to books.

However, there are more than just things or taxonomies of things in a domain. A more accurate model of a domain should include not only the things within it but also their properties, relationships and attributes. Things have properties, attributes, and relationships, according to the ER model. The first movement within documentation and information to recognize this fact was Faceted Classification (Ranganathan, Gopinath 1967). Facets are the properties of a class of things of interest for information recovery (Giunchiglia et al 2014; _____ and Dias, 2020). Besides things, conceptual models also embody properties of things, their attributes, and relationships; Including properties of things Recognizing this fact results in a more accurate representation of a domain, a conceptual model, with richer semantic expressiveness (Almeida, Souza and Fonseca 2011) than a taxonomy (Reviewer 2).

After the pioneering FRBR model (Ifla, 1998), the International Council of Museums (ICOM) adopted the CIDOC Conceptual Reference Model (CIDOC 2014), IFLA released the Library Reference Model (LRM) integrating the FRBR, FRAD, FRSAD models (Riva, Le Boeuf, and Žumer 2017) and more recently the International Council of Archives (ICA) adopted the Records in Context Conceptual Model (Ric-CM) (International Council on Archives 2019). Since the publication of the FRBR model in 1998, KO has been changing its representation activities and methodologies, from records describing documents and their

subjects to conceptual modeling, that is, representing entities, their attributes and relationships (Prasad, Giunchiglia, Devika 2007). Knowledge organization and representation is part of the digital research data curation effort. Such domains of application also uses conceptual models to integrate heterogeneous research data sources as publications, research data, patents, projects, events, funding agencies, etc. (CERIF in Brief 2014)

Conceptual models, when designating things in a domain, define a metadata vocabulary. They are aligned together with different types of KOS by Almeida, Souza and Fonseca (2011, 196), ordered according to their semantic expressiveness. Semantic expressiveness can be understood, in the context of the previous quote, as the ability of each type of KOS to distinguish and describe, that is, identify the properties and represent the different things that exist in a domain of that reality.

Conceptual model elements - entities, attributes and relationships - are expressed linguistically by a vocabulary. Most types of Vocabularies are semantic control devices, formed by systematised sets of semiotic, triadic entities (PEIRCE 1994), concepts (Dahlberg 1978) (Dextre Clarke and Zeng 2012), units of meaning that relate something (a first: object or referents), in some way (through a second: term or code), which generates or induces a third: its meaning. Vocabularies are constructed to answer the basic ontological question: what exists in a domain? They are representations or models of a domain of reality, pointing out what things there are, what their attributes are, their relationships, and how to express them linguistically, the concepts (Dahlberg 1978), and their units of meaning. Online Public Access Catalogs (OPACs) or databases are used as terminological control tools with the IRS used by institutions, with the function of standardising the terms used for the description and indexing of scientific publications, memory and cultural objects, and other items included in these systems.

2.4. 2.2. Domains

Aside from the general library classification systems such as the CDD and the CDU, KOS are developed and used concerning specific domains. The domain notion commonly used in KO is that of a specialized knowledge area.

Hjørland and Albrechtsen (1995, 400), in the text in which they propose the analysis of domains as the foundation of KO, define domains as: "thought or discourse communities, which are parts of society's division of labour." They also label a domain as a "specialty/discipline/domain/environment" (Hjørland and Albrechtsen 1995, 401).

Hjørland (2002, 422) conceptualizes domains associated with specialized libraries, questioning what knowledge would be necessary for information professionals to work in "in a specific subject field like medicine, sociology or music?" In Hjørland and Hartel (2003, 239), this view of domains as systems of thought, theories, is reaffirmed.

> Domains are basically of three kinds of theories and concepts: (1) ontological theories and concepts about the objects of human activity; (2) epistemological theories and concepts about knowledge and the ways to obtain knowledge, implying methodological principles about the ways objects are investigated; and (3) sociological concepts about the groups of people concerned with the objects.

The oldest thesaurus were intended to enable subject-based retrieval in the context of IRS because their records were representations of objects that had subjects as one of their properties, that is, documents. Today, it is not just about retrieving documents (or their representations) but digital representations of anything, as exemplified in the IoT. These representations are no longer just access points for documents, but also information resources themselves, complex descriptions of these objects, and sources of knowledge

about them, represented in such a way that they can be processed/intelligible by both machines and humans. Such representations allow machines to make inferences about the knowledge thus represented.

KO today is being called upon to model different domains of knowledge to build new "semantic" vocabularies, i.e, vocabularies compliant with the Semantic Web and LOD technologies. For this, it is necessary to expand the traditional notion of a domain as a discipline or subject. In the area of software development the notion of a domain has a broader scope: it is 'a sphere of activity or interest: field" [Webster]. In the context of software engineering, it is most often understood as an application area, a field for which software systems are developed (Prieto Diáz 1990, 50).

Since a vocabulary KOS is a terminological system that represents the "things" of interest in a domain of action to the community of agents/users in that domain, then to create a vocabulary KOS (an artifact, similar to software) several aspects and questions must be considered: What things are in a domain? How should they be represented? These are the questions of ontology and semiotics. They must be answered to create a representation, or a conceptual model, of a domain.

We must A first step is to determine what things exist in a domain and which are relevant to this community, what rules exist about these things or are created/approved/agreed on about these things, and how this community uses them to act in this domain. and, fFinally, how the conceptualizations and their agreed terms (Dahlberg 1978), generating as one of the by-products of this process a set of terms, are to be systematised, for example, in a domain model to serve as bases for the construction of vocabularies such as thesaurus or computational ontologies.

What things are in a domain? How should they be represented? These are the questions of ontology and semiotics. They must be answered to create a representation, or a conceptual model, of a domain.

2.5. Vocabularies as representations of a domain.

As shown, a vocabularyies can be (Reviewer 2) is a representations of a domains,. A domain vocabulary can be used regardless of its use, either to assign subjects to documents: a) vocabulary for indexing, which identifies what things exist in a domain (e.g. MeSH categories describing the entities within the Healthcare domain, https://meshb.nlm.nih.gov/treeView, or b) to describe objects in this domain, descriptive metadata standards that, in addition to identify what things exist in a domain, also describe their properties – attributes and relationships. Among the things within a domain some vocabularies focus on specific facets for special purposes: archival science and records management uses functional classification plans in an organization to assign the organizational provenance or the function or organizational process that generated or used a record.

2.1. 2.3. Data as Representations

What is Big Data? What is its relationship with data? What is data and how is it related to metadata? How should semantics be assigned to data? As noted in the ISO/IEC 20546/2019 Standard, "The big data paradigm is a rapidly changing field with rapidly changing technologies," later suggesting a definition: "extensive datasets (3.1.11) — primarily in the data (3.1.5) characteristics of volume, variety, velocity, and/or variability — that require a scalable technology for efficient storage, manipulation, management, and analysis."

The conceptualizations of Big Data originating from KO are few (_____, et al 2021) and replicate those originating in computer science, define it as a phenomenon that involves large amounts of data, the heterogeneity of that data, a continuous flow of generation and updating, and a need for large processing capacity so that the data reveal patterns or trends (De Mauro et al 2015). However, the same is not true for the conceptualizations of data originating from KO. Data is mentioned frequently in the literature, along with its relationships with information and knowledge (Buckland 1991), often called the data, information, knowledge, wisdom (DIKW) hierarchy (Rowley 2007). In Floridi (2019), information is related to data and semantics.

An important exception is from Hjørland (2018), who proposes a conceptualization of Big Data arising from definitions of data, a phenomenon much better known and conceptualized within the area KO. Data is in the essence of the Big Data phenomenon, it could not exist without data. In this work, Hjørland lists several similar conceptualizations of data and highlights that of Fox and Levitin:

> Within this framework, we define a datum or data item, as a triple <e, a, v>, where e is an entity in a conceptual model, a is an attribute of entity e, and v is a value from the domain of attribute a. A datum asserts that entity and has value v for attribute a. Data are the members of any collection of data items.

Such conceptualization is clarified by the following example: “2018.” What does 2018 mean? Others would say it’s a given. Let us note, however, this statement: “Giovana was born in 2018.” In it we can identify the entity we are talking about: a child called “Giovana,” an attribute or property of this entity, she is “born,” and the value of this attribute or property, her birth year, “2018.” To achieve a formal representation it is very important to clearly identify the entity being described. Although a data set usually has a title or

description identifying the entity it represents that is not always the case. A metadata set may mix metadata elements of different entities as for example the MARC21 format field 245 – Title Statement; while MARC21 format describes a bibliographic entity, e.g., a book, field 245 subfield code $c describes another entity, the responsible for the book, and field 245 subfield $f its attributes birth and death dates.

In the ontological scheme that goes back to Aristotle (2000), the reality is constituted of the first substances, the things that have real existence in space and time, and second substances, the conceptualizations we make of the first substances to think, reason, make sense of, and communicate about the things in reality. Second substances are in turn subdivided into essences, concepts designating things that have properties whose loss implies the non-existence of that individual and have existential independence (Fonseca, Porello, Guizzardi, Almeida, and Guarino 2019, 29), and accidents, concepts that designate things that are existentially dependent on other substances. Things having existential independence are commonly recognized in one of the most well-known ontological schemes, the entity-relationships (ER) model (Chen 1976) as entities, while those that are existentially dependent, as properties. Properties, in turn, are subdivided into attributes of an entity, relationships between an existentially independent entity and the value of one of its properties, and relationships, involving two or more individuals of the same, or of different existentially independent entitiesy , or of more than one existentially independent entity (Orilia and Paoletti 2020).

Classifying concepts in vocabularies as entities and their properties, attributes or relationships is a practice that has become common in the specification of vocabulary compliant with LOD technologies; see, for example, the DC Terms vocabulary, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/, the PROV-O ontology,

https://www.w3.org/TR/prov-o/, and DCAT metadata vocabualry, https://www.w3.org/TR/vocab-dcat-3/.

Data is about representations of something else. A piece of data unit, a datum (Hjørland 2018), even in the context of Big Data, then, makes no sense without referencing the entity and one of its properties, the metadata (Reviewer 2). The three concepts are inseparable and cannot be understood separately. They correspond to a descriptive, representational element of an entity, describing one of its properties. They correspond linguistically to a claim, a basic unit of knowledge to which, according to Aristotle (2000, 39), values of truth or falsity can be attributed.

The statements represented by triples constituted by an entity, one of its properties, and the value of this property correspond to the representation of informational resources in the context of LOD, using the RDF data model (RDF Primer 2014). RDF is a Semantic Web standard for describing resources. Everything that is available on the Web can be accessed through a link, or a Uniform Resource Identifier (URI). Today URI evolved towards IRI, the Internationalised Resource Identifier, which strings incorporate characters from alphabets others than the Latin alphabet. This representational model describes such a resource through triples formed by a subject, the resource being described; a predicate, a property that describes the resource; and an object, the value of this property for this resource. The RDF model assumes a minimum semantics, that is, three elements with specific roles, the subject, the predicate, and the object that form the triple are identified and appear in this order.

Semiotic and ontological analysis identifies a piece of data as an artificial and intentional artefact that represents something. The foundational types of the things that exist are entities - existentially independent things - and their properties: relationships between two existentially independent individuals, and attributes of an individual, its qualities and

quantities. Ontological Analysis of things in a domain, classifying and assigning types to these things makes the terms in a domain vocabulary consistent, as they inherit the ontological nature of their types and enable their representations to be machine-processable.

3. A Comprehensive view of vocabularies

In this section, a comprehensive view of vocabularies based on the previous discussion in section 2 and on contributions by Hjørland (2018) and Zeng (2019) was compiled and developed.

3.1. Vocabularies, Web of Data, Linked Open Data, and Big Data

LOD technologies are an integral part of the Web of Data project. Although this is its best-known name, the project is also known as Web of Data, a name that describes it better, since semantics concerns meanings (Chierchia, 2003), and the ability of the Web of Data to convey meanings is quite limited and different from the sense in our understanding of expressions in natural language.

The project was initially formulated by computer scientist Tim Berners-Lee, the creator of the Web, among others. According to its formulators, the Semantic Web aims to propose "A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities" (Berners-Lee et al 2001). To its authors, "Most of the Web's content today is designed for humans to read, not for computer programs to manipulate meaningfully." The Semantic Web then "will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users."

The Web of Data then refers to content represented in such a way that it can be understood by both machines and people. The current Web is made up of pages, such as

http://www.uff.br, formatted in Hypertext Markup Language (HTML), accessible and interconnected with each other through links. Navigating these pages through these links is done by browsers, such as Internet Explorer, Google Chrome, or Mozilla Firefox. HTML is a content markup language; it formats the content of a text of a page through a predefined set of markups, which instruct browsers to display them on computer screens for human users. The content of HTML pages is interpreted by browsers to make it readable and visually pleasing to people.

The proposed Web of Data is quite different. The Web will no longer be constituted of pages to be read by people, but of content, called informational resources, digital representations of things: concrete, like me, you, an industrial product, a monument, a geographical accident; abstract, like a musical genre, a scientific discipline; or just has a digital existence, such as a photo in a JPG file or a scientific article in a PDF file. These are the entities in the proposal by Hjørland (2018). Each of these resources is uniquely identified by a link, or a URI. A resource, identified/accessed by its URI, is described in a structured way through triples, each one formed by the URI of the resource, by each of its properties, and by the corresponding values of each of these properties. An example of how this representational model works is the Leonardo Da Vinci resource on Wikidata, https://www.wikidata.org/wiki/Q762.

This model of structuring data through the description of resources formed by one or more linguistic claims made up of triples <Subject> <Predicate> <Object> is RDF (RDF Primer, 2004). From an ontological point of view, subject, predicate, and object can be understood as an entity, a property, and the value of this property.

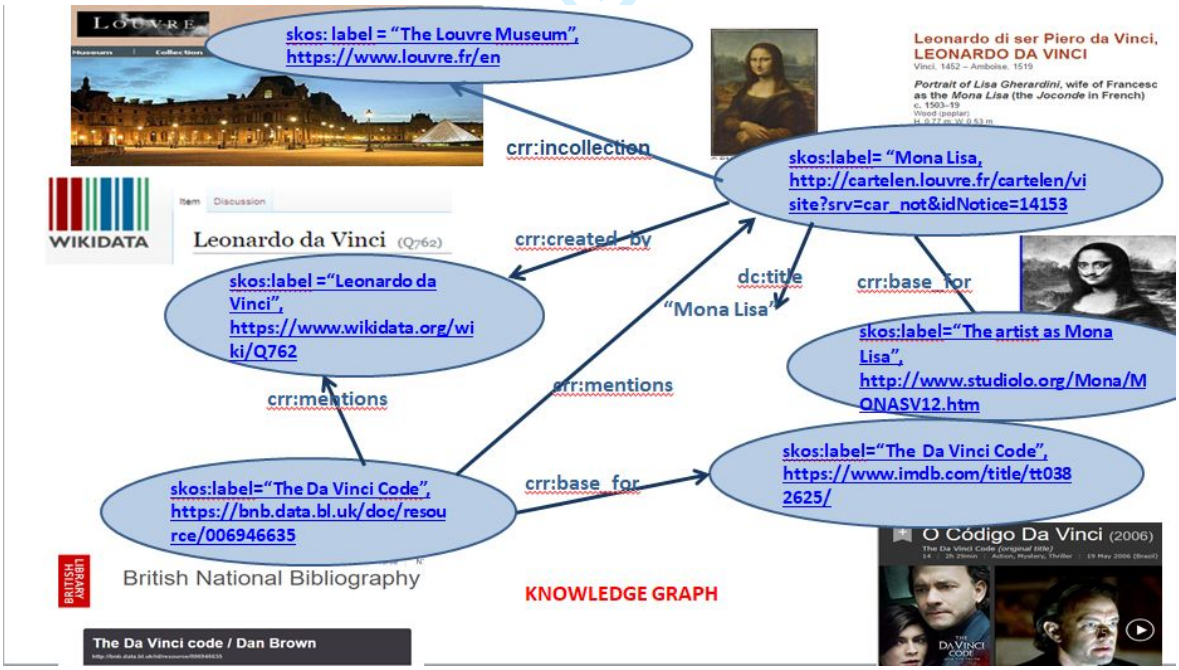
Looking in more detail at structuring a triple; for example,

“The page http://www.uff.br is authored by _____.”

Such a claim consists of three elements: the subject, “http://www.uff.br,” the predicate, “has as author” and the object, “_____”

The RDF model presupposes a minimum semantics, derived from its corresponding linguistic claim. That is, they are identified and appear in this order: the subject, the predicate and the object of the claim that form the triple (Resource Description Framework (RDF) Model and Syntax Specification 1998). A triple describes a specific piece of data from the resource description (what Hjørland calls a “datum:” a unit of data). Sets of triples with the same subject describe the same resource. Sets of interlinked triples describing a resource form a graph. Next, we see the graphical representation of an RDF graph.

FIGURE 2. Graphical representation of an RDF graph



Every computerized system rests on the data processing model. A computer system has as basic components data, processed by programs. While the RDF model describes data, the counterpart in terms of processing are programs that perform inferences on RDF graphs. The minimal semantics of the RDF model allows these programs to navigate through the graphs formed by the triples and infer one or two of the subject(s), predicate(s), or object(s) when they are unknown, such as:

- Who is the author of the page http://www.uff.br?

  - < http://www.uff.br > < authored > < ???>.

- What role does _____ have in relation to the page http://www.uff.br?

  - < http://www.uff.br > < ??? > < _____>.

- What are all the claims about the page http://www.uff.br?

  - < http://www.uff.br > < ??? > < ??? >.

SPARQL is the query language that allows users to query sets of RDF triples (SPARQL 1.1 QUERY LANGUAGE 2013), navigating through the graphs formed by them and performing inferences. It is the materialization of the Web of Data proposal of a Web that can be queried as if it were a database.

