# The Role of Vocabularies in the Age of Data: The Question of Research Data

Carlos H. Marcondes

Federal Fluminense University Information Science
R. Lara Vilela, n. 126 - São Domingos CEP 24210-590,
Niterói, RJ, Brazil 24220-301 Niterói Rio de Janeiro 24210-590
<ch_marcondes@id.uff.br>

Carlos H. Marcondes has a PhD in Information Science, is Full professor of the Graduate Program on Information Science, Fluminense Federal University and of the Graduate Program on Knowledge Management and Organization, Minas Gerais Federal University, Brazil. He is a researcher with the Brazilian National Council for Scientific and Technological Development, and consultant. His research interests include knowledge representation and organization in digital environments, data semantics, conceptual modeling and ontologies.

Marcondes, Carlos H. 2022. "The Role of Vocabularies in the Age of Data: The Question of Research Data." *Knowledge Organization* 49(7) 467-482. 77 references. DOI:10.5771/0943-7444-2022-7-467.

**Abstract**: 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. 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. 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 is discussed, and is used to classify the different vocabularies; within such a classification ontologies, are shown to be a 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.

Received 26 January 2022; Revised 3 October 2022: Accepted 1 November 2022

Keywords: automatic processing, big data, controlled vocabularies

## 1.0 Introduction

> "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."
>
> Mark Musen (FAIR Compliant Biomedical Metadata Templates | CEDAR 2019).

Big Data is the name of the phenomenon that is the huge amount of digital data being created at enormous velocity, and with great heterogeneity, as the result of social, economic, scientific and cultural activities centred on the web. Today's research data also shares the characteristics of Big Data (Fillinger et al. 2019). Data is created in huge quantities and with great speed directly from monitoring devices and projects like the Hubble Space Telescope, the Human Genome research project and 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 semantics. In today's Web landscape such themes are interwoven. Research data is an important product of science, alongside scientific publications. How can we deal with the "V"s of Big Data (namely Volume, Velocity, Variety, Variability, Veracity) in research data to enhance its "V"alue and achieve insights of such data (Iafrate 2015, 3), and how can its large-scale reuse be facilitated? Within such a context and considering the statement by researcher Mark Musen, what can be the contribution of vocabularies, an important research area in knowledge organization (KO).

### 1.1 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. Headlines also call this 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 recalls 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, for example 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 such as its impact in KO epistemology and methodologies (Hajibayova and Salaba, 2018; Frické 2015). However, contributions from the area proposing practical solutions are still few. Hjørland (2013, 179) observed that "... 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 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) comment on the "opacity of the algorithms behind the platforms and systems".

The best-known product of science, one in which KO has been interested 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 which 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 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) 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 retrieval (Foskett 1996).

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 the-

sauri, 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.

## 2.0 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 and 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 2007). 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 and Gopinath 1967). Facets are the properties of a class of things of interest for information recovery (Giunchiglia et al. 2014; Giunchiglia 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 (FLA 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 and Madalli 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 DDC and the UDC, 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 the-

> ories 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 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, 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". 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), 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 identify-

ing 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 person responsible for the book, and field 245 subfield $f their attributes birth and death dates.

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 designating things that have properties whose loss implies the non-existence of that individual and have existential independence (Fonseca et al. 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 vocabulary, 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 International-ised 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.

## 3.0 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, among others, of the Web. 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 predicate, 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 person 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 RDF triples. Many old vocabularies are being restructured to be compatible with LOD technologies (Soergel 2004; Dos Santos Maculan 2015). Examples include 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 (the Art and Architecture Thesaurus, the Union List of Artists Names, the Cultural Objects Name Authority, the Getty Thesaurus of Geographic Names) https://www.getty.edu/research/tools/vocabularies/lod/, 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 has been used until the present, as "An ontology is 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, them being either 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.

In 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.

## 4.0 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 World Health Organization (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).

