

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

Journal:	<i>Knowledge Organization</i>
Manuscript ID	KO-2022-0003.R2
Manuscript Type:	Article

SCHOLARONE™
Manuscripts

1
2
3
4 **THE ROLE OF VOCABULARIES IN BIG DATA: the quest of research data¹**
5
6

7 Carlos H. Marcondes
8
9

10 PPG-GOC/UFMG - Postgraduate Program in Knowledge Management and Organization,
11
12

13 Minas Gerais Federal University, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo
14

15 Horizonte - MG, 31270-901, Brazil, ch_marcondes@id.uff.br
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55

56
57 ¹ This is a revised and extended version of the article with the same title presented at IV Seminário de
58 Pesquisas do grupo MHTX, Nov 2021, <http://eci.ufmg.br/iv-seminario-do-grupo-de-pesquisa-mhtx/>.
59
60

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

1
2
3 repeat those originating in computer science, emphasizing aspects technological aspects as
4
5 volume, variety, velocity, heterogeneity and the need of massive computer power to process
6
7 it.
8
9

10
11 The best-known product of science, to which the KO has been dedicated since its
12
13 beginnings, are scientific publications. More recently, science has been giving increasing
14
15 importance to another of its products, research data. Today, research data, practically
16
17 entirely digital, is produced in increasing quantities as a result of scientific activity carried
18
19 out with the support of information technologies. Examples of this huge amount of digital
20
21 survey data are those generated by the Hubble Space Telescope,
22
23 https://www.nasa.gov/mission_pages/hubble/main/index.html, the Human Genome research
24
25 project, <https://www.genome.gov/human-genome-project>, or the Large Hadron Collider,
26
27 <https://home.cern/science/accelerators/large-hadron-collider>, the largest and most powerful
28
29 particle accelerator in the world. This digital research data is part of the Big Data
30
31 phenomenon.
32
33
34
35

36
37 As quoted by researcher Mark Musen at the beginning of this work: we cannot
38
39 address Big Data without using computers to help us. This observation refers to the
40
41 Semantic Web project (Berners_Lee et al 2001), the proposal for a Web whose resources
42
43 would be represented in a way that had a precise and formal meaning or semantics and
44
45 would be intelligible and understandable by both people and machines.
46
47
48

49
50 In previous works, we have already discussed how to link digital representations of
51
52 objects of memory and culture through the Web (Marcondes, 2020), and how one of the
53
54 products of science, scientific publications, could be intelligible and understandable by both
55
56 people and machines (Marcondes and Costa, 2016), when represented with the technologies
57
58
59
60

1
2
3 of Linked Open Data (LOD) and the Semantic Web. Here we are interested in doing the
4
5 same for science's other great product, digital research data.
6
7

8 9 1.2. Traditional use of vocabularies to assign subjects to documents

10
11
12 Representing documents and their subjects has been foundational to the practices
13
14 developed by KO, especially when, unlike today, there was no access to full-text documents
15
16 in digital format and the descriptive and thematic representation of the documents was a
17
18 fundamental mechanism in the intermediation, and relevance assessment processes carried
19
20 out in the retrieval of information (Saracevic, 2007). KO methodologies have always
21
22 represented domains of knowledge when building KOS like controlled/standardized
23
24 vocabularies, subject headings, and classification schemas. The early KOS, such as thesauri,
25
26 were intended to enable subject-based retrieval in the context of IRS because their records
27
28 were representations of objects that had subjects as one of their properties. But not all
29
30 objects in a domain have subjects as one of their properties, like documents. We see today
31
32 that this is just a case of representing objects in digital space.
33
34
35
36
37

38
39 Today, it is not only about retrieving documents (or their representations) but also to
40
41 create digital representations of anything, such as in the "Internet of Things" (IoT), If the
42
43 documentation movement (Otlet, 2018) and then information science intended the
44
45 empowerment of "information" by separating it from books, the Semantic Web proposes to
46
47 also, in a certain sense, empower "knowledge," which is no longer just inserted into texts to
48
49 be interpreted by humans, but rather recorded directly in Resource Description Framework
50
51 (RDF) triples (RDF 1.1 PRIMER, 2014), forming representations/descriptions of "things."
52
53 The Web thus becomes a large knowledge base that can be consulted about the "things" thus
54
55 represented (SPARQL 1.1 QUERY LANGUAGE, 2013).
56
57
58
59
60

1
2
3 The objective of this work is to discuss how KO can contribute to assigning
4 computational semantics to Big Data, especially to research data, so that computers can
5 process them, allowing their reuse on a large scale. As a methodology, the work discusses
6 the conceptualizations of data and (the few of) Big Data originating in KO in an attempt to
7 make it clearer what would be data, essential elements of the Big Data phenomenon, and in
8 particular, digital research data. It then proceeds to analyze digital research data based on the
9 Case Report Form (CRF), [WHO-COVID-CRF/WHO-2019-nCoV-Clinical_CRF-2020.3-
10 eng.pdf at master · FAIRDataTeam/WHO-COVID-CRF · GitHub](#), proposed by the World
11 Health Organization (WHO) to standardize and unify the registration of cases of patients
12 with COVID-19 worldwide.
13
14
15
16
17
18
19
20
21
22
23
24
25
26

27 The work is organized as follows. After this introduction, section 2 analyzes data
28 definitions, their traditional use in KO, and develops a conceptualization of data that is
29 illustrated by an example of research datasets, relating them to the representation of things in
30 a domain and organized into vocabularies. Section 3 presents a comprehensive view of
31 vocabularies based on Semantic Web and LOD technologies and discusses which
32 functionalities vocabularies must incorporate to integrate with these technologies. Section 4
33 raises research questions to be developed and presents final considerations.
34
35
36
37
38
39
40
41
42
43

44 **2. A Semiotic and Ontological view of data**

45
46
47 None of the most common Big Data definitions exclude the data component. It
48 seems reasonable, then, that understanding what Big Data is and how to operationalize
49 solutions to the problem begins by elucidating what data is. This section proposes a semiotic
50 and ontological analysis of data, understood as the essential component of Big Data. This
51 analysis begins with the question of elucidating how to assign semantics to the data. Then
52 we discuss what data is, from an ontological point of view. From the elucidation of these
53
54
55
56
57
58
59
60