RDF can be serialized in several formats, such as RDF/XML, N Triples, JSON, or TURTLE (RDF Primer, 2004). Of course, RDF triples coded in these formats are not as human-friendly or as clearly readable as HTML pages when viewed by browsers. But they contain elements that allow browsers to understand these formats and display them in a human-friendly manner, if applicable. The main objective of the resources described in RDF is that

they can be processed by machines (including their user-friend visualisation), thus helping to organise, retrieve, and make these resources accessible.

Naturally, given a triple like < http://www.uff.br > < is authored > < _____ >, a machine cannot do much more than identify the subject, the predicate or the object of the triple. In this example, predicate and object are names, strings of characters understandable only by people, and holders of a set of contextual and cultural information, accumulated throughout their life histories. RDF Semantics is limited to its model of structuring triples as subject, predicate, and object.

The way to extend these semantics beyond the limits of the RDF model is also to make predicates and/or objects into URI and that these URI refer to concepts of vocabularies with specific semantics. According to RDF Semantics (2004) "There are several aspects of meaning in RDF which are ignored by these semantics; in particular, it treats URI references as simple names, ignoring aspects of meaning encoded in particular URI forms." A URI in the RDF model is just a name, an identifier. The advantage of a URI over a natural language identifier such as the linguistic term "author", is its uniqueness, its validity, since a URI is valid and unique throughout the web space, and its persistence, that is, the commitment of whoever assigns it. a URI to never change it (Berners-Lee 1998).

The previous example can be extended by using URI for the subject, the predicade, and the object of the triple.

<http://www.uff.br> <http://purl.org/dc/elements/1.1/creator> <https://orcid.org/0000-0003-0929-8475>

In this example, the original predicate "author" is replaced by the URI referenced by the "creator" element of the well-known Dublin Core (DC) metadata standard. In its context,

dc:creator has specific semantics. It is defined as "An entity responsible for making the resource." The triple's object, the value or content of dc:creator, has been replaced by the Open Researcher and Contributor ID (ORCID), https://orcid.org, of the page's author.

It is with the semantics in specific vocabularies that the limited semantic expressiveness of the RDF model can be expanded, as seen in the example of the CRF. Once specified in elements of a vocabulary, the semantics can be processed by programs. While the features provided in the Web of Data, represented in markup languages such as XML, RDF, HTML, etc. are contents, programs are procedures or algorithms according to the data processing model. Programs only know how to process content, For this, they need to be clearly instructed (programmed) on what to do with certain content in a certain situation. Specially formatted vocabularies, the LOV (Mendez and Greenberg 2012) used to assign semantics to LOD (Zeng 2019) must clearly define, restrict, and specify the semantics of their concepts. For example, the DC metadata vocabulary clearly defines the semantics of each of its concepts (called elements in the DC initiative); for example, dc:creator, is the creator/author or responsible for a resource, e.g., a digital scientific paper. Furthermore, the dc:creator element has itself, a unique persistent identifier, a link, a URI: http://purl.org/dc/elements/1.1/creator. This persistent identifier, unique throughout the Web space, works as a guarantee of the metadata element semantics, allowing a developer to create a specific program to process this element of the DC vocabulary unambiguously, using the semantics specified and standardize d in the DC vocabulary to the dc:creator element.

Here is another example of what was just explained. Let the following RDF triples be:

TABLE 1. Two triples with the same precicates

<libro0237>	<title>	<Don Quixote>.
<http://catalogo.bne.es/libro0237>	<http://purl.org/dc/elements/1.1/title>	<Don Quixote>. And
<emp0027>	<title>	<President>.
<http://www.company.com/0027>	<http://www.w3c.org/2006/vcard/ns/title>	<President> .

The predicates of both triples are apparently identical as "title." They only differ by the "link" to the vocabulary. In the first example, it is http://purl.org/dc/elements/1.1/title, and in the second it is http://www.w3c.org/2006/vcard/ns/title. These links to different vocabularies, also called namespaces—a kind of delimitation of a scope where those identifiers, with those specific meanings, are valid—allow programs that process the triples to uniquely identify the different concepts in the different vocabularies that serve as predicates for the two triples and even process the two triples simultaneously without confusing their semantics. It is because they are not restricted to the eventual informal meaning of "title" but to this meaning within the scope ("namespaces") of the DC and Vcard, https://devguide.calconnect.org/vCard/vcard-4, vocabularies.

This allows programs to do more than just process inferences about a graph, a set of RDF triples. These are programs oriented by ontology or models, such as Application Program Interfaces (APIs) from the Europeana Library, https://pro.europeana.eu/page/apis.

3.2. Functionalities for vocabularies to be used to assign semantics to data within the context of the Web of Data and LOD

Through unique and persistent identifiers, metadata and data vocabularies can be used to assign machine-understandable semantics to predicates and objects in triples RDF. Many old vocabularies are being restructured to be compatible with LOD technologies (Soergel, 2004; Dos Santos Maculan, 2015), such as the UNESCO Thesaurus, http://vocabularies.unesco.org/browser/thesaurus/en/, the FAO Thesaurus, http://aims.fao.org/aos/agrovoc/c_8003.html, the AGROVOC Thesaurus, https://agrovoc.fao.org/browse/agrovoc/en/, the Paul Getty Foundation Vocabularies, https://www.getty.edu/research/tools/vocabularies/lod/, the Art and Architecture Thesaurus, the Union List of Artists Names, the Cultural Objects Name Authority, the Getty Thesaurus of Geographic Names, the DeCS/MeSH, Health Science Descriptors, https://decs.bvsalud.org/ths/, the Library of Congress Subject Headings (LCSH), https://id.loc.gov/authorities/subjects.html, in addition to many others.

Vocabularies used with LOD need to meet requirements such as having their concepts persistently and uniquely identified through valid URIs on the internet, being represented in machine-readable formats such as RDF, containing precise definitions of the semantics of their concepts, and generally, being multilingual. Many of these vocabularies that meet the principles of LOD can be found in the aforementioned LOV vocabulary registry service. By meeting the requirements for use with LOD as described above, vocabularies, an area of study, research, and practical use of KO, can contribute to addressing the issues brought about by Big Data.

Elements of data or metadata vocabularies referenced by URI account for the semantics of an individual "datum" according to (Hjørland 2018), an element of a triple. An example is

the fields of the CRF form. These vocabularies use different approaches to semantics, as pointed out in Almeida et al (2011, 195), ranging from semantics for humans, which is implicit, informal or formal, to semantics for machines, which is informal, formal, or even "powerful semantics" (Shet, 2020). In any case, used in the context of the RDF model these vocabularies allow the processing of RDF triples by machines.

3.34. Ontologies as domain models, definitions, specifications

Since 1993 Gruber (1993, 199) coined a definition of ontology which is used until nowadays as "An ontology an explicit specification of a conceptualization". Borst (1997, 12) developed Gruber's definition as "Ontologies are defined as a formal specification of a shared conceptualization". Two concepts in this last definition are of importance to the present discussion, - formal, i.e. computers' readable, and – shared, i.e., agreed by a community of agents, being them humans or computers.

The language specification OWL – Ontology Web Language Overview (2004) states that:

> OWL can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms. This representation of terms and their interrelationships is called an ontology. OWL has more facilities for expressing meaning and semantics than XML, RDF, and RDF-S, and thus OWL goes beyond these languages in its ability to represent machine interpretable content on the Web.

OWL is a standard language (meta-language in the aforementioned sense) of the W3C for representing ontologies, that is, vocabularies that specify the things existing in a domain and their interrelationships. Further on, the same specification compares the semantic expressiveness of OWL with that of other languages to represent machine-interpretable

content such as XML, XML Schema, RDF, and RDFS (ONTOLOGY WEB LANGUAGE OVERVIEW, 2004). It can thus be concluded that, with current technologies, a computational ontology developed in OWL is the most expressive type of KOS, because the "facilities" provided by OWL allow restricting, specifying, and expressing the intended meaning (Guarino 1994, 560) of the conceptual model of a domain obtained by the modeling process.

Each concept of an ontology vocabulary is typed; it is a class, or a property of a class or an instance, an individual of a class. Among these facilities are the possibility of specifying data properties (attributes, in Chen's ER model), object properties (relationships in Chen's ER model), domain and scope of the two types of properties, and cardinality constraints of each class involved in an object property, transitivity and reflexivity of properties, the disjunction between individuals of different classes, axioms for restricting the inclusion of instances in a class (ONTOLOGY WEB LANGUAGE OVERVIEW 2004), etc. These facilities can make conceptual models implicit in a computational OWL ontology more faithful to reality. Ontologies also do not distinguish thematic versus descriptive representation; every concept is described by its properties, whether thematic or descriptive.

As seen earlier, the Web of Data project, the large-scale reuse of Big Data and research data available in increasing amounts on the Web, depends on the one hand on the most expressive vocabularies that describe them, and on the other hand, on programs capable of making inferences, or at least algorithmic processing, on these representations. In this context, specific domain models, intelligible by machines and represented with the maximum possible semantic expressiveness such as computational ontologies gain importance, which, in the current stage of technology, are computational ontologies.

Another important aspect related to this issue; Bergman (2011) discusses ODapps: The Ontology-Driven Application Approach, an automatic program development methodology based heavily on ontologies, a set of them, from high-level ontologies, task ontologies, domain ontologies, to specific application ontologies (Guarino 1997, 145). In the context of ODApps, domain computational ontologies, with a high degree of semantic expressiveness, are an essential component for developing generic application programs, capable of processing, making inferences, discovering, and reusing the knowledge contained in the domain representation. It is therefore necessary for KO to advance in the creation of domain-specific computational ontologies domains that are increasingly semantically expressive to equip programs capable of processing these representations to make inferences about them and extract and reuse the knowledge contained therein. The research on patterns of definitions for concepts in ontologies (Campos 2010) plays a fundamental role in the specification of machine-intelligible semantics, developing the proposals of Dahlberg (1978) of a typology of definitions as well as issues of interoperability between concepts of different ontologies (Barbosa and Campos 2017), as suggested in Standard 25964-2 (2013), in the SKOS standard (2012) and Zeng (2019).

2.2. 4. Results

In the sequel the previous conceptualizations are applied to cases of research data and discussed.

4.1.Data, and Big Data,: the case for research data

We will attempt to demonstrate how the conceptualizations above helps address the issues of Big Data, especially research data. A concrete and dramatic example of the importance of research data and the adoption of principles and technologies that allow its wide dissemination and reuse is the form for collecting data from patients infected with COVID-

19, the CRF, which was proposed by the WHO. The GO FAIR initiative, https://www.go-fair.org/, addresses the WHO proposal by creating a worldwide network of catalogs referencing research data collected through the CRF and deposited in repositories and available according to the FAIR principles, https://www.go-fair.org/fair-principles/, the "FAIR Data Points." Brazil participates in this initiative through the VODAN-Br Virus Outbreak Data Network initiative (Veiga et al 2021).

The VODAN initiative is expected to collect huge datasets worldwide. The CRF standardized a set of fields of interest to COVID-19 epidemic research. Such fields must be filled with metadata and data associated with vocabularies largely agreed and standardized within the health sciences domain. This allows the interoperability of different datasets and , without which their processing by computers in order would not be possible, and consequently, neither drawing conclusions and insights from the data would have the ability to extract conclusions and insights. VODAN and FAIR Data Points are efforts to provide smart data (Kobielus 2016) to be used to control COVID-19 outbreak.

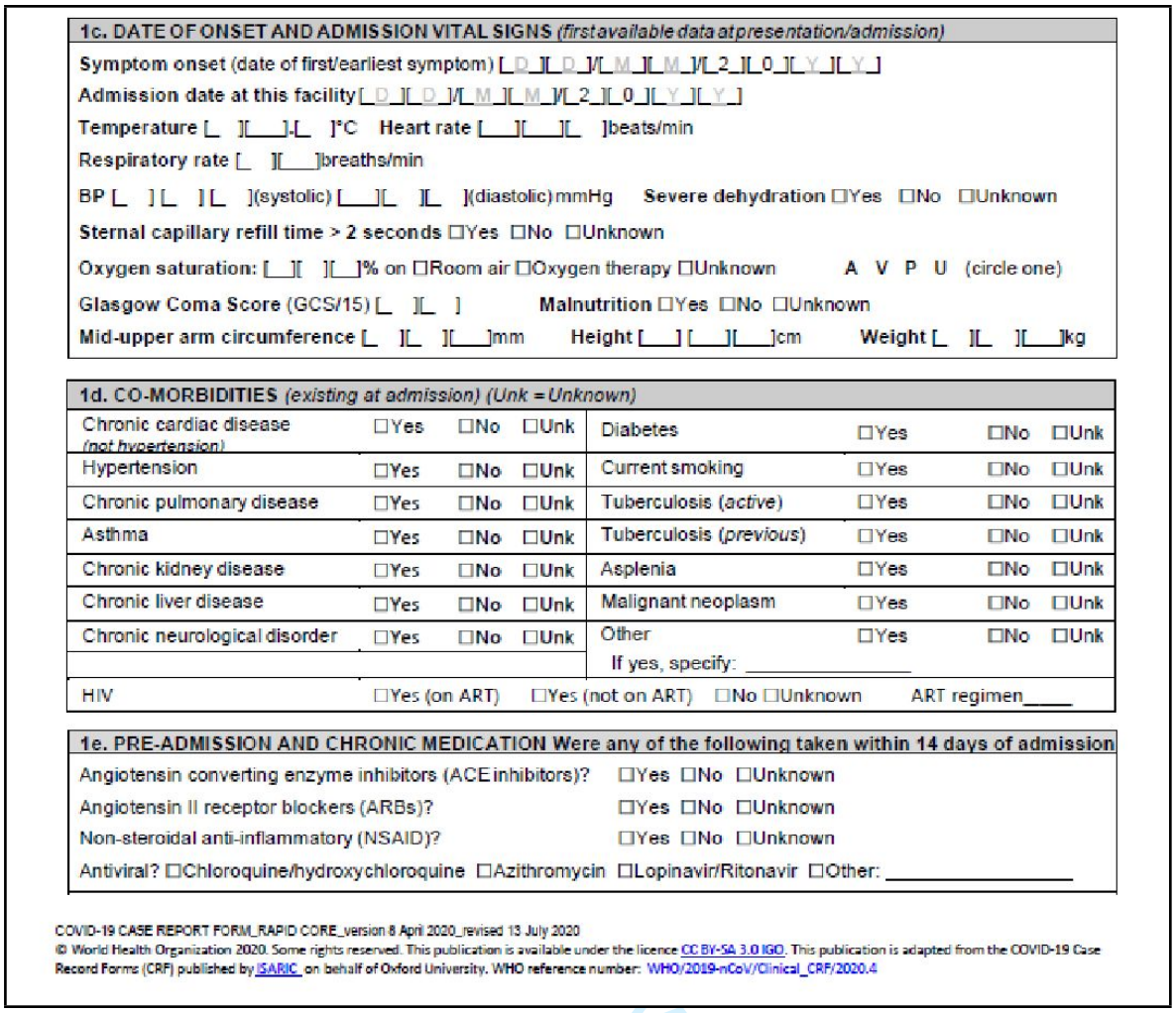
Within the RDF model, instead of the subject, predicate, and object of a triple being represented in natural language, which is ambiguous and difficult for programs to process, each can be identified by a URI. These URIs identify specific terms, both from metadata vocabularies—descriptive properties of things in a domain—and data vocabularies—values assumed by these properties for specific descriptive metadata. This unified characterization of vocabularies as, that is, sets of systematized terms that identify either the descriptive properties (metadata) of objects in a domain, or the values - the data - assumed by these properties for instance, is due to Marcia Zeng (2019).

Another important feature of using vocabularies with LOD technologies is that different vocabularies can be used simultaneously in the form fields. In Figure 1 shows we see an

excerpt from the CRF. As co-morbidity data, "CO-MORBIDITIES," of a patient (the entity) are recorded, concepts such as chronic cardiac disease (the attribute or metadata) are taken from specific biomedical ontologies or vocabularies: Yes, No, Unk (the value or data). the co-morbidity data, "CO-MORBIDITIES," of a patient (the entity); they are recorded as follows: concepts such as chronic cardiac disease (the attribute or metadata, the co-morbidity presented by the patient) are taken from specific biomedical ontologies or vocabularies that describe specific co-morbidity types; if a specific one applies, it is recorded as data as follows: Yes, No, Unk. These data have to be processed by programs so that the immense amount of records collected through the CRF around the world can serve as inputs for the planning and control of the pandemic. The question about co-morbidities has several answer options, each of which indicates a type of disease. For it to be processed by machines, each type of co-morbidity expressed in natural language must reference a concept in a vocabulary or ontology, such as SNOMED-CT, https://www.nlm.nih.gov/healthit/snomedct/index.html, for example. Another question on the CRF, such as the one related to "PRE-ADMISSION AND CHRONIC MEDICATION," has as one of its answer options "Angiotensin converting enzyme inhibitors (ACE inhibitors)?", which may be referenced in another vocabulary such as MeSH, https://meshb.nlm.nih.gov/search, the term with identifier http://id.nlm.nih.gov/mesh/D000806.

In order to have precise meaning, concepts such as those shown in the CRF must refer to specific, standardized ontologies or biomedical vocabularies to enable the processing of these data.