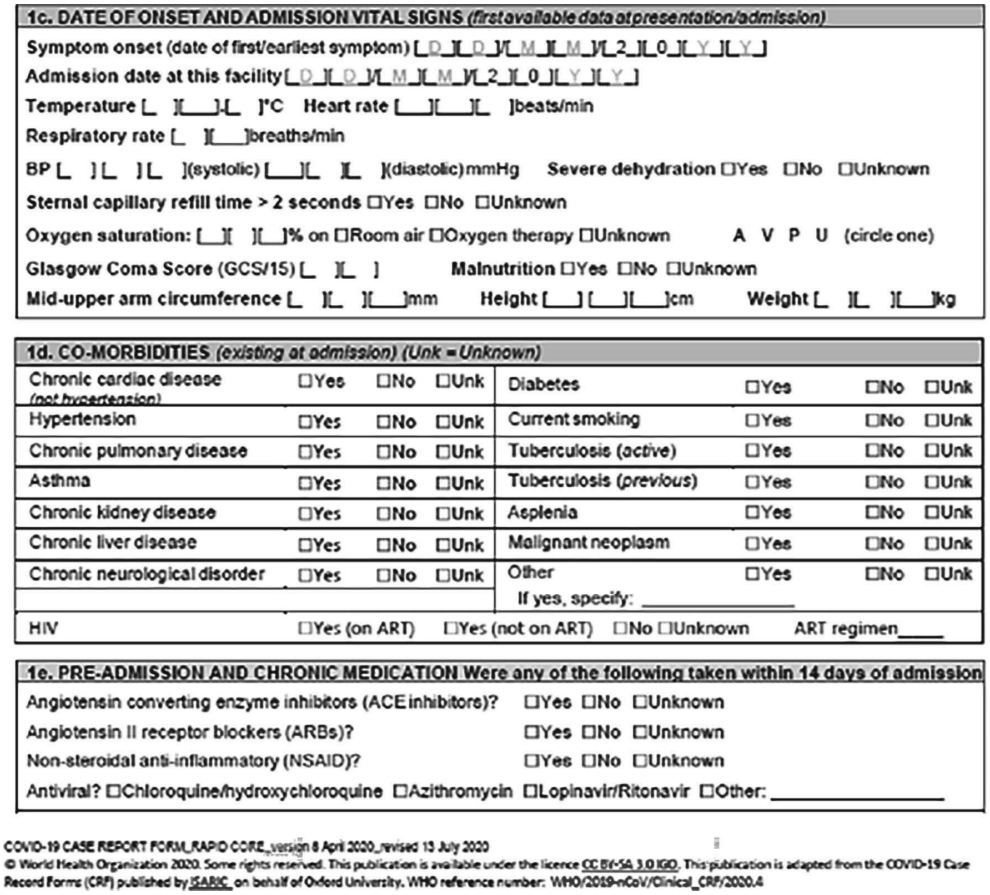
The VODAN initiative is expected to collect huge datasets worldwide. The CRF standardized a set of fields of interest to COVID-19 epidemic research. Such fields must be filled with metadata and data associated with vocabularies largely agreed and standardized within the health sciences domain. This allows 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 the 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, Unknown. These data have to be processed by programs so that the immense number 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.

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 of 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".



**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
© World Health Organization 2020. Some rights reserved. This publication is available under the licence CC BY-SA 3.0 IGO. This publication is adapted from the COVID-19 Case Record Forms (CRF) published by ISARIC on behalf of Oxford University. WHO reference number: WHO/2019-nCoV/Clinical_CRF/2020.4

*Figure 1.* Part of the CRF Form.