1
2
3 questions, concepts of research data, data concerning domains of human action, and
4
5 vocabularies as representations of domains, are developed.
6
7

8 9 2.1. Data as Representations 10

11
12 What is Big Data? What is its relationship with data? What is data and how is it
13 related to metadata? How should semantics be assigned to data? The ISO/IEC 20546/2019
14 Standard notes, “The big data paradigm is a rapidly changing field with rapidly changing
15 technologies,” later suggesting a definition: “extensive datasets (3.1.11) — primarily in the
16 data (3.1.5) characteristics of volume, variety, velocity, and/or variability — that require a
17 scalable technology for efficient storage, manipulation, management, and analysis.”
18
19
20
21
22
23
24
25

26
27 The conceptualizations of Big Data originating from KO are few (Marcondes et al
28 2021) and replicate those originating in computer science, defining it as a phenomenon that
29 involves large amounts of data, the heterogeneity of that data, a continuous flow of
30 generation and updating, and a need for large processing capacity, so that the data reveal
31 patterns or trends (De Mauro et al 2015). However, the same is not true for the
32 conceptualizations of data originating from KO. Data is mentioned frequently in the
33 literature, along with its relationships with information and knowledge (Buckland, 1991),
34 often called the data, information, knowledge, wisdom (DIKW) hierarchy (Rowley, 2007).
35 In Floridi (2019), information is related to data and semantics.
36
37
38
39
40
41
42
43
44
45
46
47
48

49 An important exception is from Hjørland (2018), who proposes a conceptualization
50 of Big Data arising from definitions of data, a phenomenon much better known and
51 conceptualized in the area. Data is the essence of the Big Data phenomenon, it could not
52 exist without data. In this work, Hjørland lists several similar conceptualizations of data and
53 highlights that of Fox and Levitin:
54
55
56
57
58
59
60

1
2
3 Within this framework, we define a datum or data item, as a triple $\langle e, a, v \rangle$,
4
5 where e is an entity in a conceptual model, a is an attribute of entity e , and v
6
7 is a value from the domain of attribute a . A datum asserts that entity and has
8
9 value v for attribute a . Data are the members of any collection of data items.
10

11
12 This conceptualization is clarified by the following example: “2018.” What does
13
14 2018 mean? Others would say it’s a given. Let us note, however, this statement: “Giovana
15
16 was born in 2018.” In it we can identify the entity we are talking about: a child called
17
18 “Giovana,” an attribute or property of this entity, she is “born,” and the value of this
19
20 attribute or property, her year of birth, “2018.”
21
22

23
24
25 In the ontological scheme that goes back to Aristotle (2000), reality is constituted of
26
27 the first substances, the things that have real existence in space and time, and second
28
29 substances, the conceptualizations we make of the first substances to think, reason, make
30
31 sense of, and communicate about the things in reality. Second substances are in turn
32
33 subdivided into essences, concepts that designate things that have existential independence,
34
35 and accidents, concepts that designate things that are existentially dependent on other
36
37 substances. Things that have existential independence are commonly recognized in one of
38
39 the most well-known ontological schemes, the entity-relationships (ER) model (Chen, 1976)
40
41 as entities, while those that are existentially dependent, as properties. Properties, in turn, are
42
43 subdivided into attributes of an entity, relationships between an existentially independent
44
45 entity and the value of one of its properties, and relationships, involving two or more
46
47 individuals of the same existentially independent entity, or of more than one existentially
48
49 independent entity (Orilia and Paoletti, 2020).
50
51
52

53
54
55 We are talking about representations. A piece of data, even in the context of Big
56
57 Data, then, makes no sense without referencing the entity and one of its properties, the
58
59
60

1
2
3 metadata. The three concepts are inseparable and cannot be understood separately. They
4
5 correspond to a descriptive, representational element of an entity, describing one of its
6
7 properties. They correspond linguistically to a claim, a basic unit of knowledge to which,
8
9 according to Aristotle (2000, p. 39), values of truth or falsity can be attributed.
10
11
12

13
14 The statements represented by triples constituted by an entity, one of its properties,
15
16 and the value of this property correspond to the representation of informational resources in
17
18 the context of LOD, using the RDF (RDF Primer, 2014). RDF is a Semantic Web standard
19
20 for describing resources. Everything that is available on the Web can be accessed through a
21
22 link, or a Uniform Resource Identifier (URI).ⁱ This representational model describes such a
23
24 resource through triples formed by subject, the resource being described; predicate, a
25
26 property that describes the resource; and object, the value of this property for this resource.
27
28 The RDF model assumes a minimum semantics, that is, the subject, the predicate, and the
29
30 object that form the triple are identified and appear in this order.
31
32
33
34
35

36 2.2. Data and Big Data: the case for research data

37
38
39 Next, we will attempt to demonstrate how the conceptualization above helps address
40
41 the issues of Big Data, especially research data. A concrete and dramatic example of the
42
43 importance of research data and the adoption of principles and technologies that allow its
44
45 wide dissemination and reuse is the form for collecting data from patients infected with
46
47 COVID-19, the CRF, which was proposed by the WHO. The GO FAIR initiative,
48
49 <https://www.go-fair.org/>, proposes the creation of a worldwide network of catalogs that can
50
51 reference research data collected through the CRF and deposited in repositories and that are
52
53 available according to the FAIR principles, the “FAIR Data Points.” Brazil participates in
54
55 this initiative through the VODAN-Br Virus Outbreak Data Network initiative (Veiga et al
56
57
58
59
60

1
2
3 2021). The form fields must be filled with metadata and data associated with vocabularies to
4
5 allow their standardization, without which their processing by computers would not be
6
7 possible, and consequently neither would the ability to extract conclusions and insights.
8
9

10
11 In the RDF model, instead of the subject, predicate, and object of a triple being
12
13 represented in natural language, which is ambiguous and difficult for programs to process,
14
15 each can be identified by a URI. These URI identify specific terms, both from metadata
16
17 vocabularies—descriptive properties of things in a domain—and from data vocabularies—
18
19 values assumed by these properties for specific instances. This unified characterization as
20
21 vocabularies, that is, sets of systematized terms that identify both the descriptive properties
22
23 (metadata) of objects in a domain, as well as the data, as the values assumed by these
24
25 properties for instance, is due to Marcia Zeng (2019).
26
27
28
29