FIGURE 1 - Part of the CRF Form



1c. DATE OF ONSET AND ADMISSION VITAL SIGNS *(first available data at presentation/admission)*

Symptom onset (date of first/earliest symptom) [_D_][_D_]/[_M_][_M_]/[_2_][_0_][_Y_][_Y_]

Admission date at this facility [_D_][_D_]/[_M_][_M_]/[_2_][_0_][_Y_][_Y_]

Temperature [_][___].[_]°C Heart rate [___][___][_]beats/min

Respiratory rate [_][___]breaths/min

BP [_] [_] [_](systolic) [___][_][_](diastolic) mmHg Severe dehydration ☐Yes ☐No ☐Unknown

Sternal capillary refill time > 2 seconds ☐Yes ☐No ☐Unknown

Oxygen saturation: [__][ ][__]% on ☐Room air ☐Oxygen therapy ☐Unknown A V P U (circle one)

Glasgow Coma Score (GCS/15) [_][_] Malnutrition ☐Yes ☐No ☐Unknown

Mid-upper arm circumference [_][_][___]mm Height [___] [___][___]cm Weight [_][_][___]kg

1d. CO-MORBIDITIES *(existing at admission) (Unk = Unknown)*

Chronic cardiac disease *(not hypertension)*	☐Yes	☐No	☐Unk	Diabetes	☐Yes	☐No	☐Unk
Hypertension	☐Yes	☐No	☐Unk	Current smoking	☐Yes	☐No	☐Unk
Chronic pulmonary disease	☐Yes	☐No	☐Unk	Tuberculosis (*active*)	☐Yes	☐No	☐Unk
Asthma	☐Yes	☐No	☐Unk	Tuberculosis (*previous*)	☐Yes	☐No	☐Unk
Chronic kidney disease	☐Yes	☐No	☐Unk	Asplenia	☐Yes	☐No	☐Unk
Chronic liver disease	☐Yes	☐No	☐Unk	Malignant neoplasm	☐Yes	☐No	☐Unk
Chronic neurological disorder	☐Yes	☐No	☐Unk	Other	☐Yes	☐No	☐Unk
				If yes, specify: _____			
HIV	☐Yes (on ART)	☐Yes (not on ART)	☐No ☐Unknown	ART regimen_____			

1e. PRE-ADMISSION AND CHRONIC MEDICATION Were any of the following taken within 14 days of admission

Angiotensin converting enzyme inhibitors (ACE inhibitors)? ☐Yes ☐No ☐Unknown

Angiotensin II receptor blockers (ARBs)? ☐Yes ☐No ☐Unknown

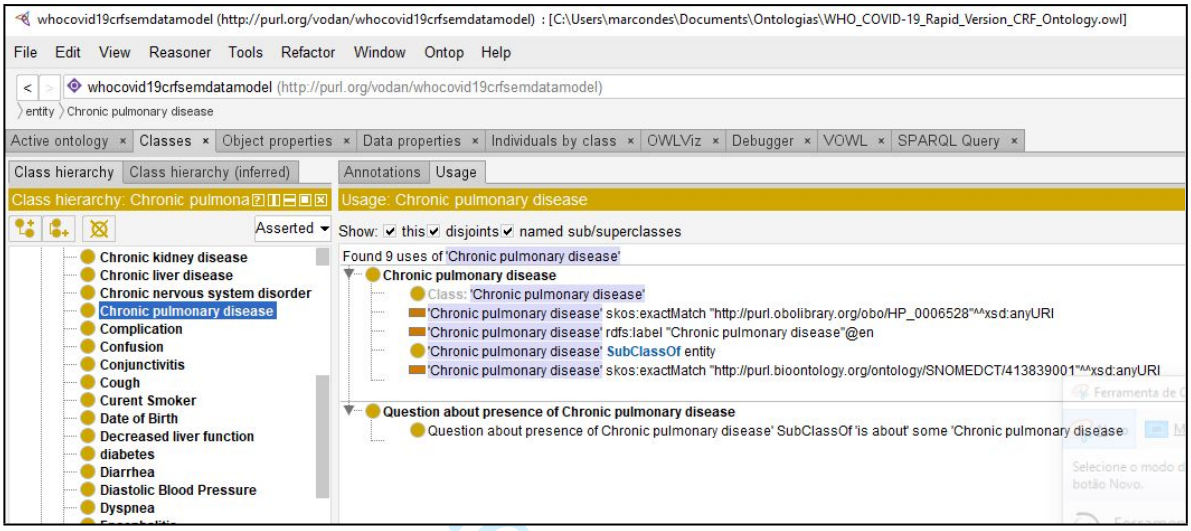
Non-steroidal anti-inflammatory (NSAID)? ☐Yes ☐No ☐Unknown

Antiviral? ☐Chloroquine/hydroxychloroquine ☐Azithromycin ☐Lopinavir/Ritonavir ☐Other: _____

COVID-19 CASE REPORT FORM_RAPID CORE_version 8 April 2020_revised 13 July 2020

The CRF is formalized by a conceptual model and owl ontology, the WHO COVID-19 Rapid Version CRF semantic data model, https://bioportal.bioontology.org/ontologies/COVIDCRFRAPID. In the following Figure 2 another feature of KOS methodologies and standards incorporated in ontologies is the mapping properties. Mapping properties of a concept in a KOS identify which concept in that KOS means the same as another concept from another KOS, i.e., the mapping one concept to another concept. The concept "chronic pulmonary disease" at Figure 1 is shown in Figure 2 as a class of the WHO COVID-19 Rapid Version CRF semantic data model; it is also shown its skos:exactMatch to the SNOMED concept "413839001".

FIGURE 2 – The class “chronic pulmonary disease” of the WHO COVID-19 Rapid Version CRF semantic data model and its SKOS mapping to the SNOMED concept



Each field in the CRF gives rise to a RDF triple in which the PARTICIPANT ID, the patient, is the subject, the field (standardized and referenced by a metadata vocabulary) is the predicate and its value (also standardized and referenced by a value vocabulary) is the object.

As previously stated, openness is essential to enable research data sharing and reuse. For data to be considered open, international recommendations rate it from 1 to 5 stars, https://5stardata.info/en/. The fourth and fifth stars are awarded when data is available in RDF format, including be accessible through a URI, their predicates and objects be referred by standardized vocabularies widely recognized by the community in a given domain, and linked together to provide rich context. For research data, which has demanded increasing attention and public policies at national and international levels, the international GO FAIR initiative recommends a set of principles for publication so that they have the attributes of FAIR: findability, accessibility, interoperability, and reuse. To achieve such attributes

research data must be accessible through a URI, represented in RDF, constituting the Linked Open Vocabularies.

The FAIR principles allow research data to be processed by machines. The M4M principle—metadata for machines—states that "There is no FAIR data without machine-actionable metadata. The overall goal of Metadata for Machines workshops (M4M) is to make routine use of machine-actionable metadata in a broad range of fields." The CRF described above is an example of the importance of research data standardization and the adoption of principles that allow its wide dissemination and reuse is the CRF form described above. Without standardization, its processing by machines would be impossible.

Applying the FAIR principles to research data causes data to be represented as RDF triples. Such a process is named "FAIRification", see https://www.go-fair.org/fair-principles/fairification-process/. FAIR compliant data is generally derived data from datasets. A distributed network of FAIR Data Points provides access to different FAIR data. That raises the question of using vocabularies to describe both the original datasets and their FAIR compliant datasets versions generated.

RDA – Research Data Alliance, https://www.rd-alliance.org/

Other vocabularies also have emerged, not to describe or provide standardized values for each piece of data, but to provide descriptive and value metadata of the datasets as a whole. Digital curation of research data is an emerging field of activity for KO professionals; one of its activities is to apply metadata to research datasets, see https://www.dcc.ac.uk/. For the curation of these datasets, metadata standards such as Data Catalog Vocabulary (DCAT) https://www.w3.org/TR/vocab-dcat-2/, or the Provenance Ontology (PROV-O) https://www.w3.org/TR/prov-o/, have been adopted to describe the provenance of the dataset. As datasets have been made available as informational resources on the Web,

information on their provenance and the record of the processing carried out on them, the extract, transform, load (ETL) and the FAIRrification processes of such data, see https://en.wikipedia.org/wiki/Extract,_transform,_load) are essential elements for research data reliability to enable sharing and reuse (See https://www.go-fair.org/fair-principles/).

The amount of research data being available every day on Coronavirus epidemic – the "V"ariety" of Big Data - makes the integration of such sources essential to control the epidemic. The Coronavirus Infectious Disease Ontology (CIDO) (He et al. 2020) stresses the essential role computational ontologies in the integration of different and heterogeneous research data sources, promoting interoperability between such sources.

These datasets, in addition to the metadata that describe their fields, are themselves of interest for the research data exploration which describe the entity represented by the dataset. They need additional metadata provided by vocabularies such as DCAT, PROV-O, as the type of licence under which data can be reused, the dataset creator, its Publisher, its format, its update date, etc, all of which are metadata for the dataset as a whole. Such metadata is provided by vocabularies as DCAT, PROV-O. They contain metadata such as the format of the dataset, the number of records, the last update date, licences to use this dataset, etc. (from DCAT), or metadata such as the entity who generate it (in this case, the dataset for which the provenance is to be registered), the agent that created the dataset, and the process that generated it (from PROV-O). Standards such as these have been used in several research data repositories to index the datasets deposited there. Indeed, digital curation is an increasingly common application by KO professionals (Poole 2013).

Digital Humanities is another growing area of application of digital research data. It grew from the wide availability of data from social activities (search and social media activity every minute, see https://www.smartinsights.com/internet-marketing-statistics/happens-online-60-seconds/) and culture, including science. Scientific articles have long been recognized as a privilege knowledge source (Swanson, 2008), see PubMed Citations per year, https://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html). Significant examples of research projects in Digital Humanities using a variety of such sources can be found in the Digging into Data Challenge program (https:// diggingintodata.org/) mentioned by Zeng (2017); in this article, the author describes in details how Digital Humanities is related to Big Data and the challenges to process such data and turn it into Smart Data (Reviewer 1).

A huge amount of such data is textual, resulting from posts on social media, emails, newspaper articles, scientific articles, and text in encyclopaedias such as Wikipedia, among others. This data is unstructured or semi-structured.

The exploitation of such potential information sources may lay on the development of vocabularies for special purposes. Their processing using techniques such as information extraction, named-entity recognition, natural language processing, text mining, machine learning, text annotation, aim at transforming such non-structured or semistructured textual data into structured.

Examples of such techniques in biomedical sciences are the National Library of Medicine Natural Language Processing tools, https://lhncbc.nlm.nih.gov/LHC-research/nlp.html, which lay on dictionaries and KOS like MeSH, the Medical Subject Headings, and UMLS, the Unified Medical Language System (Bodenreider 2004), (Aronson and Lang 2010).

3.3. 4.2. Semantics beyond the data.

Semantics is a very general concept. An operational concept of semantics applied to messages – data: in digital environment is the inference made by an agent based on a message that enables such agent to make decisions and, possibly, to act accordingly.

The concept of "powerful semantics," originally devised by Shet, Ramakrishnan, and Thomas (2005) and developed in Shet (2020, slide 42), is defined as "statistical analysis [that] allows the exploration of relationships that are not stated." Semantics is obtained from statistical patterns, not from individual datum referenced by metadata describing an entity, but rather from data sets as a whole, or Big Data. To identify this semantics, Big Data, whether structured or unstructured, has to be processed by programs. This is so-called data science (Dhar 2013).

Entities are the units to be represented by digital metadata and data within a domain, even if an entity is represented by only one of its properties. As so they are the units of meaning and correspond to what has been called a digital object. The concept of a digital object was first proposed in 1995 by Kahn and Wilensky (2006) as a set of bits that has a special interest in applications or software agents; it is related to the concept of data as a representation of an entity or phenomenon (Hjørland 2018). Digital objects of interest to research data are also just now (see https://www.fdo2022.org/) being conceptualized by initiatives such as FAIR Digital Object Framework: "In the FDOF, a digital object is a bit sequence located in a digital memory or storage that has, on its own, an informational value, i.e., the bit sequence represents an informational unit such as a document, a dataset, a photo, a service, etc", see https://fairdigitalobjectframework.org/.

Within the Web of Data context vocabularies are meaning control and standardization artefacts aimed at making knowledge records meaningful. The previous discussion poses the question of levels of meaning related to levels of data aggregation. Table 1 sketch the relationships between data aggregation levels to digital units of meaning.

DATA AGGREGATION LEVELS	DIGITAL UNITS OF MEANING
Level 1 - a datum (Hjørland 2018), the basic element of data	the value of a database field, the content or an excel cell
Level 2 - an RDF triple, a field and its content of a specific row in a database.	a proposition, state of affairs (JANSEN, 2008, 188), Hjørland (2018) (e, a, v) citing Redman, Fox and Levitin (2017, 1173), a triple of an entity, a metadata, and a datum.
Level 3 - a row in a specific database table, a digital object, a named graph	A data structure, a conceptualization, a message (CAPURRO, 2000)
Level 4 - a dataset, a database, an ontology populated with its instances	Several descriptions of different entities, a graph, a conceptualization based on a specific conceptual model, data mining on a specific dataset, an insight from processing a dataset (Dhar, 2013).
Level 5 - A research data repository as re3data, https://www.re3data.org/, described by a metadata vocabulary (Strecker et al, 2021), several heterogeneous datasets of interest for a theme or problem.	Several conceptualizations, several conceptual models. In such cases an ontology with the aid of the mapping properties specified in SKOS model (SKOS 2012) and in ISO 25964-2 Thesauri standard (ISO 25964-2 2013) may holds the agreed semantics that enable the

	integration and interoperability between such different and heterogeneous research data sources.

4. 5.Final considerations

Issues involving information technologies are obscured by the metaphorical denominations often adopted that, didactically and scientifically, make it difficult to understand and operate them, such as Big Data and the Web of Data. For an accurate understanding of current information technologies, the semantic capacity of computers has to be analysed, understood, and the real potential identified.

The Web of Data technologies bring a significant advance by incorporating more semantic expressiveness and program independence to data published on the Web. Big Data and research data also poses several issues related to the semantic of data. This article sought to demonstrate that data, which have a semiotic and ontological character and are artificial and intentional representations, cannot be understood apart from the entity to which they refer and from the metadata—the properties of this entity—that describe it.

As stressed by Ibekwe-SanJuan and Bowker (2017, 187) “In essence, Big Data will not remove the need for humanly-constructed KOSs”. This article suggests some paths towards the role of vocabularies in addressing the issues raised by research data in the age of Big Data. Web environment, Big Data, and research data together comprise a heterogeneous environment that poses the challenge of making different resources work together. Semantic interoperability is the key to achieve such goal. KOS as conceptual models and ontologies play a central role in the semantic integration of different and heterogeneous research data sources, promoting interoperability between such sources. In practical terms ontologies hold

representation of a domain while mapping properties (SKOS 2012), (ISO 25964-2 2013) and also OWL property "sameAs" (Ontology Web Language Overview (2004) enable the mapping of concepts in a data resource to concepts in another.

User-generated content, folksonomies (Hajibayova and Salaba 2018, 145)

The Web of Data technologies bring a significant advance by incorporating more semantic expressiveness and program independence to data published on the Web. Big Data and research data also poses several issues related to the semantic of data. This article sought to demonstrate that data, which have a semiotic and ontological character and are artificial and intentional representations, cannot be understood apart from the entity to which they refer and from the metadata—the properties of this entity—that describe it. Unspecified data is also an imprecise concept. It is necessary also to distinguish one piece of datum as referred to by Hjørland (2018), which is a unit that represents the value of one (of the) properties of an entity, from a record, a set of several datum describing different properties of an entity, from datasets, representing the various entities and their properties, and from databases, bringing together different datasets representing different interrelated entities. Such are different data aggregation levels, Datum as a unit is incomplete, meaningless without knowing the entity to which it refers and the specific property of that entity referred. has poor semantics, but such data aggregates having higher levels of semantics in the computational environment. Vocabularies can play an important role in addressing semantics to data at those different levels of aggregation.

The Web of Data technologies bring a significant advance by incorporating more semantic expressiveness and program independence to data published on the Web according to the RDF model. In this model, vocabularies can play a significant role, as has been suggested. Big data and research data poses several issues related to semantic of data: the

semantic of each piece of datum, the semantic of and RDF triple, of a knowledge graph, the "powerful" semantics of the different datasets (the semantic expressiveness of the aggregated datasets of other data), the semantic expressiveness embedded in textual Big Data which needs to be processed for the identification of entity names, named-entity recognition (NER) (Freitas et al 2010), for aggregating annotations and making this data structured, the semantic expressiveness given by programs. Vocabularies can play an important role in addressing them. There are however several levels of semantics in the variety and heterogeneity of data published on the Web according to its levels of aggregation: the "powerful" semantics of the different datasets (the semantic expressiveness of the aggregated datasets of other data), the semantic expressiveness embedded in textual Big Data which needs to be processed for the identification of entity names, named-entity recognition (NER) (Freitas et al 2010), for aggregating annotations and making this data structured, the semantic expressiveness given by programs according to the data processing model (for data being processed one way and not another), etc. A systematisation of these issues should be included in the KO research agenda.

Acknowledgments: This work was carried out with the support of the Brazilian agencies CAPES - Financing Code 001, and CNPq, grant number 305253/2017-4. We are also grateful to the anonymous reviewers of this work for their suggestions to improve this text.

References

Almeida, Mauricio; Souza, Renato and Fonseca, Fred. 2011. "Semantics in the Semantic Web: A Critical Evaluation". *Knowledge Organization* 38 no. 3: 187-203. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1041.7976&rep=rep1&type=pdf, accessed 25 Mar 2021.

Aristóteles. Categorias. Porto: Porto Editora Ltda, 1995.

Aronson, Alan R., and François-Michel Lang. 2010. "An overview of MetaMap: historical perspective and recent advances." *Journal of the American Medical Informatics Association* 17 no. 3: 229-236.

Barbosa, Nilson. T. and; Campos, Maria. L. de Almeida. 2017. "A questão da interoperabilidade em repositórios institucionais e sistemas de informação de pesquisas correntes (cris): uma abordagem preliminary". In *Encontro Nacional de Pesquisa em Ciência da Informação, n. XVIII ENANCIB, 2017*. http://hdl.handle.net/20.500.11959/brapci/104600, accessed 25 Dez. 2021.

Bergman, Mike. 2011. "Ontology-Driven Apps Using Generic Applications". *AI3 blog*. https://www.mkbergman.com/948/ontology-driven-apps-using-generic-applications/.

Berners-Lee, Tim. 1998. "Cool URIs don't change". https://www.w3.org/Provider/Style/URI.