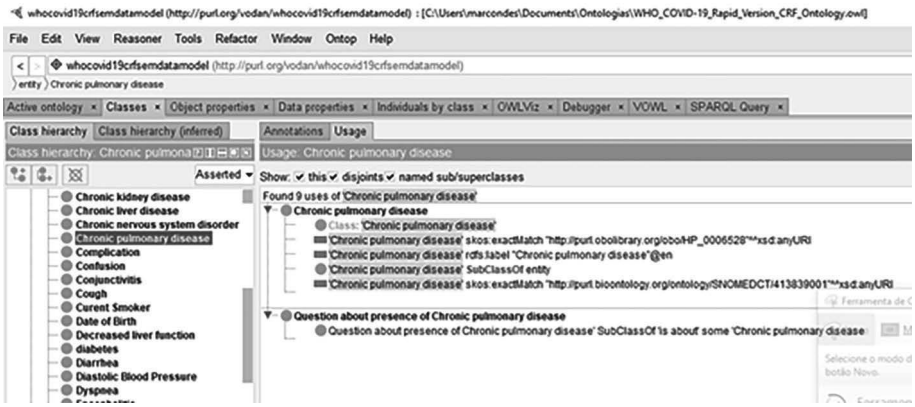



*Figure 2.* The class "chronic pulmonary disease" of the WHO COVID-19 Rapid Version CRF semantic data model and its SKOS mapping to the SNOMED concept.

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 de-

manded 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 "[t]here 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 the 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 research data exploration. They need additional metadata such 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 detail 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 depend 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, aims at transforming such non-structured or semi-structured 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 the 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 these 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 such, 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 sketches the relationships between data aggregation levels to digital units of meaning.

## 5.0 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 pose several issues related to the semantics 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) "[i]n 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 achieving such a 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 on improving the text.

## References

Almeida, Mauricio, Renato Souza and Fred Fonseca. 2011. "Semantics in the Semantic Web: A Critical Evaluation." *Knowledge Organization* 38: 187-203. doi=10.1.1.1041.7976&rep=rep1&type=pdf

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: 229-36.

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

| 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 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. | 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. |

*Table 1.* Relationships between data aggregation levels and digital units of meaning.

Bodenreider, Olivier. 2004. **"**The Unified Medical Language System (UMLS): Integrating Biomedical Terminology.**"** *Nucleic Acids Research* 32 **S**uppl_1: D267-D270.

Borst, Willem N. 1997. *Construction of Engineering Ontologies.* Enschede, Netherlands**:** Centre for Telematica and Information Technology**,** University of Twent**e**.

Capurro, R. 2000. "Angeletics - A Message Theory." In *Hierarchies of Communication* edited by H.H. Diebner and L. Ramsay. 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: 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: 142-51. 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. 2: 64-73. https://dl.acm.org/doi/pdf/10.1145/2500499

Dextre Clarke, Stella G. 2019. "The Information Retrieval Thesaurus." *Knowledge Organization* 46: 439-59. https://www.ergon-verlag.de/isko_ko/downloads/ko_46_2019_6_c.pdf

Dextre Clarke, Stella G. and Marcia Lei Zeng. 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 Alan Hopkinson. 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*, ed-

ited by Edward N. Zalta. https://plato.stanford.edu/archives/win2019/entries/information-semantic/

Foskett, A. C. 1996. *The Subject Approach to Information*. London: Facet.

Fonseca, Claudenir M., Daniele Porello, Giancarlo Guizzardi, João Paulo A. Almeida and Nicola Guarino. 2019. "Relations in Ontology-Driven Conceptual Modeling." In *Conceptual Modeling*, edited by A. Laender, B. Pernici, E. P. Lim and J. de Oliveira. Lecture Notes in Computer Science 11788. Cham: Springer. 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: 6791-800. doi:10.1007/s00216-019-02074-9

Freitas, C., P. Carvalho, H. G. Oliveira, C. Mota and D. Santos. 2010. "Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese." In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, edited by Nicoletta Calzolari et al. Valletta: European Language Resources Association, 3630-37.

Frické, Martin. 2015. "Big Data and Its Epistemology." *Journal of the Association for Information Science and Technology* 66: 651-61.