30
31 Another important feature of using vocabularies with LOD technologies is that
32
33 different vocabularies can be used simultaneously in the form fields. In Figure 1 we see an
34
35 excerpt from the CRF. As co-morbidity data, “CO-MORBIDITIES,” of a patient (the entity)
36
37 are recorded, concepts such as chronic cardiac disease (the attribute or metadata) are taken
38
39 from specific biomedical ontologies or vocabularies: Yes, No, Unk (the value or data).
40
41 These data have to be processed by programs so that the immense amount of records
42
43 collected through the CRF around the world can serve as inputs for the planning and control
44
45 of the pandemic. The question about co-morbidities has several answer options, each of
46
47 which indicates a type of disease. For it to be processed by machines, each type of co-
48
49 morbidity expressed in natural language must reference a concept in a vocabulary or
50
51 ontology, such as SNOMED-CT, <https://www.nlm.nih.gov/healthit/snomedct/index.html>,
52
53 for example. Another question on the CRF, such as the one related to “PRE-ADMISSION
54
55 AND CHRONIC MEDICATION,” has as one of its answer options “Angiotensin
56
57 converting enzyme inhibitors (ACE inhibitors)?”, which may be referenced in another
58
59
60

1
2
3 in RDF format, include predicates and objects referring to standardized vocabularies, is
4 widely recognized by the community in a given domain, and linked together to provide rich
5 context. For research data, which has demanded increasing attention and public policies at
6 national and international levels, the international GO FAIR initiative recommends a set of
7 principles for publication so that they have the attributes of FAIR: findability, accessibility,
8 interoperability, and reuse. For research data to achieve these attributes, they must be
9 accessible through a URI, represented in RDF, constituting the Linked Open Vocabularies
10 (LOV).
11
12
13
14
15
16
17
18
19
20

21
22 The idea behind the FAIR principles is to allow research data to be processed by
23 machines. The M4M principle—metadata for machines—states that “There is no FAIR data
24 without machine-actionable metadata. The overall goal of Metadata for Machines
25 workshops (M4M) is to make routine use of machine-actionable metadata in a broad range
26 of fields.” An example of the importance of research data and the adoption of principles that
27 allow its wide dissemination and reuse is the CRF form described above. Without
28 standardization, its processing by machines would be impossible.
29
30
31
32
33
34
35
36
37
38

39 2.3. Traditional use of vocabularies to assign subjects to documents, generalized use of
40 vocabularies as representations of a domain
41
42
43
44

45 Since the onset of the information explosion, thesauri have emerged, complementary
46 systems to the IRS, of which the KOS were one of its components. The development of the
47 IRS drew on sources from other traditions of librarianship, documentation and cataloging.
48 Catalog sheets and bibliographic entries served as models for computational bibliographic
49 formats in projects such as Machine Readable Cataloging (MARC), based on the AACR2
50 cataloging standard, and the UNISIST Reference Manual (Dierickx, Hopkinson, 1986),
51 based on the ISBD standard. These formats served as a model for library catalog databases
52
53
54
55
56
57
58
59
60

1
2
3 and for indexing and summary services. Significantly, bibliographic formats evolved
4
5 separately from the also nascent technology of computer databases that, from the 1970s
6
7 onwards, had the relational model as a paradigm (Codd, 1970). The Text Retrieval
8
9 Conferences (TREC) conference series illustrates this separate evolution.
10
11

12
13 Concerning thematic representation, a whole theoretical and methodological
14
15 foundation were all developed to support the development of KOS, from classification
16
17 theories, the Faceted Classification Theory (Ranganathan, Gopinath, 1967), the proposals of
18
19 the Classification Research Group (CRG) (Wilson, 1972), Concept Theory (Dahlberg,
20
21 1978), to Terminology (Cabr e, 2005). This theoretical and methodological tradition , an area
22
23 of excellence of KO, meets, with the emergence of the Semantic Web, in subdisciplines such
24
25 as systems modeling, artificial intelligence, and computational ontologies,  reas originating
26
27 from computer science. These are understood as one of the foundations of the proposed
28
29 Semantic Web. Many of these new KOS are developed by computer professionals and
30
31 scientists from different areas or specialists: biomedicine, statistics, or from curators of
32
33 digital collections in memory and culture, etc. The words of Hjørland (2008, p. 86) highlight
34
35 and warn about this approach to other areas: “(LIS) is the central discipline of KO in this
36
37 narrow sense (although seriously challenged by, among other fields, computer science).”
38
39 Will KO limit itself to developing traditional KOS and leave this space to computer science
40
41 specialists as Hjørland warns?
42
43
44
45
46
47
48

49 The technical traditions and standards developed by KO to manage the information
50
51 explosion resulted in the establishment of IRS/KOS assumptions that persist to this day. In
52
53 most discourse in the area, these assumptions are so implicit that it becomes difficult to
54
55 make them explicit, consider them, and analyze their consequences. All the theories and
56
57 methodologies of KO mentioned bring these assumptions implicitly: the IRS represent
58
59 documents in their computerized databases, MARC and the bibliographic formats that
60

1
2
3 emerged from the UNISIST Reference Manual are sets of metadata that represent different
4 (descriptive) properties of the documents, while the KOS associated with them are
5 terminological standardization instruments specifically for the subject property, the subject
6 field of the records of the IRS computerized databases. These records represent objects that
7 have, among others, the subject property. They are symbolic objects, documents. It is worth
8 adding that the records themselves, the metadata set, are also symbolic objects, representing
9 document-type objects.
10
11
12
13
14
15
16
17
18
19

20 These implicit assumptions account for the division that occurs in the teaching and
21 practice of librarianship and KO between descriptive representation and thematic
22 representation, or of subjects, of a document. To what extent do these assumptions hold up
23 today, and are they sufficient to address the challenges of the Semantic Web era, Big Data,
24 and the Internet of Things?
25
26
27
28
29
30
31

32 In the 1980s-1990s, as a consequence of the emergence of online bibliographic
33 catalog management systems and databases, the domain of information retrieval in library
34 catalogues, so familiar to us, but also so exclusive, with its diversity of objects, was modeled
35 using a methodology used in computer science to plan database management systems. The
36 conceptual model Functional Requirements for Bibliographic Records (FRBR) appeared in
37 1998, whose development was promoted by IFLA (1998).
38
39
40
41
42
43
44
45
46