Bodenreider, Olivier. (2004). "The unified medical language system (UMLS): integrating biomedical terminology." *Nucleic acids research* 32 no. suppl_1: D267-D270.

Borst, Willem N. 1997. *Construction of Engineering ontologies*. Centre for Telematica and Information Technology. University of Twenty, Enschede, The Nederalands.

Cabré, María Teresa. 2005. A Terminologia, uma disciplina em evolução: passado, presente e alguns elementos de futuro. *Debate Terminológico*. ISSN: 1813-1867, v1. https://www.seer.ufrgs.br/riterm/article/download/21286/15349, accessed 21 Set. 2020.

Campos, Maria Luiza de Almeida. 2010. "O papel das definições na pesquisa em ontologia". *Perspectivas em Ciência da Informação*, 15: 220-238 https://www.scielo.br/j/pci/a/tJr4GnX9Xp7pj5pf44gK4yD/?lang=pt&format=html.

Capurro, R. 2000. "Angeletics—A message theory. In H.H. Diebner & L. Ramsay (Eds.) (2003), Hierarchies of communication". *In inter-institutional and international symposium on aspects of communication on different scales and levels*. Karlsruhe, Germany: ZKM. Retrieved July 25, 2005, from http://www.capurro.de/angeletics_zkm.html.

CERIF in Brief. (2014). https://eurocris.org/eurocris_archive/cerifsupport.org/cerif-in-brief/index.html

Chen, Peter Pin-Shan. 1976. "The Entity-Relationship Model-Toward a Unified View of Data". *ACM Transactions on Database Systems* 1 no.1: 9-36.

Chierchia, Gennaro. 2003. *Semântica*. São Paulo: Ed. UNICAMP.

CIDOC Conceptual Reference Model Version 5.1.12. 2014. ICOM/CIDOC. http://www.cidoc-crm.org/Version/version-5.1.2, accessed May 3, 2015.

Codd, Eugene. F. 1970. "A relational model of data for large shared databanks". *Communications of The ACM*, 13(6): 377-387. https://dl.acm.org/doi/pdf/10.1145/362384.362685?casa_token=uOdxFTaktMAAAAAA:i_ewo3eO7rDNRE7VYvlBGeHn452O1VQGi69Jn13MciziUeGNMPy827WA6guuZzLkgq4DGl79ocfO4A.

Dahlberg, Ingetraut. 1978. "A referent-oriented, analytical concept theory for INTERCONCEPT". *Knowledge Organization* 5 no. 3: 142-151 https://www.ergon-verlag.de/isko_ko/downloads/ic_5_1978_3.pdf#page=20.

Dhar, Vasant. 2013. "Data science and prediction". *Communications of the ACM* 56 no. 12: 64-73. https://dl.acm.org/doi/pdf/10.1145/2500499.

Dextre Clarke, Stella G. 2019. "The Information Retrieval Thesaurus". *Knowledge Organization* 46 no. 6: 439-459. https://www.ergon-verlag.de/isko_ko/downloads/ko_46_2019_6_c.pdf.

Dextre Clarke, Stella G. and Zeng, Marcia Lei. 2012. "From ISO 2788 to ISO 25964: The evolution of thesaurus standards towards interoperability and data modelling". *Information Standards Quarterly (ISQ)* 24 no. 1. http://eprints.rclis.org/16818/1/SP_clarke_zeng_isqv24no1.pdf.

Dierickx, Harold and Hopkinson, Alan. 1986. *Reference manual for machine-readable bibliographic descriptions*. http://biblio.cerist.dz/hrbdonf5214/ouvrages/00000000000000594806000000_2.pdf.

FAIR Compliant Biomedical Metadata Templates. 2019. CEDAR, Center for Expanded Annotation and Retrieval, University of Stanford, Department of Medicine. https://medicine.stanford.edu/2019-report/cedar-to-the-rescue.html.

Floridi, Luciano. 2019. "Semantic Conceptions of Information". In *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/win2019/entries/information-semantic/.

Foskett, A. C. (1996). "*The subject approach to information*". Facet Publishing.

Fonseca, Claudenir M., Porello, Daniele, Guizzardi, Giancarlo, Almeida, João Paulo A. and Guarino, Nicola. (2019). Relations in Ontology-Driven Conceptual Modeling. In Laender,

A., Pernici, B., Lim, EP., de Oliveira, J. (eds) Conceptual Modeling. ER 2019. Lecture Notes in Computer Science 11788. Springer, Cham. https://doi.org/10.1007/978-3-030-33223-5_4.

Fillinger, Sven et al. 2019. "Challenges of big data integration in the life sciences." *Analytical and bioanalytical chemistry* 411 no. 26: 6791-6800. doi:10.1007/s00216-019-02074-9

Freitas, C.; Carvalho, P.; Oliveira, H. G.; Mota, C. and Santos, D. 2010. "Second HAREM: advancing the state of the art of named entity recognition in Portuguese". In Nicoletta Calzolari et al. (eds.), Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010). European Language Resources Association, pp. 3630-3637. Valletta, 2010.

Frické, Martin. 2015. "Big Data and Its Epistemology". *Journal of the Association for Information Science and Technology* 66 no. 4: 651-61.

Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." *International journal of information management* 35.2 (2015): 137-144.

Gershenfeld, Nel, Krikorian, Raffi, and Cohen, Danny. 2004. "The Internet of Things". *Scientific American*, October: 76-81. Available: http://cba.mit.edu/docs/papers/04.10.i0.pdf. Accessed May 5 2021.

Giunchiglia, Fausto; Dutta, Biswanath and Maltese, Vincenzo. 2014. "From knowledge organization to knowledge representation". *Knowledge Organization* 41 no. 1: 44-56. http://eprints.biblio.unitn.it/4186/1/techRep027.pdf.

Gray, Jim. 2009. "eScience: A Transformed Scientific Method". In *The FourthParadigm, Data-intensive Scientific Discovery*, ed. Tony Hey, Stewart Tansley andKristin Tolle. Redmond, Wash.: Microsoft Research, 19-33. available at: http://itre.cis.upenn.edu/myl/JimGrayOnE-Science.pdf.

Guarino, Nicola. 1997. "Semantic matching: Formal ontological distinctions for information organization, extraction, and integration". In *International Summer School on Information Extraction*. Springer, Berlin, Heidelberg, 1997. 139-170. https://kask.eti.pg.gda.pl/redmine/projects/sova/repository/revisions/5378040326bc499e118636a1d25ad667285e005c/entry/Praca_dyplomowa/materialy/10.1.1.53.939.pdf.

Guarino, Nicola; Carrara, Massimiliano an Giaretta, Pierdaniele. 1994. "Formalizing ontological commitment". In *AAAI*. 1994. p. 560-567. https://www.aaai.org/Papers/AAAI/1994/AAAI94-085.pdf.

Gruber, Thomas R. 1993. "A translation approach to portable ontology specifications." *Knowledge acquisition* 5 no. 2: 199-220.

Hajibayova, Lala, and Athena Salaba. 2018. "Critical questions for big data approach in knowledge representation and organization." *Challenges and Opportunities for Knowledge Organization in the Digital Age: Proceedings of the Fifteenth International ISKO Conference 9-11 July 2018 Porto, Portugal*, Vol. 16. Ergon Verlag.

He, Yongqun, et al. 2020. "CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis." *Scientific data* 7 no. 1: 1-5.

Hey, Tony; Trefethen, Anne. 2003. "The data deluge: An e-science perspective". In *Grid computing: Making the global infrastructure a reality*, p. 809-824. https://eprints.soton.ac.uk/257648/1/The_Data_Deluge.pdf.

Hjørland, Birger. (2018). "Data (with big data and database semantics)". *Knowledge Organization* 45 no. 8: 685-708.

Hjørland, Birger. (2002). "Domain analysis in information science: eleven approaches– traditional as well as innovative". *Journal of Documentation*, 58 no. 4: 422-462.

Hjørland, Birger. 2013. "Theories of knowledge organization — theories of knowledge.", *Knowledge Organization* 40: 169–181.

Hjørland, Birger, and Albrechtsen, Hanne. (1995). "Toward a new horizon in information science: Domain-analysis". *Journal of the American society for information science 46 no.* 6: 400-425.

Hjørland, Birger and Hartel, Jenna. 2003. "Introduction to a special issue of Knowledge Organization". *Knowledge Organization 30 no. 3/4*: 125-7.

Iafrate, Fernando. (2015). *From Big Data to Smart Data*. London: ISTE Ltd., and Hoboken, NJ: John Wiley & Sons, Inc.

Ibekwe-SanJuan, Fidelia and Geoffrey C. Bowker. 2017. "Implications of Big Data for Knowledge Organization". *Knowledge Organization* 44, no. 3: 187-98.

International Council on Archives. Experts Group on Archival Description. 2019. Records in Context: A Conceptual Model for Archival Description (Consultation Draft v0.1). ICA. https://www.ica.org/sites/default/files/ric-cm-0.2_preview.pdf, accessed December 12, 2018.

International Federation of Library Associations and Institutions (IFLA). 1998. *Study Group on Functional Requirements for Bibliographic Records: Final Report.* UBCIM Publications New Series. München: K. G. Saur.

ISO/IEC 20546:2019(en). 2019. *Information technology — Big data — Overview and vocabulary*. ISO.

ISO 25964-2 (2013). *Information and documentation — Thesauri and interoperability with other vocabularies — Part 2: Interoperability with other vocabularies*. ISO, 2013.

Kahn, Robert; Wilensky, Robert. 2006. "A Framework for Distributed Digital Objects Services". *International Journal on Digital Libraries* 6 no. 2: 115–123. https://www.doi.org/topics/2006_05_02_Kahn_Framework.pdf. Access: June 28, 2022.

Lambe, Patrick. 2014. *Organising knowledge: taxonomies, knowledge and organizational effectiveness*. Elsevier.

Leonelli, Sabina. 2012. "Classificatory Theory in Data-intensive Science: TheCase of Open Biomedical Ontologies". *International Studies in the Philosophy ofScience* 26 no. 1: 47–65.

_____ and Costa, Leonardo C. da. 2016. "A Model to Represent and Process Scientific Knowledge in Biomedical Articles with Semantic Web Technologies". *Knowledge Organization,* 43(2): 122-137. https://www.ergon-verlag.de/isko_ko/downloads/ko_43_2016_2_b.pdf, accessed Apr. 12, 2017.

_____ and Dias, Celia. 2020. "Representing facet classification in SKOS". In International ISKO Conference, Aalborg, Denmark, 16th, *Proceedings… 1. Edition.* Würzburg: Ergon Verlag. *ISBN print: 978-3-95650-775-5, ISBN online: 978-3-*

95650-776-2, Series: Advances in knowledge organization 9. Würzburg: Ergon Verlag, 254–263. https://doi.org/10.5771/9783956507762, acessed Fev. 15, 2021.

_____; Martins, Sergio. C. and Ramos Junior, Mauricio. C. 2021. The role of vocabularies for the access and reuse of Big Data. *Informação & Informação*, 26 no. 4: 146-174. https://www.uel.br/revistas/uel/index.php/informacao/article/view/44653/pdf. Access 5 Jan. 2022.

De Mauro, Andrea; Greco, Marco and GrimaldiI, Michele. 2015. "What is big data? A consensual definition and a review of key research topics". In *AIP conference proceedings*. American Institute of Physics, 2015. p. 97-104. http://big-data-fr.com/wp-content/uploads/2015/02/aip-scitation-what-is-bigdata.pdf.

Mazzocchi, Fulvio. 2018. "Knowledge organization system (KOS)". Knowledge Organization 45, no.1: 54-78. Also available in ISKO Encyclopaedia of Knowledge Organization, eds. Birger Hjørland and Claudio Gnoli, http://www.isko.org/cyclo/kos.

Méndez, Eva; Greenberg, Jane. (2012). "Linked data for open vocabularies and HIVE's global framework". *El profesional de la información* 21 no. 3: 236-244.

Mylopoulos, John. 1992. "Conceptual modelling and Telos". In *Conceptual modelling, databases, and CASE: An integrated view of information system development*, p. 49-68. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.3647&rep=rep1&type=pdf, accessed Dec. 13, 2020.

Ontology Web Language Overview. 2004. W3C. https://www.w3.org/TR/owl-features/, accessed 7 Jan. 2022.

Orilia, Francesco and Paoletti, Michele Paolini. 2020. "Properties", *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/win2020/entries/properties/, accessed 20 Sept. 2021.

Otlet, Paul. (2018). Tratado de Documentação: o livro sobre o livro, teoria e prática. Brasília: Briquet de Lemos Livros.

Peirce, Charles. S. 1869 . "On a new list of categories". *Proceedings of the American Academy of Arts and Sciences, v.* 7, p. 287-298, 1868. Disponível em: < http://www.bocc.ubi.pt/pag/peirce--charles-list-categories.pdf>. Access July 28, 2021.

Poole, Alex H. "Now is the Future Now? 2013. "The Urgency of Digital Curation in the Digital Humanities." *DHQ: Digital Humanities Quarterly* 7 no. 2.

Prasad, A. R. D., Giunchiglia, Fausto; Devika, P. Madalli. 21017. "DERA: from document centric to entity centric knowledge modelling". In: *Proceedings of the International UDC seminar 2017. Faceted classification today*. London: September, 2017. p. 169-179. *http://seminar.udcc.org*.

Prieto-Díaz, Ruben. 1990. "Domain analysis: An introduction". *ACM SIGSOFT Software Engineering Notes*, 15 no. 2: 47-54.

Ranganathan, S. R. and Gopinath, M. A. *Prolegomena to Library Classification*. 3 ed. Bombay: Asia Publishing House, 1967.

RDF semantics. W3C, 2004. http://www.w3.org/TR/rdf-mt/, acessed Mar, 10, 2010.

RDF 1.1. PRIMER. 2014. W3C. https://www.w3.org/TR/rdf11-primer/, accessed 12 Dez. 2019.

Resource Description Framework (RDF) Model and Syntax Specification. W3C, 1998. https://www.w3.org/1998/10/WD-rdf-syntax-19981008/. Accessed May 5, 2011.

Riva, Pat, Le Boeuf, Patrick, and Žumer, Maja. 2017 "*IFLA Library Reference Model: A Conceptual Model for Bibliographic Information"*. IFLA. [online] https://www.ifla.org/publications/node/11412 (Accessed 23 March 2019)

Rowley, Jennifer. 2007. "The wisdom hierarchy: representations of the DIKW hierarchy". *Journal of information science* 33 no. 2: 163-180. http://web.dfc.unibo.it/buzzetti/IUcorso2007-08/mdidattici/rowleydikw.pdf, access Jul 14 2013.

Saracevic, Tefko. 2007. "Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance". *Journal of the american society for information science and technology* 58 no. 13: 1915-1933.

Shet, Amith. 2020. "Knowledge Graphs and their central role in big data processing: Past, Present, and Future". In 7th ACM India Joint Conference on Data Science & management of Data (COD-COMAD), Indian School of Business, Hyderabad Campus, 5-7 January 2020. https://www.slideshare.net/apsheth/knowledge-graphs-and-their-central-role-in-big-data-processing-past-present-and-future, accessed Jun. 5, 2021.

Shet, Amith; Ramakrishnan, Cartic and Thomas, Christopher. 2005. "Semantics for the semantic web: The implicit, the formal and the powerful". *International Journal on Semantic Web and Information Systems (IJSWIS)*, no. 1 vol. 1:1-18. http://www.ebusinessforum.gr/old/content/downloads/JSWIS.pdf#page=19, accessed Jul 14, 2010.

Shiri, Ali. 2013. "Linked data meets big data: A knowledge organization systems perspective." *Advances in Classification Research Online* 24 no. 1: 16-20.

SKOS – Simple Knowledge Organization System Namespace Document. W3C, 2012. https://www.w3.org/2009/08/skos-reference/skos.html#, accessed Aug 10, 1013.

Soergel, Dagobert. 2015. "Unleashing the Power of Data Through Organization: Structure and Connections for Meaning, Learning and Discovery." *Knowledge Organization* 42 no. 6: 401-427.

SPARQL 1.1 QUERY LANGUAGE, 2013. W3C. https://www.w3.org/TR/sparql11-query/, accessed 12 Fev. 2010.

Strecker, Dorothea et al. 2021. *Metadata Schema for the Description of Research Data Repositories*. Re3data, 2021. Available at: https://doi.org/10.48440/re3.010. Access 08 Jul. 2022.

Swanson, Don R. (2008). "Literature-based discovery? The very idea." In *Literature-based discovery*. Springer, Berlin, Heidelberg. 3-11.

Veiga, Viviane Santos de Oliveira; Campos, Maria Luiza; Silva, Carlos Roberto Lyra; Henning, Patricia and Moreira, João. 2021. "Vodan br: a gestão de dados no enfrentamento da pandemia coronavirus". *Páginas A&B, Arquivos e Bibliotecas (Portugal),* n. Especial: 51-58. http://hdl.handle.net/20.500.11959/brapci/157353, accessed Out 7, 2021.

Wilson, Thomas. D. 1972. "The work of the British Classification Research Group". In Wellish, H. (ed). Subject retrieval in the seventies. Westport: Greenword Publishing Co., 62-71.

Zeng, Marcia Lei. 2019. "Interoperability". *Knowledge Organization* 46, no. 2: 122-146. Also available in Hjørland, Birger and Gnoli, Claudio eds. *ISKO Encyclopedia of Knowledge Organization*, http://www.isko.org/cyclo/interoperability.

Zeng, Marcia. L. (2017). "Smart data for digital humanities". *Journal of Data and Information Science* 2 no. 1: 1-12. DOI: 10.1515/jdis-2017-0001.

THE ROLE OF VOCABULARIES IN THE AGE OF DATA: the question of research data

Abstract

Objective: The objective of this work is to discuss how vocabularies, can contribute to assigning computational semantics to digital research data within the context of Big Data, so that computers can process them, allowing their reuse on large scale.

Methodology: A conceptualization of data is developed in an attempt to make it clearer what would be data, as an essential element of the Big Data phenomenon, and in particular, digital research data. It then proceeds to analyse digital research data uses and cases and their relation to semantics and vocabularies.