Gandomi, Amir and Murtaza Haider. 2015. "Beyond the Hype: Big Data Concepts, Methods, and Analytics." *International Journal of Information Management* 35: 137-44.

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

Giunchiglia, Fausto, Biswanath Dutta and Vincenzo Maltese. 2014. "From Knowledge Organization to Knowledge Representation." *Knowledge Organization* 41: 44-56. http://eprints.biblio.unitn.it/4186/1/techRep027.pdf

Gray, Jim. 2009. "eScience: A Transformed Scientific Method." In *The Fourth Paradigm, Data-intensive Scientific Discovery*, edited by 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*. Berlin; Heidelberg: Springer, 139-70. https://kask.eti.pg.gda.pl/redmine/projects/sova/repository/revisions/5378040326bc499e118636a1d25ad667285e005c/entry/Praca_dyplomowa/materialy/10.1.1.53.939.pdf

Guarino, Nicola, Massimiliano Carrara and Pierdaniele Giaretta. 1994. "Formalizing Ontological Commitment." In *Proceedings of AAAI 1994*, 560-7. https://www.aaai.org/Papers/AAAI/1994/AAAI94-085.pdf

Gruber, Thomas R. 1993. "A Translation Approach to Portable Ontology Specifications." *Knowledge Acquisition* 5: 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*, Advances in Knowledge Organization Vol. 16, edited by Fernanda Ribeiro and Maria Elisa Cerveira. Baden-Baden: Ergon.

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 and Anne Trefethen. 2003. "The Data Deluge: An E-Science Perspective." In *Grid Computing: Making the Global Infrastructure a Reality*. London: Wiley, 809-24. https://eprints.soton.ac.uk/257648/1/The_Data_Deluge.pdf

Hjørland, Birger. 2018. "Data (With Big Data and Database Semantics)." *Knowledge Organization* 45: 685-708.

Hjørland, Birger. 2002. "Domain Analysis in Information Science: Eleven Approaches–Traditional as Well as Innovative." *Journal of Documentation* 58: 422-62.

Hjørland, Birger. 2013. "Theories of Knowledge Organization - Theories of Knowledge." *Knowledge Organization* 40: 169–81.

Hjørland, Birger and Hanne Albrechtsen. 1995. "Toward a New Horizon in Information Science: Domain-Analysis." *Journal of the American Society for Information Science* 46: 400-25.

Hjørland, Birger and Jenna Hartel. 2003. "Introduction to a Special Issue of *Knowledge Organization*." *Knowledge Organization* 30: 125-7.

Iafrate, Fernando. 2015. *From Big Data to Smart Data*. London: ISTE; Hoboken, NJ: John Wiley.

Ibekwe-SanJuan, Fidelia and Geoffrey C. Bowker. 2017. "Implications of Big Data for Knowledge Organization." *Knowledge Organization* 44: 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.

ISO/IEC 20546:2019 (en). 2019. *Information Technology - Big data - Overview and Vocabulary*. ISO.

Kahn, Robert and Robert Wilensky. 2006. "A Framework for Distributed Digital Objects Services." *International Journal on Digital Libraries* 6: 115–123. https://www.doi.org/topics/2006_05_02_Kahn_Framework.pdf

Lambe, Patrick. 2007. *Organising Knowledge: Taxonomies, Knowledge and Organizational Effectiveness*. Oxford: Chandos.

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

Leonelli, Sabina and Celia Dias. 2020. "Representing Facet Classification in SKOS." In *Knowledge Organization at the Interface. Proceedings of the 16th International ISKO Conference, Aalborg, Denmark*, edited by M. Lykke, T. Svarre, N. Skov and D. Martínez-Ávila. Advances in Knowledge Organization 9. Baden-Baden: Ergon, 254–63. https://doi.org/10.5771/9783956507762

De Mauro, Andrea, Marco Greco and Michele Grimaldi. 2015. "What is Big Data? A Consensual Definition and a Review of Key Research Topics." In *AIP Conference Proceedings*. American Institute of Physics, 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: 54-78. Also available in ISKO Encyclopaedia of Knowledge Organization, edited by Birger Hjørland and Claudio Gnoli. http://www.isko.org/cyclo/kos

Méndez, Eva and Jane Greenberg. 2012. "Linked Data for Open Vocabularies and HIVE's Global Framework." *El Profesional de la Información* 21: 236-44.