47 According to Mylopoulos (1992, p. 3) “Conceptual modeling is the activity of
48 formally describing some aspects of the physical and social world around us for purposes of
49 understanding and communication.” For Mylopoulos,
50
51
52

53
54
55 the descriptions that arise from conceptual modeling activities are intended
56 to be used by humans, not machines. . . [and] The adequacy of a conceptual
57 modeling notation rests on its contribution to the construction of models of
58
59
60

1
2
3 reality that promote a common understanding of that reality among their
4
5 human users.
6
7

8 A conceptual model sets an agreement between users of a system on what kinds of
9 things exist and will be represented in the system, or entities (also called classes) in a given
10 domain of reality, e.g. documents of historical value, the properties of these entities and how
11 they relate (relationships) to each other. Thus, a conceptual model is a representation, in the
12 form of an abstract and generic description, independent of computational implementations
13 (hardware, operating systems, languages, database management systems) of a given domain
14 of reality. To understand this reality, reason about it and establish a common understanding
15 of this reality, a conceptual model answers questions such as: What different things exist in
16 a given domain? How are they distinguished from each other? How do they relate? What are
17 your properties?
18
19
20
21
22
23
24
25
26
27
28
29
30
31

32 A conceptual model as a representation is expressed, communicated, externalized
33 through a language, or more specifically a meta-language or meta-model (Guizzardi, 2007,
34 23), which is a language to express the languages that express things in specific domains.
35 Examples of these meta-languages are either natural language (through a system
36 requirements document), which functions as the most general of all meta-languages, or a
37 diagrammatic meta-language, such as entity-relationship (meta) Model. The Unified
38 Modeling Language (UML), <https://www.uml.org/>, class diagram, in which domain-specific
39 ER models or class diagrams are expressed, both an ER model and a class diagram can
40 define a language that designates things in a domain or a specific vocabulary to that domain.
41
42
43
44
45
46
47
48
49
50
51
52
53

54 In the descriptive representation, once established and consolidated practical
55 standards of representation such as MARC, UNISIST, AACR2, and ISDB, the KO started to
56 question what things were implicit within them, their conceptual models.
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

1
2
3 the pioneering FRBR model (Ifla, 1998), the International Council of Museums (ICOM)
4 adopted the CIDOC Conceptual Reference Model (CIDOC, 2014), and more recently the
5
6 International Council of Archives (ICA) adopted the Records in Context Conceptual Model
7
8 (Ric-CM) (International Council on Archives, 2019).
9
10
11
12

13 Conceptual models, when designating things in a domain, define a vocabulary for
14 metadata. They are aligned within different types of KOS (Almeida et al 2011, 196), ordered
15 according to their semantic expressiveness. Semantic expressiveness can be understood, in
16 the context of the previous quote, as the ability of each type of KOS to distinguish and
17 describe, that is, identify the properties and represent the different things that exist in a
18 domain of that reality.
19
20
21
22
23
24
25
26
27

28 Vocabularies of most types are semantic control devices, formed by systematized
29 sets of semiotic, triadic entities (PEIRCE, 1994), concepts (Dextre Clarke and Zeng, 2012),
30 units of meaning that relate something (a first: object or referents), in some way (through a
31 second: term or code), which generates or induces a third: its meaning.
32
33
34
35
36
37

38 Vocabularies are constructed to answer the basic ontological question: what exists
39 in a domain? They are representations or models of a domain of reality, taking of what
40 things there are, what their attributes are, their relationships, and how to express them
41 linguistically, the concepts (Dahlberg, 1978), and their units of meaning. Online Public
42 Access Catalogs (OPACs) or databases are used as terminological control tools with the IRS
43 used by institutions, with the function of standardizing the terms used for the description and
44 indexing of scientific publications, memory and cultural objects, and other items included in
45 these systems.
46
47
48
49
50
51
52
53
54

55 56 57 2.4. Domains 58 59 60

1
2
3 Aside from the general library classification systems such as the CDD and the CDU,
4
5 KOS are developed and used concerning a specific domain. The domain notion commonly
6
7 used in KO is that of a specialized knowledge area.
8
9

10
11 Hjørland and Albrechtsen (1995, 400), in the text in which they propose the analysis
12
13 of domains as the foundation of KO, define domains as: “thought or discourse communities,
14
15 which are parts of society’s division of labor.” They also label a domain as a
16
17 “specialty/discipline/domain/environment” (Hjørland and Albrechtsen, 1995, 401).
18
19

20
21 Hjørland (2002, 422) conceptualizes domains associated with specialized libraries,
22
23 questioning what knowledge would be necessary for information professionals to work in
24
25 “in a specific subject field like medicine, sociology or music?” In Hjørland and Hartel
26
27 (2003, 239), this view of domains as systems of thought, theories, is reaffirmed.
28
29

30
31 Domains are basically of three kinds of theories and concepts: (1)
32
33 ontological theories and concepts about the objects of human activity; (2)
34
35 epistemological theories and concepts about knowledge and the ways to
36
37 obtain knowledge, implying methodological principles about the ways
38
39 objects are investigated; and (3) sociological concepts about the groups of
40
41 people concerned with the objects.
42
43
44

45 The KOS of the early years of KO, such as thesauri, were intended to enable subject-
46
47 based retrieval in the context of IRS because their records were representations of objects
48
49 that had subjects as one of their properties, that is, documents. Today, it is not just about
50
51 retrieving documents (or their representations) but digital representations of anything, as
52
53 exemplified in the IoT. These representations are no longer just access points for documents,
54
55 but also information resources themselves, complex descriptions of these objects, sources of
56
57 knowledge about them, represented in such a way that they can be processed/intelligible by
58
59
60

1
2
3 both machines and humans. Such representations allow machines to make inferences about
4
5 the knowledge thus represented.
6
7

8
9 KO today is being called upon to model different domains of knowledge to build
10
11 new semantic vocabularies. For this, it is necessary to expand the traditional notion of a
12
13 domain as a discipline or subject. In the area of software development, the notion of a
14
15 domain has a broader scope: it is ‘a sphere of activity or interest: field’ [Webster]. In the
16
17 context of software engineering, it is most often understood as an application area, a field for
18
19 which software systems are developed (Prieto Díaz, 1990, 50).
20
21
22