Results: Data is conceptualized as an artificial, intentional construction that represents a property of an entity within a specific domain and serves as the essential component of Big Data. The concept of semantic expressivity and use it to classify the different vocabularies and within such classification ontologies are shown to be the type of knowledge organization system with a higher degree of semantic expressivity. Features of vocabularies that may be used within the context of the Semantic Web and the Linked Open Data to assign machine-processable semantics to Big Data are suggested. It is shown that semantics may be assigned at different data aggregation levels.

> The ultimate Big Data challenge lies not in the data, but in the metadata— the machine-readable descriptions that provide data about the data. It is not enough to simply put data online; data are not usable until they can be 'explained' in a manner that both humans and computers can process." Researcher Mark Musen Declaration (FAIR Compliant Biomedical Metadata Templates | CEDAR, 2019).

1. Introduction

Big Data has been called the phenomenon describing the huge amount of digital data that is being created at enormous velocity, great heterogeneity as the result of social, economic, scientific and cultural activities centred on the web. Today's research data has also the characteristics of Big Data (Fillinger et al. 2019). Data is created in huge quantities and velocity directly from monitoring devices and projects, like the Hubble Space Telescope, the Human Genome research project, the Large Hadron Collider. Besides the data created directly by scientific activities, Big Data in itself is of interest for scientific research. Shiri (2013, 18) claims that Big Data is made up of research data, open data, linked data and semantic. In today's Web landscape such themes are interwived. Research data is an important product of science, along with scientific publications. How to deal with the "V"s of Big Data, Volume, Velocity, Variety, Variability, Veracity, in research data to enhance its "V"alue and achieve insights of such data (Iafrate 2015, 3)? How can its large-scale reuse be facilitated? Within such a context and considering the statement by the researcher Mark Musen, what can be the contribution of vocabularies, an important research area in Knowledge Organization (KO).

1.1. The Big Data

Big Data, the term for a recent phenomenon describing the amount of data produced in digital format, its explosive growth, and the difficulties of storing, processing, and reusing the data, is increasingly present in information technology media. The headlines also call the phenomenon "information deluge," "data deluge," or "tsunami of data" (Hey and Trefethen 2003). According to these sources, it is impacting business, government, culture, science, and society.

Big Data reminds the so-called "information explosion," a phenomenon connected to the rise of Information Science and KO. In response, KO created knowledge organization systems (KOS) that work in conjunction with information retrieval systems (IRS), computerized databases containing representations of scientific documents. Such KOS, the "information retrieval thesaurus" (Dextre Clarke 2016, 138), control and standardize the natural language used both for indexing the documents entered in the IRS and the keywords used in the user's queries.

Most conceptualizations of Big Data tend to emphasize technological aspects such as volume, variety, velocity, heterogeneity, and the need for massive computer power to process it (Gandomi and Haider 2015). Big Data has also been sparking interest in KO (Ibekwe-SanJuan and Bowker 2017, 192), raising questions like its impact in KO epistemology and methodologies (Hajibayova and Salaba, 2018), (Frické 2015). However contributions from the area to propose practical solutions are still few; Hjørland (2013, 179) stressed: "But such progress is brought to us from the outside; it is not something the field of KO has provided". The availability today of huge datasets recording user interactions with different systems, their interests, and preferences, gave rise to the development of data-driven methodologies to guide the interactions between users and such systems, including IRS, an area of application of KO. Nonetheless, methodologies and tools created on their bases have been developed by private enterprises such as Google, Amazon, Netflix. Hajibayova and Salaba (2018, 147) stress the "opacity of the algorithms behind the platforms and systems".

The best-known product of science, to which the KO has been dedicated since its beginnings, is scientific publications. More recently, science has been giving increasing importance to another of its products, research data. Today, research data, practically entirely digital, is produced in increasing quantities as a result of scientific activity carried out with the support

of information technologies. Examples of this huge amount of digital survey data are those generated by the Hubble Space Telescope, https://www.nasa.gov/mission_pages/hubble/main/index.html, the Human Genome research project, https://www.genome.gov/human-genome-project, or the Large Hadron Collider, https://home.cern/science/accelerators/large-hadron-collider, the largest and most powerful particle accelerator in the world. A large amount of digital research data now available has even raised debates concerning scientific methodology (Gray 2009), (Leonelli 2012), (Frické 2015).

Research data is defined as "factual records (numerical scores, textual records, images, and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings." (OECD, 2007, 13). Share and reuse of research data presupposes its openness but not only that. As quoted by researcher the Mark Musen at the beginning of this work: "the metadata — the machine-readable descriptions that provide data about the data", has been gaining increasing importance. Vocabularies, i.e., data vocabularies or metadata vocabularies (Zeng 2019), is an important research area in KO. Musen's observation refers to the Semantic Web project (Berners_Lee et al 2001), the proposal for a Web whose resources would be represented in a way that had a precise and formal meaning or semantics and would be intelligible and understandable by both people and machines.

1.2. The document centered vision of vocabularies

The technical traditions and standards developed by KO to manage the information explosion rest on assumptions that persist to this day. In most discourses in the area, these assumptions are so implicit that it becomes difficult to make them explicit, consider them, and analyse their consequences. All the theories and methodologies of KO mentioned bringing these

assumptions implicitly: the IRS represent documents in their computerized databases; MARC and the bibliographic formats that emerged from the UNISIST Reference Manual for machine-readable bibliographic descriptions (Dierickx and Hopkinson, 1986) (Reviewer 2) are metadata sets that represent different descriptive properties of the documents.

KOS associated with IRS confirm such assumptions; they "have been designed to support the organization of knowledge and information to make their management and retrieval easier" (Mazzocchi 2018). They are terminological control instruments used to standardize the records' subject and authorities fields in IRS computerized databases, so useful for users' subject-based (Foskett 1996) retrieval.

Representing documents and their subjects is a practice with a long tradition in KO. In the past such documents surrogates were a fundamental mechanism to provide access to information and enable processes of relevance assessment carried out by libraries and IRS (Saracevic 2007). KO methodologies have always represented domains of knowledge when building KOS like controlled/standardized vocabularies, subject headings, and taxonomies KOS, such as thesauri, were intended to enable subject-based retrieval in the context of IRS because their records were representations of objects that have as one of their properties subjects. But not all objects in a domain have subjects as one of their properties like documents. We see now that this is just one among many cases of representing different objects in digital space.

To what extent do these assumptions hold up today, and are they sufficient to address the challenges of the Semantic Web era, Big Data, research data, and the Internet of Things? Today, it is not only the case of retrieving documents (or their representations) but also to create digital representations of anything, as demanded by the "Internet of Things" (IoT) (Gershenfeld, Krikorian, and Cohen 2004). If the documentation movement (Otlet 2018) and

then Information Science empowered information by separating it from books, the Semantic Web proposal and Big Data did the same with the knowledge (Soergel 2015). It is no longer just inserted into texts to be interpreted by humans, but rather serialized in Resource Description Framework (RDF) triples (RDF 1.1 PRIMER 2014), forming representations/descriptions of "things".

The objective of this work is to discuss how vocabularies, in the sense used within LOD Technologies i.e., value vocabularies, or KOS, and metadata vocabularies (Zeng 2019), can contribute to assigning computational semantics to digital research data within the context of Big Data, so that computers can process them, allowing their reuse on large scale. Descriptive metadata sets represent specific entities, or resources in the Web context; value vocabularies assign standardized data values to specific descriptive items of entity instances described by metadata vocabularies.

As a methodology, the work develops a conceptualization of data in an attempt to make it clearer what would be data, as an essential element of the Big Data phenomenon, and in particular, digital research data. It then proceeds to analyse digital research data uses and cases and their relation to semantics and vocabularies.

The work is organized as follows. After this introduction, section 2 analyses data from a semiotic and ontological point of view. Section 3 presents a comprehensive view of vocabularies within the context of Semantic Web and LOD. Within such a context Section 4 develops a conceptualization of data that is illustrated by examples of research data, research datasets, and related initiatives, and shows how research data at different levels of aggregation yields semantics. Section 5 draws conclusions, raises research questions to be developed and presents final considerations.

2. Semiotic and ontological view of data

None of the most common Big Data definitions exclude the data component. It seems reasonable, then, that to understand what Big Data is and how to operationalize solutions to the problem begins by elucidating what is data. After presenting the traditional use of vocabularies to represent and assign subjects to documents this section proposes a semiotic and ontological analysis of data, understood as the essential component of Big Data and research data. This analysis begins with the question of conceptual models and domains and goes on to analyse how conceptual models of domains are expressed linguistically as vocabularies. Then data is discussed from a semiotic and ontological point of view.

2.1. Vocabularies as representations of domains

In the 1980s-1990s, as a consequence of the emergence of online bibliographic catalog management systems and databases, the domain of information retrieval in library catalogues, so familiar to us but also so exclusive, with its diversity of objects, was first modelled using a methodology used in computer science to plan database management systems. The Functional Requirements for Bibliographic Records conceptual model (FRBR) based on Chen (1976) Entity-Relationship (E-R) model, appeared in 1998, whose development was promoted by IFLA (1998).

According to Mylopoulus (1992, 3) "Conceptual modeling is the activity of formally describing some aspects of the physical and social world around us for purposes of understanding and communication." For Mylopoulus,

> the descriptions that arise from conceptual modeling activities are intended to be used by humans, not machines. . . [and] The adequacy of a conceptual modeling notation rests on its contribution to the construction of models of reality that promote a common understanding of that reality among their human users.

A conceptual model sets an agreement between users of a system on what kinds of things exist and will be represented in the system, or entities (also called classes) in a given domain of reality, e.g. documents of historical value, the properties of these entities and how they relate to each other (relationships). Thus, a conceptual model is a representation, in the form of an abstract and generic description, independent of computational implementations (hardware, operating systems, languages, database management systems) of a given domain of reality. It aims at understand this reality, reason about it, and establish a common view of this reality; a conceptual model answers questions such as: What different things exist in a given domain? How are they distinguished from each other? How do they relate? What are their properties?

As a representation, a conceptual model is expressed, communicated, and externalized through a language, or more specifically a meta-language or meta-model (Guizzardi 2007, 23), which is a language to express the vocabulary (concepts, terms) that express things in specific domains. Examples of these meta-languages are either natural language (through a system requirements document), which functions as the most general of all meta-languages, or a diagrammatic meta-language, such as entity-relationship (meta) Model or the Unified Modelling Language (UML), https://www.uml.org/, class diagram, in which domain-specific ER models or class diagrams are expressed.

Within descriptive representation, once established and consolidated practical standards such as MARC, UNISIST, AACR2 and ISDB, the question of what are the "things" represented is raised, a view with a higher level of abstraction of a domain.

Conceptual models in the area of documentation and information have made things like documents, authors, and subjects explicit. They evolved from the previously mentioned standards for creating automated bibliographic records, starting with the pioneering FRBR (Ifla 1998). FRBR, as a conceptual model of the bibliographic domain, is not intended for

describing or indexing documents, but for formalizing, identifying, agreeing, and standardizing objects, actors, and processes and their relationships within such domain.

Universal bibliographic classification systems such as the Dewey Decimal Classification (DCC) – and the Universal Decimal Classification (UDC) are used for thematic representation, for assigning subjects – as discipline names - to books. They model the universe of knowledge as a set of taxonomies, each having as a root a discipline. The use of taxonomies to organize a domain is typically used today for information management within corporations and to organize the content of websites (Lambe 2014). Taxonomies only organize the things in a domain in class-subclass relationships. The things being organized in a universal bibliographic classification are discipline names to be used as subjects to books.

However, there are more than just things or taxonomies of things in a domain. A more accurate model of a domain should include also their properties, relationships and attributes, according to the ER model. The first movement within documentation and information to recognize this fact was Faceted Classification (Ranganathan, Gopinath 1967). Facets are the properties of a class of things of interest for information recovery (Giunchiglia et al 2014; _____ and Dias, 2020). Including properties of things results in a more accurate representation of a domain, a conceptual model, with richer semantic expressiveness (Almeida, Souza and Fonseca 2011) than a taxonomy.

After the pioneering FRBR model (Ifla, 1998), the International Council of Museums (ICOM) adopted the CIDOC Conceptual Reference Model (CIDOC 2014), IFLA released the Library Reference Model (LRM) integrating the FRBR, FRAD, FRSAD models (Riva, Le Boeuf, and Žumer 2017) and more recently the International Council of Archives (ICA) adopted the Records in Context Conceptual Model (Ric-CM) (International Council on Archives 2019). Since the publication of the FRBR model in 1998, KO has been changing its representation

activities and methodologies, from records describing documents and their subjects to conceptual modeling, that is, representing entities, their attributes and relationships (Prasad, Giunchiglia, Devika 2007). Knowledge organization and representation is part of the digital research data curation effort. Such domains of application also uses conceptual models to integrate heterogeneous research data sources as publications, research data, patents, projects, events, funding agencies, etc. (CERIF in Brief 2014)

Conceptual models are aligned together with different types of KOS by Almeida, Souza and Fonseca (2011, 196), ordered according to their semantic expressiveness. Semantic expressiveness can be understood, in the context of the previous quote, as the ability of each type of KOS to distinguish and describe, that is, identify the properties and represent the different things that exist in a domain of that reality.

Conceptual model elements - entities, attributes and relationships - are expressed linguistically by a vocabulary. Vocabularies are semantic control devices, formed by systematised sets of semiotic, triadic entities (PEIRCE 1994), concepts (Dahlberg 1978), units of meaning that relate something (a first: object or referents), in some way (through a second: term or code), which generates or induces a third: its meaning.

2.2. Domains

Aside from the general library classification systems such as the CDD and the CDU, KOS are developed and used concerning specific domains. The domain notion commonly used in KO is that of a specialized knowledge area.

Hjørland and Albrechtsen (1995, 400), in the text in which they propose the analysis of domains as the foundation of KO, define domains as: “thought or discourse communities,

which are parts of society's division of labour." They also label a domain as a "specialty/discipline/domain/environment" (Hjørland and Albrechtsen 1995, 401).

Hjørland (2002, 422) conceptualizes domains associated with specialized libraries, questioning what knowledge would be necessary for information professionals to work in "in a specific subject field like medicine, sociology or music?" In Hjørland and Hartel (2003, 239), this view of domains as systems of thought, theories, is reaffirmed.

> Domains are basically of three kinds of theories and concepts: (1) ontological theories and concepts about the objects of human activity; (2) epistemological theories and concepts about knowledge and the ways to obtain knowledge, implying methodological principles about the ways objects are investigated; and (3) sociological concepts about the groups of people concerned with the objects.

The oldest thesaurus were intended to enable subject-based retrieval in the context of IRS because their records were representations of objects that had subjects as one of their properties, that is, documents. Today, it is not just about retrieving documents (or their representations) but digital representations of anything, as exemplified in the IoT. These representations are no longer just access points for documents, but also information resources themselves, complex descriptions of these objects, and sources of knowledge about them, represented in such a way that they can be processed/intelligible by both machines and humans. Such representations allow machines to make inferences about the knowledge thus represented.

KO today is being called upon to model different domains of knowledge to build new "semantic" vocabularies, i.e, vocabularies compliant with the Semantic Web and LOD technologies. For this, it is necessary to expand the traditional notion of a domain as a

discipline or subject. In the area of software development the notion of a domain has a broader scope: it is 'a sphere of activity or interest: field" [Webster]. In the context of software engineering, it is most often understood as an application area, a field for which software systems are developed (Prieto Diáz 1990, 50).

Since a vocabulary is a terminological system that represents the "things" of interest in a domain of action to the community of agents/users in that domain, then to create a vocabulary (an artifact, similar to software) several aspects and questions must be considered: What things are in a domain? How should they be represented? These are the questions of ontology and semiotics. They must be answered to create a representation, or a conceptual model, of a domain.

A first step is to determine what things exist in a domain and which are relevant to this community, what rules exist about these things or are created/approved/agreed on about these things, and how this community uses them to act in this domain. Finally, how the conceptualizations and their agreed terms (Dahlberg 1978), one of the by-products of this process, are to be systematised in a domain model to serve as bases for the construction of vocabularies such as thesaurus or computational ontologies.

As shown, vocabularies can be representations of domains. A domain vocabulary can be used either to assign subjects to documents: a) (e.g. MeSH categories describing the entities within the Healthcare domain, https://meshb.nlm.nih.gov/treeView, or b) to describe objects in this domain, descriptive metadata standards that, in addition to identify what things exist in a domain, also describe their properties – attributes and relationships. Among the things within a domain some vocabularies focus on specific facets for special purposes: archival science and records management uses functional classification plans in an organization to assign the

organizational provenance or the function or organizational process that generated or used a record.

2.3. Data as Representations

What is Big Data? What is its relationship with data? What is data and how is it related to metadata? How should semantics be assigned to data? As noted in the ISO/IEC 20546/2019 Standard, "The big data paradigm is a rapidly changing field with rapidly changing technologies," later suggesting a definition: "extensive datasets (3.1.11) — primarily in the data (3.1.5) characteristics of volume, variety, velocity, and/or variability — that require a scalable technology for efficient storage, manipulation, management, and analysis."

The conceptualizations of Big Data define it as a phenomenon that involves large amounts of data, the heterogeneity of that data, a continuous flow of generation and updating, and a need for large processing capacity so that the data reveal patterns or trends (De Mauro et al 2015). However, the same is not true for the conceptualizations of data originating from KO. Data is mentioned frequently in the literature, along with its relationships with information and knowledge (Buckland 1991), often called the data, information, knowledge, wisdom (DIKW) hierarchy (Rowley 2007). In Floridi (2019), information is related to data and semantics.