Mylopoulos, John. 1992. "Conceptual Modelling and Telos." In *Conceptual Modelling, Databases, and CASE: An Integrated View of Information System Development*, edited by Pericles Loucopoulos and Roberto Zicari. London: Wiley, 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 Michele Paolini Paoletti. 2020. "Properties." In *The Stanford Encyclopedia of Philosophy* edited by Edward N. Zalta. 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.

Peirce, Charles. S. 1869. "On a New List of Categories." *Proceedings of the American Academy of Arts and Sciences* 7: 287-98. http://www.bocc.ubi.pt/pag/peirce--charles-list-categories.pdf

Poole, Alex H. 2013. "Now is the Future Now? The Urgency of Digital Curation in the Digital Humanities." *DHQ: Digital Humanities Quarterly* 7 no. 2. www.digitalhumanities.org/dhq/vol/7/2/000163/000163.html

Prasad, A. R. D., Fausto Giunchiglia and Devika P. Madalli. 2017. "DERA: from Document Centric to Entity Centric Knowledge Modelling." In: *Proceedings of the International UDC Seminar 2017. Faceted Classification Today: Theory, Technology and End Users London, 14-15 September, 2017* edited by Aida Slavic and Claudio Gnoli. Wurzburg: Ergon, 169-79.

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

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

*Resource Description Framework RDF Semantics*. 2004. W3C Recommendation 10 February 2004. http://www.w3.org/TR/rdf-mt/

*Resource Description Framework RDF 1.1. Primer*. 2014. W3C Working Group Note 24 June 2014. https://www.w3.org/TR/rdf11-primer/

*Resource Description Framework (RDF) Model and Syntax Specification*. 1998. W3C Working Draft 08 October 1998. https://www.w3.org/1998/10/WD-rdf-syntax-19981008/

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

Rowley, Jennifer. 2007. "The Wisdom Hierarchy: Representations of the DIKW Hierarchy". *Journal of Information Science* 33: 163-80. 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: 1915-33.

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

Shet, Amith, Cartic Ramakrishnan and Christopher Thomas. 2005. "Semantics for the Semantic Web: The Implicit, the Formal and the Powerful." *International Journal on Semantic Web and Information Systems (IJS WIS)* 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: 16-20.

*SKOS Simple Knowledge Organization System Namespace Document*. 2012. 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: 401-27.

*SPARQL 1.1 Query Language.* 2013. W3C Recommendation 21 March 2013. https://www.w3.org/TR/sparql11-query/

Strecker, Dorothea, Roland Bertelmann, Helena Cousijn, Kirsten Elger et al. 2021. *Metadata Schema for the Description of Research Data Repositories*. *Version 3.1.* https://doi.org/10.48440/re3.010

Swanson, Don R. 2008. "Literature-based Discovery? The Very Idea." In *Literature-based Discovery*, edited by P. Bruza and M. Weeber. Berlin; Heidelberg: Springer, 3-11.

Veiga, Viviane Santos de Oliveira, Maria Luiza Campos, Carlos Roberto Lyra Silva, Patricia Henning and João Moreira. 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: 122-46. Also available in *ISKO Encyclopedia of Knowledge Organization*, edited by Birger Hjorland and Claudio Gnoli. http://www.isko.org/cyclo/interoperability

Zeng, Marcia. L. 2017. "Smart Data for Digital Humanities." *Journal of Data and Information Science* 2: 1-12. DOI: 10.1515/jdis-2017-0001