23
24 If we consider that a KOS is a terminological system that represents the “things” of
25
26 interest in a domain of action for the community of agents/users of that domain, to create a
27
28 KOS (an artifact, similar to software) several aspects must be considered. We must first
29
30 determine what things exist in a domain and which are relevant to this community, what
31
32 rules exist about these things or are created/approved/agreed on about these things, how this
33
34 community uses them to act in this domain and, finally, how the conceptualizations
35
36 (Dahlberg, 1978), generating as one of the by-products of this process as a set of terms, are
37
38 to be systematized, for example, in a thesaurus.
39
40
41
42

43
44 What things are in a domain? How should they be represented? These are the
45
46 questions of ontology and semiotics. They must be answered to create a representation, or a
47
48 conceptual model, of a domain.
49
50

51 2.5. Vocabularies as representations of a domain. 52 53

54
55 As shown, a vocabulary is a representation of a domain, regardless of its use, either
56
57 to assign subjects to documents: a) vocabulary for indexing, which identifies what things
58
59 exist in a domain (e.g. MeSH categories describing the entities within the Healthcare
60

1
2
3 domain, <https://meshb.nlm.nih.gov/treeView>), or to b) vocabulary to describe objects in this
4
5 domain, descriptive metadata standards that, in addition to identifying what things exist in a
6
7 domain, also describe and list their properties. Among the things within a domain, there are
8
9 vocabularies of specific facets for special purposes: archival science and records
10
11 management uses functional classification plans in an organization to assign the
12
13 organizational provenance or the function that generated or used a record.
14
15
16
17

18 3. A Comprehensive view of vocabularies

19
20
21 In this section we compiled and developed a comprehensive view of vocabularies
22
23 based on previous discussion in section 2 and on contributions by Hjørland (2018) and Zeng
24
25 (2019).
26
27

28 3.1. Vocabularies, Semantic Web, Linked Open Data, and Big Data

29
30
31 LOD technologies are an integral part of the Semantic Web project. Although this is
32
33 its best-known name, the project is also known as Web of Data, a name that describes it
34
35 better, since semantics concerns meanings (Chierchia, 2003), and the ability of the Web of
36
37 Data to convey meanings is quite limited and different from the sense in our understanding
38
39 of expressions in natural language.
40
41
42
43

44 The project was initially formulated by computer scientist Tim Berners-Lee, the
45
46 creator of the Web, among others. According to its formulators, the Semantic Web aims to
47
48 propose “A new form of Web content that is meaningful to computers will unleash a
49
50 revolution of new possibilities” (Berners-Lee et al 2001). To its authors, “Most of the Web’s
51
52 content today is designed for humans to read, not for computer programs to manipulate
53
54 meaningfully.” The Semantic Web then “will bring structure to the meaningful content of
55
56 Web pages, creating an environment where software agents roaming from page to page can
57
58 readily carry out sophisticated tasks for users.”
59
60

1
2
3 The Web of Data then refers to content represented in such a way that it can be
4 understood by both machines and people. The current Web is made up of pages, such as
5 <http://www.uff.br>, formatted in Hypertext Markup Language (HTML), accessible and
6 interconnected with each other through links. Navigating these pages through these links is
7 done by browsers, such as Internet Explorer, Google Chrome, or Mozilla Firefox. HTML is
8 a content markup language; it formats the content through a pre-defined set of markups,
9 which instruct browsers to display them on computer screens for human users. The content
10 of HTML pages is interpreted by browsers to make it readable and visually pleasing to
11 people.
12
13
14
15
16
17
18
19
20
21
22
23
24

25 The proposed Web of Data is quite different. The Web will no longer be constituted
26 of pages to be read by people, but of content, called informational resources, digital
27 representations of things: concrete, like me, you, an industrial product, a monument, a
28 geographical accident; abstract, like a musical genre, a scientific discipline; or just has a
29 digital existence, such as a photo in a JPG file or a scientific article in a PDF file. These are
30 the entities in the proposal by Hjørland (2018). Each of these resources is uniquely identified
31 by a link, or a URI. A resource, identified/accessed by its URI, is described in a structured
32 way through triples, each one formed by the URI of the resource, by each of its properties,
33 and by the corresponding values of each of these properties. An example of how this
34 representational model works is the Leonardo Da Vinci resource on Wikidata,
35 <https://www.wikidata.org/wiki/Q762>.
36
37
38
39
40
41
42
43
44
45
46
47
48
49

50 This model of structuring data through the description of resources formed by one or
51 more linguistic claims made up of triples <Subject> <Predicate> <Object> is RDF (RDF
52 Primer, 2004). From an ontological point of view, subject, predicate, and object can be
53 understood as an entity, a property, and the value of this property.
54
55
56
57
58
59
60

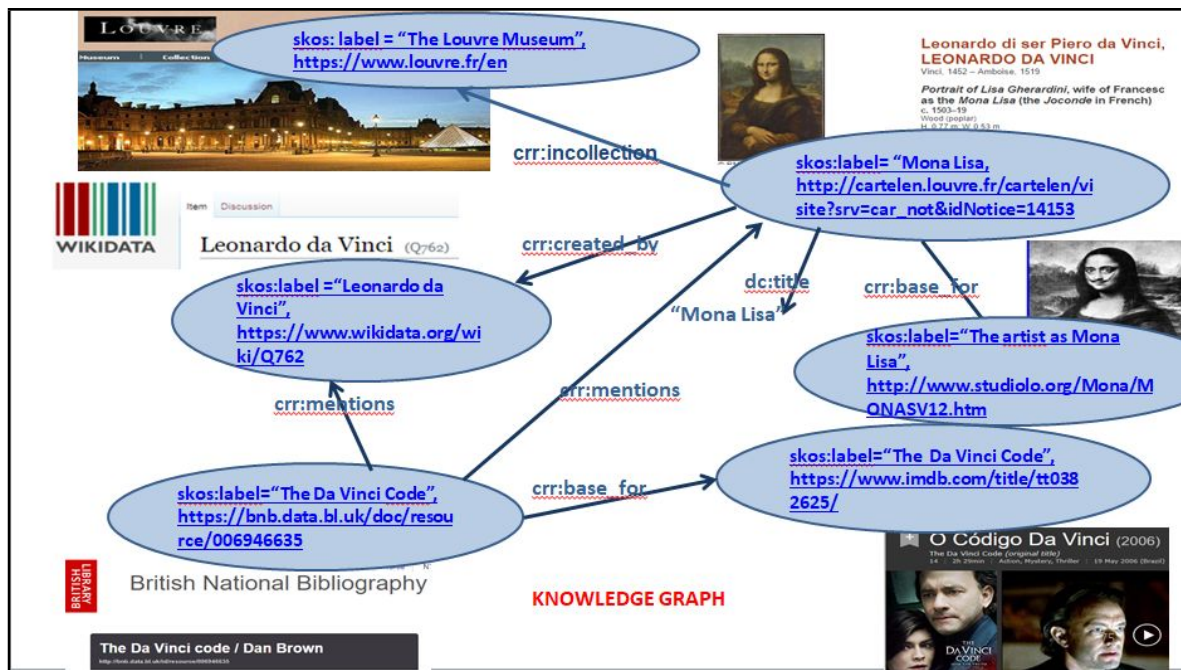
1
2
3 Looking in more detail at structuring a triple; for example,
4
5