An important exception is from Hjørland (2018), who proposes a conceptualization of Big Data arising from definitions of data, a phenomenon much better known and conceptualized within KO. Data is in the essence of the Big Data phenomenon, it could not exist without data. In this work, Hjørland lists several similar conceptualizations of data and highlights that of Fox and Levitin:

> Within this framework, we define a datum or data item, as a triple <e, a, v>, where e is an entity in a conceptual model, a is an attribute of entity e, and v

is a value from the domain of attribute a. A datum asserts that entity and has value v for attribute a. Data are the members of any collection of data items.

Such conceptualization is clarified by the following example: "2018." What does 2018 mean? Others would say it's a given. Let us note, however, this statement: "Giovana was born in 2018." In it we can identify the entity we are talking about: a child called "Giovana," an attribute or property of this entity, she is "born," and the value of this attribute or property, her birth year, "2018." To achieve a formal representation it is very important to clearly identify the entity being described. Although a data set usually has a title or description identifying the entity it represents that is not always the case. A metadata set may mix metadata elements of different entities as for example the MARC21 format field 245 – Title Statement; while MARC21 format describes a bibliographic entity, e.g., a book, field 245 subfield code $c describes another entity, the responsible for the book, and field 245 subfield $f its attributes birth and death dates.

In the ontological scheme that goes back to Aristotle (2000), the reality is constituted of the first substances, the things that have real existence in space and time, and second substances, the conceptualizations we make of the first substances to think, reason, make sense of, and communicate about the things in reality. Second substances are in turn subdivided into essences, concepts designating things that have properties whose loss implies the non-existence of that individual and have existential independence (Fonseca, Porello, Guizzardi, Almeida, and Guarino 2019, 29), and accidents, concepts that designate things that are existentially dependent on other substances. Things having existential independence are commonly recognized in one of the most well-known ontological schemes, the entity-relationships (ER) model (Chen 1976) as entities, while those that are existentially dependent, as properties. Properties, in turn, are subdivided into attributes of an entity, relationships between an existentially independent entity and the value of one of its properties, and

relationships, involving two or more individuals of the same, or of different existentially independent entities (Orilia and Paoletti 2020).

Classifying concepts in vocabularies as entities and their properties, attributes or relationships is a practice that has become common in the specification of vocabulary compliant with LOD technologies; see, for example, the DC Terms vocabulary, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/, the PROV-O ontology, https://www.w3.org/TR/prov-o/, and DCAT metadata vocabualry, https://www.w3.org/TR/vocab-dcat-3/.

Data is about representations of something else. A data unit, a datum (Hjørland 2018), even in the context of Big Data, then, makes no sense without referencing the entity and one of its properties, the metadata. The three concepts are inseparable and cannot be understood separately. They correspond to a descriptive, representational element of an entity, describing one of its properties. They correspond linguistically to a claim, a basic unit of knowledge to which, according to Aristotle (2000, 39), values of truth or falsity can be attributed.

The statements represented by triples constituted by an entity, one of its properties, and the value of this property correspond to the representation of informational resources in the context of LOD, using the RDF data model (RDF Primer 2014). RDF is a Semantic Web standard for describing resources. Everything that is available on the Web can be accessed through a link, or a Uniform Resource Identifier (URI). Today URI evolved towards IRI, the Internationalised Resource Identifier, which strings incorporate characters from alphabets others than the Latin alphabet. This representational model describes such a resource through triples formed by a subject, the resource being described; a predicate, a property that describes the resource; and an object, the value of this property for this resource. The RDF model

assumes a minimum semantics, that is, three elements with specific roles, the subject, the predicate, and the object that form the triple and appear in this order.

Semiotic and ontological analysis identifies a piece of data as an artificial and intentional artefact that represents something. The foundational types of the things that exist are entities - existentially independent things - and their properties: relationships between two existentially independent individuals, and attributes of an individual, its qualities and quantities. Ontological Analysis of things in a domain, classifying and assigning types to these things makes the terms in a domain vocabulary consistent, as they inherit the ontological nature of their types and enable their representations to be machine-processable.

3. A Comprehensive view of vocabularies

In this section, a comprehensive view of vocabularies based on the previous discussion in section 2 and on contributions by Hjørland (2018) and Zeng (2019) was compiled and developed.

3.1. Vocabularies, Web of Data, Linked Open Data, and Big Data

LOD technologies are an integral part of the Web of Data project. Although this is its best-known name, the project is also known as Web of Data, a name that describes it better, since semantics concerns meanings (Chierchia, 2003), and the ability of the Web of Data to convey meanings is quite limited and different from the sense in our understanding of expressions in natural language.

The project was initially formulated by computer scientist Tim Berners-Lee, the creator of the Web, among others. According to its formulators, the Semantic Web aims to propose "A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities" (Berners-Lee et al 2001). To its authors, "Most of the Web's content today is

designed for humans to read, not for computer programs to manipulate meaningfully." The Semantic Web then "will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users."

The Web of Data then refers to content represented in such a way that it can be understood by both machines and people. The current Web is made up of pages, such as http://www.uff.br, formatted in Hypertext Markup Language (HTML), accessible and interconnected with each other through links. Navigating these pages through these links is done by browsers, such as Internet Explorer, Google Chrome, or Mozilla Firefox. HTML is a content markup language; it formats the content of a text of a page through a predefined set of markups, which instruct browsers to display them on computer screens for human users. The content of HTML pages is interpreted by browsers to make it readable and visually pleasing to people.

The proposed Web of Data is quite different. The Web will no longer be constituted of pages to be read by people, but of content, called informational resources, digital representations of things: concrete, like me, you, an industrial product, a monument, a geographical accident; abstract, like a musical genre, a scientific discipline; or just has a digital existence, such as a photo in a JPG file or a scientific article in a PDF file. These are the entities in the proposal by Hjørland (2018). Each of these resources is uniquely identified by a link, or a URI. A resource, identified/accessed by its URI, is described in a structured way through triples, each one formed by the URI of the resource, by each of its properties, and by the corresponding values of each of these properties. An example of how this representational model works is the Leonardo Da Vinci resource on Wikidata, https://www.wikidata.org/wiki/Q762.

This model of structuring data through the description of resources formed by one or more linguistic claims made up of triples <Subject> <Predicate> <Object> is RDF (RDF Primer,

2004). From an ontological point of view, subject, predicate, and object can be understood as an entity, a property, and the value of this property.

Looking in more detail at structuring a triple; for example,

"The page http://www.uff.br is authored by _____."

Such a claim consists of three elements: the subject, "http://www.uff.br," the predicate, "has as author" and the object, "_____"

The RDF model presupposes a minimum semantics, derived from its corresponding linguistic claim. That is, they are identified and appear in this order: the subject, the predicate and the object of the claim that form the triple (Resource Description Framework (RDF) Model and Syntax Specification 1998). A triple describes a specific piece of data from the resource description (what Hjørland calls a "datum:" a unit of data). Sets of triples with the same subject describe the same resource. Sets of interlinked triples describing a resource form a graph.

SPARQL is the query language that allows users to query sets of RDF triples (SPARQL 1.1 QUERY LANGUAGE 2013), navigating through the graphs formed by them and performing inferences. It is the materialization of the Web of Data proposal of a Web that can be queried as if it were a database.

RDF can be serialized in several formats, such as RDF/XML, N Triples, JSON, or TURTLE (RDF Primer, 2004). Of course, RDF triples coded in these formats are not as human-friendly or as clearly readable as HTML pages when viewed by browsers. But they contain elements that allow browsers to understand these formats and display them in a human-friendly manner, if applicable. The main objective of the resources described in RDF is that they can be processed by machines (including their user-friend visualisation), thus helping to organise, retrieve, and make these resources accessible.

The way to extend these semantics beyond the limits of the RDF model is also to make predicates and/or objects into URI and that these URI refer to concepts of vocabularies with specific semantics. According to RDF Semantics (2004) "There are several aspects of meaning in RDF which are ignored by these semantics; in particular, it treats URI references as simple names, ignoring aspects of meaning encoded in particular URI forms." A URI in the RDF model is just a name, an identifier. The advantage of a URI over a natural language identifier such as the linguistic term "author", is its uniqueness, its validity, since a URI is valid and unique throughout the web space, and its persistence, that is, the commitment of whoever assigns it. a URI to never change it (Berners-Lee 1998).

The previous example can be extended by using URI for the subject, the predicade, and the object of the triple.

<http://www.uff.br> <http://purl.org/dc/elements/1.1/creator> <https://orcid.org/0000-0003-0929-8475>

In this example, the original predicate "author" is replaced by the URI referenced by the "creator" element of the well-known Dublin Core (DC) metadata standard. In its context, dc:creator has specific semantics. It is defined as "An entity responsible for making the resource." The triple's object, the value or content of dc:creator, has been replaced by the Open Researcher and Contributor ID (ORCID), https://orcid.org, of the page's author.

It is with the semantics in specific vocabularies that the limited semantic expressiveness of the RDF model can be expanded. Once specified in elements of a vocabulary, the semantics can be processed by programs. While the features provided in the Web of Data, represented in markup languages such as XML, RDF, HTML, etc. are contents, programs are procedures. Programs only know how to process content, they need to be clearly instructed (programmed)

on what to do with certain content in a certain situation. Specially formatted vocabularies, the LOV (Mendez and Greenberg 2012) used to assign semantics to LOD (Zeng 2019) must clearly define, restrict, and specify the semantics of their concepts. For example, the DC metadata vocabulary clearly defines the semantics of each of its concepts (called elements in the DC initiative); for example, dc:creator, is the creator/author or responsible for a resource, e.g., a digital scientific paper. Furthermore, the dc:creator element has itself, a unique persistent identifier, a link, a URI: http://purl.org/dc/elements/1.1/creator. This persistent identifier, unique throughout the Web space, works as a guarantee of the metadata element semantics, allowing a developer to create a specific program to process this element of the DC vocabulary unambiguously, using the semantics specified and standardize d in the DC vocabulary to the dc:creator element.

3.2. Functionalities for vocabularies to be used within the context of the Web of Data and LOD

Through unique and persistent identifiers, metadata and data vocabularies can be used to assign machine-understandable semantics to predicates and objects in triples RDF. Many old vocabularies are being restructured to be compatible with LOD technologies (Soergel, 2004; Dos Santos Maculan, 2015), such as the UNESCO Thesaurus, http://vocabularies.unesco.org/browser/thesaurus/en/, the FAO Thesaurus, http://aims.fao.org/aos/agrovoc/c_8003.html, the AGROVOC Thesaurus, https://agrovoc.fao.org/browse/agrovoc/en/, the Paul Getty Foundation Vocabularies, https://www.getty.edu/research/tools/vocabularies/lod/, the Art and Architecture Thesaurus, the Union List of Artists Names, the Cultural Objects Name Authority, the Getty Thesaurus of Geographic Names, the DeCS/MeSH, Health Science Descriptors,

https://decs.bvsalud.org/ths/, the Library of Congress Subject Headings (LCSH), https://id.loc.gov/authorities/subjects.html, in addition to many others.

Vocabularies used with LOD need to meet requirements such as having their concepts persistently and uniquely identified through valid URIs on the internet, being represented in machine-readable formats such as RDF, containing precise definitions of the semantics of their concepts, and generally, being multilingual. Many of these vocabularies that meet the principles of LOD can be found in the aforementioned LOV vocabulary registry service. By meeting the requirements for use with LOD as described above, vocabularies, an area of study, research, and practical use of KO, can contribute to addressing the issues brought about by Big Data.

Elements of data or metadata vocabularies referenced by URI account for the semantics of an individual "datum" (Hjørland 2018), an element of a triple. These vocabularies use different approaches to semantics, as pointed out in Almeida et al (2011, 195), ranging from semantics for humans, which is implicit, informal or formal, to semantics for machines, which is informal, formal, or even "powerful semantics" (Shet, 2020). In any case, used in the context of the RDF model these vocabularies allow the processing of RDF triples by machines.

3.3. Ontologies as domain models

Since 1993 Gruber (1993, 199) coined a definition of ontology which is used until nowadays as "An ontology an explicit specification of a conceptualization". Borst (1997, 12) developed Gruber's definition as "Ontologies are defined as a formal specification of a shared conceptualization". Two concepts in this last definition are of importance to the present

discussion, - formal, i.e. computers' readable, and – shared, i.e., agreed by a community of agents, being them humans or computers.

The language specification OWL – Ontology Web Language Overview (2004) states that:

> OWL can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms. This representation of terms and their interrelationships is called an ontology. OWL has more facilities for expressing meaning and semantics than XML, RDF, and RDF-S, and thus OWL goes beyond these languages in its ability to represent machine interpretable content on the Web.

OWL is a standard language (meta-language in the aforementioned sense) of the W3C for representing ontologies, that is, vocabularies that specify the things existing in a domain and their interrelationships. Further on, the same specification compares the semantic expressiveness of OWL with that of other languages to represent machine-interpretable content such as XML, XML Schema, RDF, and RDFS (ONTOLOGY WEB LANGUAGE OVERVIEW, 2004). It can thus be concluded that, with current technologies, a computational ontology developed in OWL is the most expressive type of KOS, because the "facilities" provided by OWL allow restricting, specifying, and expressing the intended meaning (Guarino 1994, 560) of the conceptual model of a domain.

Each concept of an ontology vocabulary is typed; it is a class, or a property of a class or an instance, an individual of a class. Among these facilities are the possibility of specifying data properties (attributes, in Chen's ER model), object properties (relationships in Chen's ER model), domain and scope of the two types of properties, and cardinality constraints of each class involved in an object property, transitivity and reflexivity of properties, the disjunction between individuals of different classes, axioms for restricting the inclusion of instances in a

class (ONTOLOGY WEB LANGUAGE OVERVIEW 2004), etc. These facilities can make conceptual models implicit in a computational OWL ontology more faithful to reality. Ontologies also do not distinguish thematic versus descriptive representation; every concept is described by its properties, whether thematic or descriptive.

As seen earlier, the Web of Data project, the large-scale reuse of Big Data and research data available in increasing amounts on the Web, depends on the one hand on the most expressive vocabularies that describe them, and on the other hand, on programs capable of making inferences, or at least algorithmic processing, on these representations. In this context, specific domain models, intelligible by machines and represented with the maximum possible semantic expressiveness such as computational ontologies gain importance.

Another important aspect related to this issue; Bergman (2011) discusses ODapps: The Ontology-Driven Application Approach, an automatic program development methodology based heavily on ontologies, a set of them, from high-level ontologies, task ontologies, domain ontologies, to specific application ontologies (Guarino 1997, 145). In the context of ODApps, domain computational ontologies, with a high degree of semantic expressiveness, are an essential component for developing generic application programs, capable of processing, making inferences, discovering, and reusing the knowledge contained in the domain representation. It is therefore necessary to advance in the creation of domain-specific computational ontologies domains that are increasingly semantically expressive to equip programs capable of processing these representations to make inferences about them and extract and reuse the knowledge contained therein.

4. Results

In the sequel the previous conceptualizations are applied to cases of research data and discussed.

4.1.Data, Big Data, research data

A concrete and dramatic example of the importance of research data and the adoption of principles and technologies that allow its wide dissemination and reuse is the form for collecting data from patients infected with COVID-19, the CRF – Case Report Form, proposed by the WHO. The GO FAIR initiative, https://www.go-fair.org/, addresses the WHO proposal by creating a worldwide network of catalogs referencing research data collected through the CRF and deposited in repositories and available according to the FAIR principles, https://www.go-fair.org/fair-principles/, the "FAIR Data Points." Brazil participates in this initiative through the VODAN-Br Virus Outbreak Data Network initiative (Veiga et al 2021).

The VODAN initiative is expected to collect huge datasets worldwide. The CRF standardized a set of fields of interest to COVID-19 epidemic research. Such fields must be filled with metadata and data associated with vocabularies largely agreed and standardized within the health sciences domain. This allows the interoperability of different datasets and their processing by computers in order to drawing conclusions and insights from the data. VODAN and FAIR Data Points are efforts to provide smart data (Kobielus 2016) to be used to control COVID-19 outbreak.

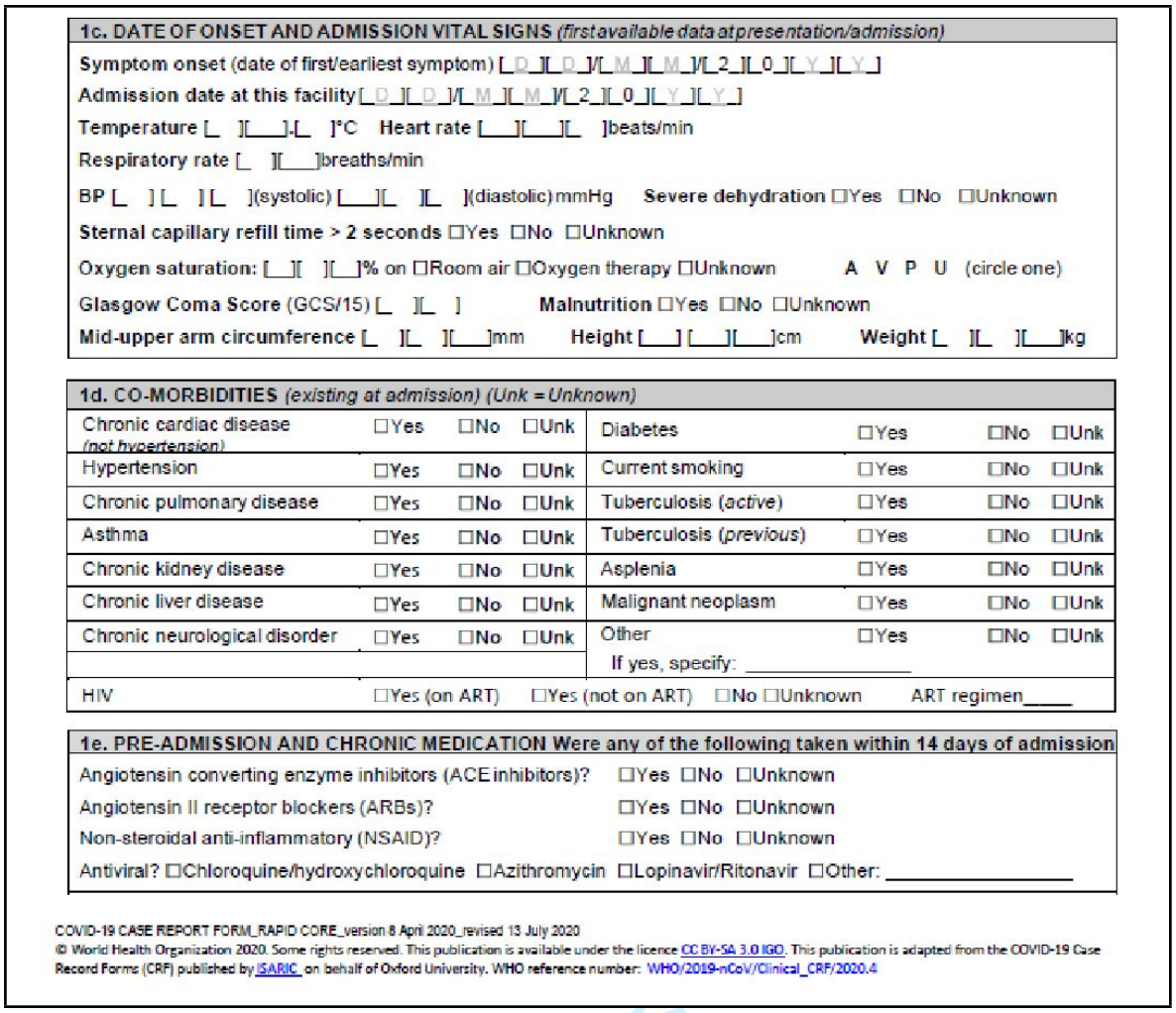
Within the RDF model, the subject, predicate, and object of a triple can be identified by a URI. These URIs identify specific terms, both from metadata vocabularies—descriptive properties of things in a domain—and data vocabularies—values assumed by these properties for specific descriptive metadata.