6 “The page <http://www.uff.br> is authored by _____.”
7
8

9 We have then a claim that consists of three elements: the subject,
10 “<http://www.uff.br>,” the predicate, “has as author” and the object, “_____”
11
12
13

14 The RDF model presupposes a minimum semantics, derived from its corresponding
15 linguistic claim. That is, they are identified and appear in this order: the subject, the
16 predicate and the object of the claim that form the triple (Resource Description Framework
17 (RDF) Model and Syntax Specification, 1998). A triple describes a specific piece of data
18 from the resource description (what Hjørland calls a “datum:” a unit of data). Sets of triples
19 with the same subject describe the same resource. Sets of linked interlinked triples
20 describing a resource form a graph. Next, we see the graphical representation of an RDF
21 graph.
22
23
24
25
26
27
28
29
30
31
32
33

34 FIGURE 2. Graphical representation of an RDF graph
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



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>?

1
2
3 - < http://www.uff.br > < ??? > < ??? >.
4
5

6 (SPARQL 1.1 QUERY LANGUAGE, 2013) is the query language that allows users
7
8 to query sets of RDF triples, navigating through the graphs formed by them and performing
9
10 inferences. It is the materialization of the Web of Data proposal for a Web that can be
11
12 consulted as if it were a database.
13
14
15

16
17 RDF can be represented (“serialized” in computing technical language) in several
18
19 formats, such as RDF/XML, N Triples, JSON, or TURTLE (RDF Primer, 2004). Of course,
20
21 RDF triples represented in these formats are not as human-friendly or as clearly readable as
22
23 HTML pages when viewed by browsers. But they contain elements that allow browsers to
24
25 understand these formats and display them in a human-friendly manner, if applicable. The
26
27 main objective of the resources described in RDF is that they can be processed by machines
28
29 (including their user-friendly visualization), thus helping to organize, retrieve, and make these
30
31 resources accessible.
32
33
34
35

36 Naturally, given a triple like < http://www.uff.br > < is authored > < _____
37
38 >, a machine cannot do much more than identify the subject, the predicate or the object of
39
40 the triple. In this example, predicate and object are names, strings of characters
41
42 understandable only by people, holders of a set of contextual and cultural information,
43
44 accumulated throughout their life histories. RDF Semantics is limited to its model of
45
46 structuring triples as subject, predicate, object.
47
48
49
50

51 The way to extend these semantics beyond the limits of the RDF model is also to
52
53 make predicates and/or objects into URI and that these URI refer to concepts of vocabularies
54
55 with specific semantics. According to RDF Semantics (2004) “There are several aspects of
56
57 meaning in RDF which are ignored by these semantics; in particular, it treats URI references
58
59
60

1
2
3 as simple names, ignoring aspects of meaning encoded in particular URI forms.” A URI in
4
5 the RDF model is just a name, an identifier. The advantage of a URI over a natural language
6
7 identifier as the linguistic term “author,” is its uniqueness (other properties can be identified
8
9 as the natural language term “author,” synonyms as creator, for example), its validity, as a
10
11 URI, throughout the whole webspace, and its persistence, that is, the commitment of
12
13 whoever assigns a URI to never change it (Berners-Lee, 1998).
14
15
16
17

18 Extending the previous example, we have:

19
20
21 <http://www.uff.br> <http://purl.org/dc/elements/1.1/creator> <https://orcid.org/0000-0003-
22
23 0929-8475>
24
25
26

27 In this example, the original predicate “author” is replaced by the URI referenced by
28
29 the “creator” element of the well-known Dublin Core (DC) metadata standard. In its context,
30
31 dc:creator has a specific semantics. It is defined as “An entity responsible for making the
32
33 resource.” The triple’s object, the value or content of dc:creator, has been replaced by the
34
35 Open Researcher and Contributor ID (ORCID), <https://orcid.org>, of the page’s author.
36
37
38
39

40 It is with the semantics in specific vocabularies that the limited semantic
41
42 expressiveness of the RDF model can be expanded, as seen in the example of the CRF. Once
43
44 specified in elements of a vocabulary, the semantics can be processed by programs. While
45
46 the features provided in the Web of Data, represented in markup languages such as XML,
47
48 RDF, HTML, etc. are contents, programs are procedures or algorithms according to the data
49
50 processing model. Programs only know how to process content. For this, they need to be
51
52 clearly instructed (programmed) on what to do with certain content in a certain situation.
53
54 LOV used to assign semantics to LOD (Zeng, 2019) must clearly define, restrict, and specify
55
56 the semantics of their concepts. For example, the DC metadata vocabulary clearly defines
57
58
59
60

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 Webspaces, 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 predicates

<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 >	<President> .
>	/title>	

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

1
2
3 meaning of “title” but to this meaning within the scope (“namespaces”) of the DC and
4
5 Vcard, <https://devguide.calconnect.org/vCard/vcard-4/>, vocabularies.
6
7

8 This allows programs to do more than just process inferences about a graph, a set of
9
10 RDF triples. These are programs oriented by ontology or models, such as Application
11
12 Program Interfaces (APIs) from the Europeana Library, <https://pro.europeana.eu/page/apis>.
13
14