Another important feature of using vocabularies with LOD technologies is that different vocabularies can be used simultaneously in the form fields. Figure 1 shows an excerpt from the CRF, the co-morbidity data, "CO-MORBIDITIES," of a patient (the entity); they are recorded as follows: concepts such as chronic cardiac disease (the attribute or metadata, the

co-morbidity presented by the patient) are taken from specific biomedical ontologies or vocabularies that describe specific co-morbidity types; if a specific one applies, it is recorded as data as follows: Yes, No, Unk. These data have to be processed by programs so that the immense amount of records collected through the CRF around the world can serve as inputs for the planning and control of the pandemic. The question about co-morbidities has several answer options, each of which indicates a type of disease. For it to be processed by machines, each type of co-morbidity expressed in natural language must reference a concept in a vocabulary or ontology, such as SNOMED-CT, https://www.nlm.nih.gov/healthit/snomedct/index.html. Another question on the CRF, such as the one related to "PRE-ADMISSION AND CHRONIC MEDICATION," has as one of its answer options "Angiotensin converting enzyme inhibitors (ACE inhibitors)?", which may be referenced in another vocabulary such as MeSH, https://meshb.nlm.nih.gov/search, the term with identifier http://id.nlm.nih.gov/mesh/D000806.

In order to have precise meaning, concepts such as those shown in the CRF must refer to specific, standardized ontologies or biomedical vocabularies to enable the processing of these data.

FIGURE 1 - Part of the CRF Form



1c. DATE OF ONSET AND ADMISSION VITAL SIGNS *(first available data at presentation/admission)*

Symptom onset (date of first/earliest symptom) [_D_][_D_]/[_M_][_M_]/[_2_][_0_][_Y_][_Y_]

Admission date at this facility [_D_][_D_]/[_M_][_M_]/[_2_][_0_][_Y_][_Y_]

Temperature [_][___].[_]°C Heart rate [___][___][_]beats/min

Respiratory rate [_][___]breaths/min

BP [_] [_] [_](systolic) [___][_][_](diastolic) mmHg Severe dehydration ☐Yes ☐No ☐Unknown

Sternal capillary refill time > 2 seconds ☐Yes ☐No ☐Unknown

Oxygen saturation: [__][ ][__]% on ☐Room air ☐Oxygen therapy ☐Unknown A V P U (circle one)

Glasgow Coma Score (GCS/15) [_][_] Malnutrition ☐Yes ☐No ☐Unknown

Mid-upper arm circumference [_][_][___]mm Height [___] [___][___]cm Weight [_][_][___]kg

1d. CO-MORBIDITIES *(existing at admission) (Unk = Unknown)*

Chronic cardiac disease *(not hypertension)*	☐Yes	☐No	☐Unk	Diabetes	☐Yes	☐No	☐Unk
Hypertension	☐Yes	☐No	☐Unk	Current smoking	☐Yes	☐No	☐Unk
Chronic pulmonary disease	☐Yes	☐No	☐Unk	Tuberculosis (*active*)	☐Yes	☐No	☐Unk
Asthma	☐Yes	☐No	☐Unk	Tuberculosis (*previous*)	☐Yes	☐No	☐Unk
Chronic kidney disease	☐Yes	☐No	☐Unk	Asplenia	☐Yes	☐No	☐Unk
Chronic liver disease	☐Yes	☐No	☐Unk	Malignant neoplasm	☐Yes	☐No	☐Unk
Chronic neurological disorder	☐Yes	☐No	☐Unk	Other	☐Yes	☐No	☐Unk
				If yes, specify: _____			
HIV	☐Yes (on ART)	☐Yes (not on ART)	☐No ☐Unknown	ART regimen_____			

1e. PRE-ADMISSION AND CHRONIC MEDICATION Were any of the following taken within 14 days of admission

Angiotensin converting enzyme inhibitors (ACE inhibitors)? ☐Yes ☐No ☐Unknown

Angiotensin II receptor blockers (ARBs)? ☐Yes ☐No ☐Unknown

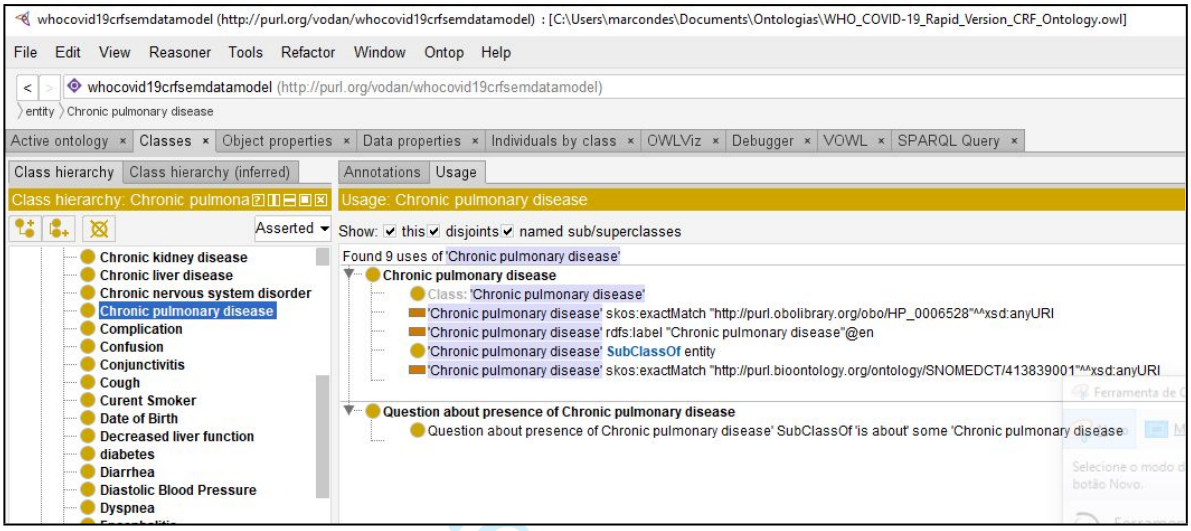
Non-steroidal anti-inflammatory (NSAID)? ☐Yes ☐No ☐Unknown

Antiviral? ☐Chloroquine/hydroxychloroquine ☐Azithromycin ☐Lopinavir/Ritonavir ☐Other: _____

COVID-19 CASE REPORT FORM_RAPID CORE_version 8 April 2020_revised 13 July 2020

The CRF is formalized by a conceptual model and owl ontology, the WHO COVID-19 Rapid Version CRF semantic data model, https://bioportal.bioontology.org/ontologies/COVIDCRFRAPID. In the following Figure 2 another feature of KOS methodologies and standards incorporated in ontologies is the mapping properties. Mapping properties of a concept in a KOS identify which concept in that KOS means the same as another concept from another KOS, i.e., the mapping one concept to another concept. The concept "chronic pulmonary disease" at Figure 1 is shown in Figure 2 as a class of the WHO COVID-19 Rapid Version CRF semantic data model; it is also shown its skos:exactMatch to the SNOMED concept "413839001".

FIGURE 2 – The class “chronic pulmonary disease” of the WHO COVID-19 Rapid Version CRF semantic data model and its SKOS mapping to the SNOMED concept.



Each field in the CRF gives rise to a RDF triple in which the PARTICIPANT ID, the patient, is the subject, the field (standardized and referenced by a metadata vocabulary) is the predicate and its value (also standardized and referenced by a value vocabulary) is the object.

As previously stated, openness is essential to enable research data sharing and reuse. For data to be considered open, international recommendations rate it from 1 to 5 stars, https://5stardata.info/en/. The fourth and fifth stars are awarded when data is available in RDF format, including be accessible through a URI, their predicates and objects be referred by standardized vocabularies widely recognized by the community in a given domain, and linked together to provide rich context. For research data, which has demanded increasing attention and public policies at national and international levels, the international GO FAIR initiative recommends a set of principles for publication so that they have the attributes of FAIR: findability, accessibility, interoperability, and reuse.

The FAIR principles allow research data to be processed by machines. The M4M principle—metadata for machines—states that "There is no FAIR data without machine-actionable metadata. The overall goal of Metadata for Machines workshops (M4M) is to make routine use of machine-actionable metadata in a broad range of fields." The CRF described above is an example of the importance of research data standardization and the adoption of principles that allow its wide dissemination and reuse.

Applying the FAIR principles to research data causes data to be represented as RDF triples. Such a process is named "FAIRification", see https://www.go-fair.org/fair-principles/fairification-process/. FAIR compliant data is generally derived data from datasets. A distributed network of FAIR Data Points provides access to different FAIR data. That raises the question of using vocabularies to describe both the original datasets and their FAIR compliant datasets versions generated.

Other vocabularies also have emerged, not to describe or provide standardized values for each piece of data, but to provide descriptive and value metadata of the datasets as a whole. Digital curation of research data is an emerging field of activity for KO professionals; one of its activities is to apply metadata to research datasets, see https://www.dcc.ac.uk/. For the curation of these datasets, metadata standards such as Data Catalog Vocabulary (DCAT) https://www.w3.org/TR/vocab-dcat-2/, or the Provenance Ontology (PROV-O) https://www.w3.org/TR/prov-o/, have been adopted to describe the provenance of the dataset. As datasets have been made available as informational resources on the Web, information on their provenance and the record of the processing carried out on them, the extract, transform, load (ETL), see https://en.wikipedia.org/wiki/Extract,_transform,_load, and the FAIRrification processes of such data, are essential elements for research data reliability to enable sharing and reuse.

The amount of research data being available every day on Coronavirus epidemic – the "V"ariety" of Big Data - makes the integration of such sources essential to control the epidemic. The Coronavirus Infectious Disease Ontology (CIDO) (He et al. 2020) stresses the essential role computational ontologies in the integration of different and heterogeneous research data sources, promoting interoperability between such sources.

These datasets, in addition to the metadata that describe their fields, are themselves of interest for the research data exploration. They need additional metadata as the type of licence under which data can be reused, the dataset creator, its publisher, its format, its update date, etc, all of which are metadata for the dataset as a whole. They contain metadata such as the format of the dataset, the number of records, the last update date, licences to use this dataset, etc. (from DCAT), or metadata such as the agent that created the dataset, and the process that generated it (from PROV-O). Standards such as these have been used in several research data repositories to index the datasets deposited there. Indeed, digital curation is an increasingly common application by KO professionals (Poole 2013).

Digital Humanities is another growing area of application of digital research data. It grew from the wide availability of data from social activities (search and social media activity every minute, see https://www.smartinsights.com/internet-marketing-statistics/happens-online-60-seconds/) and culture, including science. Scientific articles have long been recognized as a privilege knowledge source (Swanson, 2008), see PubMed Citations per year, https://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html). Significant examples of research projects in Digital Humanities using a variety of such sources can be found in the Digging into Data Challenge program (https:// diggingintodata.org/) mentioned by Zeng

(2017); in this article, the author describes in details how Digital Humanities is related to Big Data and the challenges to process such data and turn it into Smart Data.

A huge amount of such data is textual, resulting from posts on social media, emails, newspaper articles, scientific articles, and text in encyclopaedias such as Wikipedia, among others. This data is unstructured or semi-structured.

The exploitation of such potential information sources may lay on the development of vocabularies for special purposes. Their processing using techniques such as information extraction, named-entity recognition, natural language processing, text mining, machine learning, text annotation, aim at transforming such non-structured or semistructured textual data into structured.

Examples of such techniques in biomedical sciences are the National Library of Medicine Natural Language Processing tools, https://lhncbc.nlm.nih.gov/LHC-research/nlp.html, which lay on dictionaries and KOS like MeSH, the Medical Subject Headings, and UMLS, the Unified Medical Language System (Bodenreider 2004), (Aronson and Lang 2010).

4.2. Semantics beyond the data.

Semantics is a very general concept. An operational concept of semantics applied to messages – data: in digital environment is the inference made by an agent based on a message that enables such agent to make decisions and, possibly, to act accordingly.

The concept of "powerful semantics," originally devised by Shet, Ramakrishnan, and Thomas (2005) and developed in Shet (2020, slide 42), is defined as "statistical analysis [that] allows the exploration of relationships that are not stated." Semantics may be obtained from statistical

patterns, not from individual datum referenced by metadata describing an entity, but rather from data sets as a whole, or Big Data. To identify this semantics, Big Data, whether structured or unstructured, has to be processed by programs. This is so-called data science (Dhar 2013).

Entities are the units to be represented by digital metadata and data within a domain, even if an entity is represented by only one of its properties. As so they are the units of meaning and correspond to what has been called a digital object. The concept of a digital object was first proposed in 1995 by Kahn and Wilensky (2006) as a set of bits that has a special interest in applications or software agents; it is related to the concept of data as a representation of an entity or phenomenon (Hjørland 2018). Digital objects of interest to research data are also just now (see https://www.fdo2022.org/) being conceptualized by initiatives such as FAIR Digital Object Framework: "In the FDOF, a digital object is a bit sequence located in a digital memory or storage that has, on its own, an informational value, i.e., the bit sequence represents an informational unit such as a document, a dataset, a photo, a service, etc", see https://fairdigitalobjectframework.org/.

Within the Web of Data context vocabularies are meaning control and standardization artefacts aimed at making knowledge records meaningful. The previous discussion poses the question of levels of meaning related to levels of data aggregation. Table 1 sketch the relationships between data aggregation levels to digital units of meaning.

DATA AGGREGATION LEVELS	DIGITAL UNITS OF MEANING
Level 1 - a datum (Hjørland 2018), the basic element of data	the value of a database field, the content or an excel cell

Level 2 - a proposition, state of affairs (JANSEN, 2008, 188), Hjørland (2018) (e, a, v) citing Redman, Fox and Levitin (2017, 1173)an RDF triple, a field and its content of a specific row in a database.	a proposition, state of affairs (JANSEN, 2008, 188), Hjørland (2018) (e, a, v) citing Redman, Fox and Levitin (2017, 1173), a RDF triple of an entity, a metadata, and a datum, a field and its content of a specific row in a database, an ontology instance property value, a XML leaf <a>hghghsag</<a>
Level 3 - A data structure, a conceptualization, a message (CAPURRO, 2000) a row in a specific database table, a digital object, a named graph	a row in a specific database table, a digital object, a named graph A data structure, a conceptualization, a message (CAPURRO, 2000)
Level 4 - Several descriptions of different entities, a graph, a conceptualization based on a specific conceptual model a dataset, a database, an ontology populated with its instances	Several descriptions of different entities, a graph, a conceptualization based on a specific conceptual model, a dataset, a database, an ontology populated with its instances, data mining on a specific dataset, an insight from processing a dataset (Dhar, 2013).
Level 5 - Several conceptualizations, several conceptual models. In such cases an ontology with the aid of the mapping properties specified in SKOS model (SKOS 2012) and in ISO 25964-2 Thesauri standard (ISO 25964-2 2013) may holds the agreed semantics that enable the integration and interoperability between such different and heterogeneous research data sources. A research data	A research data repository as re3data, https://www.re3data.org/, described by a metadata vocabulary (Strecker et al, 2021), several heterogeneous datasets of interest for a theme or problem.Several conceptualizations, several conceptual models. In such cases an ontology with the aid of the mapping properties specified in SKOS model (SKOS 2012) and in ISO 25964-2 Thesauri standard (ISO 25964-2 2013) may holds the agreed

repository as re3data, https://www.re3data.org/, described by a metadata vocabulary (Strecker et al, 2021), several heterogeneous datasets of interest for a theme or problem.	semantics that enable the integration and interoperability between such different and heterogeneous research data sources.

5. Final considerations

Issues involving information technologies are obscured by the metaphorical denominations often adopted that, didactically and scientifically, make it difficult to understand and operate them, such as Big Data and the Web of Data. For an accurate understanding of current information technologies, the semantic capacity of computers has to be analysed, understood, and the real potential identified.

The Web of Data technologies bring a significant advance by incorporating more semantic expressiveness and program independence to data published on the Web. Big Data and research data also poses several issues related to the semantic of data. This article sought to demonstrate that data, which have a semiotic and ontological character and are artificial and intentional representations, cannot be understood apart from the entity to which they refer and from the metadata—the properties of this entity—that describe it.