15 16 17 3.2. Functionalities for vocabularies to be used to assign semantics to data within the context 18 19 of the Semantic Web and LOD 20

21
22
23 Through unique and persistent identifiers, metadata and data vocabularies can be
24
25 used to assign machine-understandable semantics to predicates and objects in triples RDF.
26
27 Many old vocabularies are being restructured to be compatible with LOD technologies
28
29 (Soergel, 2004; Dos Santos Maculan, 2015), such as the UNESCO Thesaurus,
30
31 <http://vocabularies.unesco.org/browser/thesaurus/en/>, the FAO Thesaurus,
32
33 http://aims.fao.org/aos/agrovoc/c_8003.html, the AGROVOC Thesaurus,
34
35 <https://agrovoc.fao.org/browse/agrovoc/en/>, the Paul Getty Foundation Vocabularies,
36
37 <https://www.getty.edu/research/tools/vocabularies/lod/>, the Art and Architecture Thesaurus,
38
39 the Union List of Artists Names, the Cultural Objects Name Authority, the Getty Thesaurus
40
41 of Geographic Names, the DeCS/MeSH, Health Science Descriptors,
42
43 <https://decs.bvsalud.org/th/>, the Library of Congress Subject Headings (LCSH),
44
45 <https://id.loc.gov/authorities/subjects.html>, in addition to many others.
46
47
48
49
50
51
52
53

54 Vocabularies used with LOD need to meet requirements such as having their
55
56 concepts persistently and univocally identified through valid URIs on the internet, being
57
58 represented in machine-readable formats such as RDF, containing precise definitions of the
59
60

1
2
3 semantics of their concepts, and generally, being multilingual. Many of these vocabularies
4 that meet the principles of LOD can be found in the aforementioned LOV vocabulary
5 registry service. By meeting the requirements for use with LOD as described above,
6 vocabularies, an area of study, research, and practical use of KO, can contribute to
7 addressing the issues brought about by Big Data.
8
9
10
11
12
13
14
15
16

17 Elements of data or metadata vocabularies referenced by URI account for the
18 semantics of an individual datum, an element of a triple, the “datum” according to Hjørland
19 (2018). An example is the fields of the CRF form. These vocabularies use different
20 approaches to semantics, as pointed out in Almeida et al (2011, p. 195), ranging from a
21 semantics for humans, that is implicit, informal or formal, to semantics for machines, that is
22 informal, formal or even “powerful semantics.” In any case, used in the context of the RDF
23 model these vocabularies already allow the processing of RDF triples by machines.
24
25
26
27
28
29
30
31
32
33
34

35 3.3. Semantics beyond the data. 36 37 38 39

40 The concept of “powerful semantics,” originally devised by Shet, Ramakrishnan, and
41 Thomas (2005) and developed in Shet (2020, slide 42), is defined as “statistical analysis
42 [that] allows the exploration of relationships that are not stated.” Semantics is obtained from
43 statistical patterns, not from individual datum referenced by metadata describing an entity,
44 but rather from data sets, or Big Data. Naturally, to identify this semantics, Big Data,
45 whether structured or unstructured, has to be processed by programs. This is so-called data
46 science (Dhar, 2013).
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 Other vocabularies also have emerged, not to describe or provide standardized values
4
5 for each piece of data, but to provide descriptive and value metadata of the datasets as a
6
7 whole. Digital curation is an emerging field of activity for KO professionals.see
8
9
10 <https://www.dcc.ac.uk/>. For the curation of these datasets, metadata standards such as Data
11
12 Catalog Vocabulary (DCAT) <https://www.w3.org/TR/vocab-dcat-2/>, or the Provenance
13
14 Ontology (PROV-O) <https://www.w3.org/TR/prov-o/>, have been adopted to describe the
15
16 provenance of the dataset. As datasets have been made available as informational resources
17
18 on the Web, information on their provenance and the record of the processing carried out on
19
20 them, the extract, transform, load (ETL) processes (See
21
22 https://en.wikipedia.org/wiki/Extract,_transform,_load) are essential elements for the data to
23
24 have credibility and to be able to be reused (See <https://www.go-fair.org/fair-principles/>).

25
26
27
28
29
30
31 These datasets, in addition to the metadata that describe their fields, which describe
32
33 the entity represented by the dataset, have additional metadata provided by vocabularies
34
35 such as DCAT and PROV-O for the dataset as a whole. For example, they contain metadata
36
37 such as its format, its quantity, its update date, licenses to use this dataset, etc. (from
38
39 DCAT), or metadata such as the entity (in this case, the dataset for which the provenance is
40
41 to be registered), the agent that created the dataset, and the process that generated it (from
42
43 PROV-O). Standards such as these have been used in several research data repositories to
44
45 index the files deposited there, an increasingly common application by KO professionals.
46
47
48
49
50

51 3.4. Ontologies as domain models, definitions, specifications

52
53
54 The language specification OWL – Ontology Web Language Overview (2004) states
55
56 that:
57
58
59
60

1
2
3 OWL can be used to explicitly represent the meaning of terms in
4 vocabularies and the relationships between those terms. This representation
5 of terms and their interrelationships is called an ontology. OWL has more
6 facilities for expressing meaning and semantics than XML, RDF, and RDF-
7 S, and thus OWL goes beyond these languages in its ability to represent
8 machine interpretable content on the Web.
9
10
11
12
13
14
15

16 OWL then is a standard language (meta-language in the aforementioned sense) of the
17 W3C for representing ontologies, that is, vocabularies that specify the things existing in a
18 domain and their interrelationships. Further on, the same specification compares the
19 semantic expressiveness of OWL with that of other languages to represent machine-
20 interpretable content such as XML, XML Schema, RDF, and RDFS (ONTOLOGY WEB
21 LANGUAGE OVERVIEW, 2004). It can thus be concluded that, with current technologies,
22 a computational ontology developed in OWL is the most expressive type of KOS, because
23 the “facilities” provided by OWL allow restricting, specifying, and expressing the intended
24 meaning – “intended meaning –” (Guarino, 1994, 560) of the conceptual model of a domain
25 obtained by the modeling process.
26
27
28
29
30
31
32
33
34
35
36
37
38
39