As stressed by Ibekwe-SanJuan and Bowker (2017, 187) “In essence, Big Data will not remove the need for humanly-constructed KOSs”. This article suggests some paths towards the role of vocabularies in addressing the issues raised by research data in the age of Big Data. Web environment, Big Data, and research data together comprise a heterogeneous environment that poses the challenge of making different resources work together. Semantic interoperability is the key to achieve such goal. KOS as conceptual models and ontologies

play a central role in the semantic integration of different and heterogeneous research data sources, promoting interoperability between such sources. In practical terms ontologies hold representation of a domain while mapping properties (SKOS 2012), (ISO 25964-2 2013) and also OWL property "sameAs" (Ontology Web Language Overview (2004) enable the mapping of concepts in a data resource to concepts in another.

It is necessary also to distinguish one piece of datum as referred to by Hjørland (2018), a unit that represents the value of one (of the) properties of an entity, from a record, a set of several datum describing different properties of an entity, from datasets, representing the various entities and their properties, and from databases, bringing together different datasets representing different interrelated entities. Such are different data aggregation levels, having higher levels of semantics in the computational environment. Vocabularies can play an important role in addressing semantics to data at those different levels of aggregation.

Acknowledgments: This work was carried out with the support of the Brazilian agencies CAPES - Financing Code 001, and CNPq, grant number 305253/2017-4. We are also grateful to the anonymous reviewers of this work for their suggestions to improve this text.

References

Almeida, Mauricio; Souza, Renato and Fonseca, Fred. 2011. "Semantics in the Semantic Web: A Critical Evaluation". *Knowledge Organization* 38 no. 3: 187-203. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1041.7976&rep=rep1&type=pdf, accessed 25 Mar 2021.

Aristóteles. 1995. *Categorias*. Porto: Porto Editora Ltda.

Aronson, Alan R., and François-Michel Lang. 2010. "An overview of MetaMap: historical perspective and recent advances." *Journal of the American Medical Informatics Association* 17 no. 3: 229-236.

4- Bergman, Mike. 2011. "Ontology-Driven Apps Using Generic Applications". *AI3 blog*. https://www.mkbergman.com/948/ontology-driven-apps-using-generic-applications/.

Berners-Lee, Tim. 1998. "Cool URIs don't change". https://www.w3.org/Provider/Style/URI.

Bodenreider, Olivier. 2004. "The unified medical language system (UMLS): integrating biomedical terminology." *Nucleic acids research* 32 no. suppl_1: D267-D270.

Borst, Willem N. 1997. *Construction of Engineering ontologies*. Centre for Telematica and Information Technology. University of Twenty, Enschede, The Nederalands.

Capurro, R. 2000. "Angeletics—A message theory". In H.H. Diebner & L. Ramsay (Eds.), *Hierarchies of communication*. Karlsruhe, Germany: ZKM. http://www.capurro.de/angeletics_zkm.html.

CERIF in Brief. 2014. https://eurocris.org/eurocris_archive/cerifsupport.org/cerif-in-brief/index.html

Chen, Peter Pin-Shan. 1976. "The Entity-Relationship Model-Toward a Unified View of Data". *ACM Transactions on Database Systems* 1 no.1: 9-36.

Chierchia, Gennaro. 2003. *Semântica*. São Paulo, Ed. UNICAMP.

CIDOC Conceptual Reference Model Version 5.1.12. 2014. ICOM/CIDOC. http://www.cidoc-crm.org/Version/version-5.1.2.

Dahlberg, Ingetraut. 1978. "A referent-oriented, analytical concept theory for INTERCONCEPT". *Knowledge Organization* 5 no. 3: 142-151. https://www.ergon-verlag.de/isko_ko/downloads/ic_5_1978_3.pdf#page=20.

Dhar, Vasant. 2013. "Data science and prediction". *Communications of the ACM* 56 no. 12: 64-73. https://dl.acm.org/doi/pdf/10.1145/2500499.

Dextre Clarke, Stella G. 2019. "The Information Retrieval Thesaurus". *Knowledge Organization* 46 no. 6: 439-459. https://www.ergon-verlag.de/isko_ko/downloads/ko_46_2019_6_c.pdf.

Dextre Clarke, Stella G. and Zeng, Marcia Lei. 2012. "From ISO 2788 to ISO 25964: The evolution of thesaurus standards towards interoperability and data modelling". *Information Standards Quarterly (ISQ)* 24 no. 1. http://eprints.rclis.org/16818/1/SP_clarke_zeng_isqv24no1.pdf.

Dierickx, Harold and Hopkinson, Alan. 1986. *Reference manual for machine-readable bibliographic descriptions*. http://biblio.cerist.dz/hrbdonf5214/ouvrages/00000000000000594806000000_2.pdf.

FAIR Compliant Biomedical Metadata Templates. 2019. CEDAR, Center for Expanded Annotation and Retrieval, University of Stanford, Department of Medicine. https://medicine.stanford.edu/2019-report/cedar-to-the-rescue.html.

Floridi, Luciano. 2019. "Semantic Conceptions of Information". In *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/win2019/entries/information-semantic/.

Foskett, A. C. 1996. "*The subject approach to information"*. Facet Publishing.

Fonseca, Claudenir M., Porello, Daniele, Guizzardi, Giancarlo, Almeida, João Paulo A. and Guarino, Nicola. 2019. "Relations in Ontology-Driven Conceptual Modeling". In Laender, A., Pernici, B., Lim, EP., de Oliveira, J. (eds) *Conceptual Modeling*. ER 2019. Lecture Notes in Computer Science 11788. Springer, Cham. https://doi.org/10.1007/978-3-030-33223-5_4.

Fillinger, Sven et al. 2019. "Challenges of big data integration in the life sciences." *Analytical and bioanalytical chemistry* 411 no. 26: 6791-6800. doi:10.1007/s00216-019-02074-9

Freitas, C.; Carvalho, P.; Oliveira, H. G.; Mota, C. and Santos, D. 2010. "Second HAREM: advancing the state of the art of named entity recognition in Portuguese". In Nicoletta Calzolari et al. (eds.), Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010). European Language Resources Association, Valletta, 2010. pp. 3630-3637.

Frické, Martin. 2015. "Big Data and Its Epistemology". *Journal of the Association for Information Science and Technology* 66 no. 4: 651-61.

Gandomi, Amir, and Murtaza Haider. 2015. "Beyond the hype: Big data concepts, methods, and analytics." *International journal of information management* 35 no.2: 137-144.

Gershenfeld, Nel, Krikorian, Raffi, and Cohen, Danny. 2004. "The Internet of Things". *Scientific American*, October: 76-81. http://cba.mit.edu/docs/papers/04.10.i0.pdf.

Giunchiglia, Fausto; Dutta, Biswanath and Maltese, Vincenzo. 2014. "From knowledge organization to knowledge representation". *Knowledge Organization* 41 no. 1: 44-56. http://eprints.biblio.unitn.it/4186/1/techRep027.pdf.

Gray, Jim. 2009. "eScience: A Transformed Scientific Method". In *The FourthParadigm, Data-intensive Scientific Discovery*, ed. Tony Hey, Stewart Tansley and Kristin Tolle. Redmond, Washington, Microsoft Research, 19-33. http://itre.cis.upenn.edu/myl/JimGrayOnE-Science.pdf.

Guarino, Nicola. 1997. "Semantic matching: Formal ontological distinctions for information organization, extraction, and integration". In *International Summer School on Information Extraction*. Springer, Berlin, Heidelberg, 1997. 139-170. https://kask.eti.pg.gda.pl/redmine/projects/sova/repository/revisions/5378040326bc499e118636a1d25ad667285e005c/entry/Praca_dyplomowa/materialy/10.1.1.53.939.pdf.

Guarino, Nicola; Carrara, Massimiliano an Giaretta, Pierdaniele. 1994. "Formalizing ontological commitment". In *AAAI*. 1994 : 560-567. https://www.aaai.org/Papers/AAAI/1994/AAAI94-085.pdf.

Gruber, Thomas R. 1993. "A translation approach to portable ontology specifications." *Knowledge acquisition* 5 no. 2: 199-220.

Hajibayova, Lala, and Athena Salaba. 2018. "Critical questions for big data approach in knowledge representation and organization." *Challenges and Opportunities for Knowledge Organization in the Digital Age: Proceedings of the Fifteenth International ISKO Conference 9-11 July 2018 Porto, Portugal*, Vol. 16. Ergon Verlag.

He, Yongqun, et al. 2020. "CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis." *Scientific data* 7 no. 1: 1-5.

Hey, Tony; Trefethen, Anne. 2003. "The data deluge: An e-science perspective". In *Grid computing: Making the global infrastructure a reality*, p. 809-824. https://eprints.soton.ac.uk/257648/1/The_Data_Deluge.pdf.

Hjørland, Birger. 2018. "Data (with big data and database semantics)". *Knowledge Organization* 45 no. 8: 685-708.

Hjørland, Birger. 2002. "Domain analysis in information science: eleven approaches– traditional as well as innovative". *Journal of Documentation*, 58 no. 4: 422-462.

Hjørland, Birger. 2013. "Theories of knowledge organization — theories of knowledge.", *Knowledge Organization* 40: 169−181.

Hjørland, Birger, and Albrechtsen, Hanne. 1995. "Toward a new horizon in information science: Domain-analysis". *Journal of the American Society for Information Science 46 no.* 6: 400-425.

Hjørland, Birger and Hartel, Jenna. 2003. "Introduction to a special issue of Knowledge Organization". *Knowledge Organization 30 no. 3/4*: 125-7.

Iafrate, Fernando. 2015. *From Big Data to Smart Data*. London: ISTE Ltd., and Hoboken, NJ: John Wiley & Sons, Inc.

Ibekwe-SanJuan, Fidelia and Geoffrey C. Bowker. 2017. "Implications of Big Data for Knowledge Organization". *Knowledge Organization* 44, no. 3: 187-98.

International Council on Archives. Experts Group on Archival Description. 2019. Records in Context: A Conceptual Model for Archival Description (Consultation Draft v0.1). ICA. https://www.ica.org/sites/default/files/ric-cm-0.2_preview.pdf.

International Federation of Library Associations and Institutions (IFLA). 1998. *Study Group on Functional Requirements for Bibliographic Records: Final Report.* UBCIM Publications New Series. München: K. G. Saur.

ISO 25964-2. 2013. *Information and documentation — Thesauri and interoperability with other vocabularies — Part 2: Interoperability with other vocabularies*. ISO, 2013.

ISO/IEC 20546:2019(en). 2019. *Information technology — Big data — Overview and vocabulary*. ISO.

Kahn, Robert; Wilensky, Robert. 2006. "A Framework for Distributed Digital Objects Services". *International Journal on Digital Libraries* 6 no. 2: 115–123. https://www.doi.org/topics/2006_05_02_Kahn_Framework.pdf.

Lambe, Patrick. 2014. *Organising knowledge: taxonomies, knowledge and organizational effectiveness*. Elsevier.

Leonelli, Sabina. 2012. "Classificatory Theory in Data-intensive Science: TheCase of Open Biomedical Ontologies". *International Studies in the Philosophy ofScience* 26 no. 1: 47–65.

_____ and Dias, Celia. 2020. "Representing facet classification in SKOS". In International ISKO Conference, Aalborg, Denmark, 16th, *Proceedings… 1. Edition.* Würzburg: Ergon Verlag. *ISBN print: 978-3-95650-775-5, ISBN online: 978-3-*

95650-776-2, Series: Advances in knowledge organization 9. Würzburg: Ergon Verlag, 254–263. https://doi.org/10.5771/9783956507762.

De Mauro, Andrea; Greco, Marco and GrimaldiI, Michele. 2015. "What is big data? A consensual definition and a review of key research topics". In *AIP conference proceedings*. American Institute of Physics, 2015. p. 97-104. http://big-data-fr.com/wp-content/uploads/2015/02/aip-scitation-what-is-bigdata.pdf.

Mazzocchi, Fulvio. 2018. "Knowledge organization system (KOS)". Knowledge Organization 45, no.1: 54-78. Also available in ISKO Encyclopaedia of Knowledge Organization, eds. Birger Hjørland and Claudio Gnoli. http://www.isko.org/cyclo/kos.

Méndez, Eva; Greenberg, Jane. 2012. "Linked data for open vocabularies and HIVE's global framework". *El profesional de la información* 21 no. 3: 236-244.

Mylopoulos, John. 1992. "Conceptual modelling and Telos". In *Conceptual modelling, databases, and CASE: An integrated view of information system development*, p. 49-68. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.3647&rep=rep1&type=pdf.

Ontology Web Language Overview. 2004. W3C. https://www.w3.org/TR/owl-features/.

Orilia, Francesco and Paoletti, Michele Paolini. 2020. "Properties", *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/win2020/entries/properties/.

Otlet, Paul. 2018. Tratado de Documentação: o livro sobre o livro, teoria e prática. Brasília: Briquet de Lemos Livros.

Peirce, Charles. S. 1869. “On a new list of categories”. *Proceedings of the American Academy of Arts and Sciences, v.* 7, p. 287-298, 1868. http://www.bocc.ubi.pt/pag/peirce--charles-list-categories.pdf.

Poole, Alex H. "Now is the Future Now? 2013. “The Urgency of Digital Curation in the Digital Humanities." *DHQ: Digital Humanities Quarterly* 7 no. 2.

Prasad, A. R. D., Giunchiglia, Fausto; Devika, P. Madalli. 21017. “DERA: from document centric to entity centric knowledge modelling”. In: Proceedings of the International UDC seminar 2017. Faceted classification today. London: September, 2017. p. 169-179. http://seminar.udcc.org.

Prieto-Díaz, Ruben. 1990. “Domain analysis: An introduction”. *ACM SIGSOFT Software Engineering Notes*, 15 no. 2: 47-54.

Ranganathan, S. R. and Gopinath, M. A. 1967. *Prolegomena to Library Classification*. 3 ed. Bombay: Asia Publishing House.

RDF semantics. W3C, 2004. http://www.w3.org/TR/rdf-mt/.

RDF 1.1. PRIMER. 2014. W3C. https://www.w3.org/TR/rdf11-primer/. 2019.

Resource Description Framework (RDF) Model and Syntax Specification. 1998. W3C. https://www.w3.org/1998/10/WD-rdf-syntax-19981008/.

Riva, Pat, Le Boeuf, Patrick, and Žumer, Maja. 2017 “*IFLA Library Reference Model: A Conceptual Model for Bibliographic Information”*. IFLA. [online] https://www.ifla.org/publications/node/11412.

Rowley, Jennifer. 2007. "The wisdom hierarchy: representations of the DIKW hierarchy". *Journal of information science* 33 no. 2: 163-180. http://web.dfc.unibo.it/buzzetti/IUcorso2007-08/mdidattici/rowleydikw.pdf.

Saracevic, Tefko. 2007. "Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance". *Journal of the american society for information science and technology* 58 no. 13: 1915-1933.

Shet, Amith. 2020. "Knowledge Graphs and their central role in big data processing: Past, Present, and Future". In 7th ACM India Joint Conference on Data Science & management of Data (COD-COMAD), Indian School of Business, Hyderabad Campus, 5-7 January 2020. https://www.slideshare.net/apsheth/knowledge-graphs-and-their-central-role-in-big-data-processing-past-present-and-future, accessed Jun. 5, 2021.

Shet, Amith; Ramakrishnan, Cartic and Thomas, Christopher. 2005. "Semantics for the semantic web: The implicit, the formal and the powerful". *International Journal on Semantic Web and Information Systems (IJSWIS)*, no. 1 vol. 1:1-18. http://www.ebusinessforum.gr/old/content/downloads/JSWIS.pdf#page=19.

Shiri, Ali. 2013. "Linked data meets big data: A knowledge organization systems perspective." *Advances in Classification Research Online* 24 no. 1: 16-20.

SKOS – Simple Knowledge Organization System Namespace Document. 2012. W3C. https://www.w3.org/2009/08/skos-reference/skos.html#.

Soergel, Dagobert. 2015. "Unleashing the Power of Data Through Organization: Structure and Connections for Meaning, Learning and Discovery." *Knowledge Organization* 42 no. 6: 401-427.

SPARQL 1.1 QUERY LANGUAGE, 2013. W3C. https://www.w3.org/TR/sparql11-query/.

Strecker, Dorothea et al. 2021. *Metadata Schema for the Description of Research Data Repositories*. Re3data, 2021. Available at: https://doi.org/10.48440/re3.010. Access 08 Jul. 2022.

Swanson, Don R. 2008. "Literature-based discovery? The very idea." In *Literature-based discovery*. Springer, Berlin, Heidelberg. 3-11.

Veiga, Viviane Santos de Oliveira; Campos, Maria Luiza; Silva, Carlos Roberto Lyra; Henning, Patricia and Moreira, João. 2021. "Vodan br: a gestão de dados no enfrentamento da pandemia coronavirus". *Páginas A&B, Arquivos e Bibliotecas (Portugal),* n. Especial: 51-58. http://hdl.handle.net/20.500.11959/brapci/157353.

Zeng, Marcia Lei. 2019. "Interoperability". *Knowledge Organization* 46, no. 2: 122-146. Also available in Hjørland, Birger and Gnoli, Claudio eds. *ISKO Encyclopedia of Knowledge Organization*, http://www.isko.org/cyclo/interoperability.

Zeng, Marcia. L. 2017. "Smart data for digital humanities". *Journal of Data and Information Science* 2 no. 1: 1-12. DOI: 10.1515/jdis-2017-0001.

KO-2022-0003 - revised-v3 - Answers to the Reviewers comments

October 1, 2022

Dear Reviewers

Thank you for your valuable comments to our text.

The text was edited removing the revised v2 version markup in red and yellow made to attend to the reviewer comments. Text added to this version was highlighted blue.

Explanation of value vocabularies and metadata vocabularies were expanded according to the reviewer's suggestion in section 4.1.

The reference list was also checked and revised.