40 Among these facilities is the possibility of specifying data properties (attributes, in
41 Chen's ER model), object properties (relationships in Chen's ER model), domain and scope
42 of the two types of properties, cardinality constraints of each class involved in an object
43 property, transitivity and reflexivity of properties, the disjunction between individuals of
44 different classes, axioms for restricting the inclusion of instances in a class (ONTOLOGY
45 WEB LANGUAGE OVERVIEW, 2004), etc. These facilities can make the conceptual
46 models implicit in a computational ontology in OWL more faithful to reality.
47
48
49
50
51
52
53
54
55
56

57 As seen earlier, the Web of Data project, the large-scale reuse of Big Data available
58 in increasing amounts on the Web, depends on the one hand on the most expressive
59
60

1
2
3 vocabularies that describe them, and on the other hand on programs capable of make
4
5 inferences, or at least algorithmic processing, on these representations. In this context,
6
7 specific domain models, intelligible by machines and represented with the maximum
8
9 possible semantic expressiveness, gain importance, which, in the current stage of
10
11 technology, are computational ontologies.
12
13
14

15 Another important aspect related to this issue; Bergman (2011) discusses ODapps:
16
17 The Ontology-Driven Application Approach, an automatic program development
18
19 methodology based heavily on ontologies, a set of them, from high-level ontologies, task
20
21 ontologies, domain ontologies, to specific application ontologies (Guarino, 1997, 145). In
22
23 the context of ODapps, domain computational ontologies, with a high degree of semantic
24
25 expressiveness, are an essential component for developing generic application programs,
26
27 capable of processing, making inferences, discovering, and reusing the knowledge contained
28
29 in the domain representation.
30
31
32
33

34
35 It is, therefore, necessary for KO to advance in the creation of computational
36
37 ontologies of specific domains that are increasingly semantically expressive to equip
38
39 programs capable of processing these representations to make inferences about them and
40
41 extract and reuse the knowledge contained therein. The research on patterns of definitions
42
43 for concepts in ontologies (Campos, 2010) plays a fundamental role in the specification of
44
45 machine-intelligible semantics, developing the proposals of Dahllberg (1978) of a typology
46
47 of definitions; just as issues of interoperability between concepts of different ontologies
48
49 (Barbosa and Campos, 2017), as suggested in Standard 25964-2 (2013), in the SKOS
50
51 standard (2012) and Zeng (2019).
52
53
54
55

56 **4. Final considerations**

57
58
59
60

1
2
3 Issues involving information technologies are obscured by the metaphorical
4 denominations often adopted that, didactically and scientifically, make it difficult to
5 understand and operate them, such as Big Data and the Semantic Web. For an accurate
6 understanding of current information technologies, the semantic capacity of computers has
7 to be analyzed, understood, and the real potential identified.
8
9
10
11
12
13
14

15 This article sought to demonstrate that data, which have a semiotic and ontological
16 character and are artificial and intentional representations, cannot be understood apart from
17 the entity to which they refer and from the metadata—the properties of this entity—that
18 describe it. Unspecified data is also an imprecise concept. It is necessary to distinguish one
19 piece of datum as referred to by Hjørland (2018), which is a unit that represents the value of
20 one (of the) properties of an entity, from a record, a set of several datum describing various
21 properties of an entity, from datasets, representing the various entities and their properties,
22 and from databases, bringing together different datasets representing different interrelated
23 entities. The datum as a unit has its own semantics, but records, datasets, and databases
24 already have other levels of semantics in the computational environment.
25
26
27
28
29
30
31
32
33
34
35
36
37
38

39 The Web of Data technologies bring a significant advance by incorporating more
40 semantic expressiveness and program independence to data published on the Web according
41 to the RDF model. In this model, vocabularies can play a significant role, as has been
42 suggested. There are however, several levels of semantics in the variety and heterogeneity of
43 data published on the Web: the “powerful” semantics of the different datasets (the semantic
44 expressiveness of the aggregated datasets of other data), the semantic expressiveness
45 embedded in textual Big Data which needs to be processed for the identification of entity
46 names, named-entity recognition (NER) (Freitas et al 2010), for aggregating annotations
47 and making this data structured, the semantic expressiveness given by programs according
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 to the data processing model (for data being processed one way and not another), etc. A
4
5 systematization of these issues should be included in the KO research agenda.
6
7

8 9 **References**

10
11
12 Ameida, Mauricio; Souza, Renato and Fonseca, Fred. 2011. "Semantics in the Semantic
13
14 Web: A Critical Evaluation". *Knowledge Organization*, 38(3):187-203.
15
16 <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1041.7976&rep=rep1&type=pdf>,
17
18 accessed 25 Mar 2021.
19
20

21
22
23 Aristóteles. *Categorias*. Porto: Porto Editora Ltda, 1995.
24
25

26
27
28 Barbosa, Nilson. T. and ; CAMPOS, Maria. L. de Almeida. 2017. "A questão da
29
30 interoperabilidade em repositórios institucionais e sistemas de informação de pesquisas
31
32 correntes (cris): uma abordagem preliminary". In *Encontro Nacional de Pesquisa em Ciência*
33
34 *da Informação, n. XVIII ENANCIB, 2017*. <http://hdl.handle.net/20.500.11959/brapci/104600>,
35
36 accessed 25 Dez. 2021.
37
38

39
40
41
42 Bergman, Mike. 2011. "Ontology-Driven Apps Using Generic Applications". *AI3 blog*.
43
44 <https://www.mkbergman.com/948/ontology-driven-apps-using-generic-applications/>.
45
46

47
48
49 Berners-Lee, Tim. 1998. "Cool URIs don't change".
50
51 <https://www.w3.org/Provider/Style/URI>.
52
53
54
55
56
57
58
59
60

1
2
3 Cabré, María Teresa. 2005. A Terminologia, uma disciplina em evolução: passado, presente
4 e alguns elementos de futuro. *Debate Terminológico*. ISSN: 1813-1867, v1.
5
6 <https://www.seer.ufrgs.br/riterm/article/download/21286/15349>, accessed 21 Set. 2020.
7
8
9

10
11
12 Campos, Maria Luiza de Almeida. 2010. “O papel das definições na pesquisa em
13 ontologia”. *Perspectivas em Ciência da Informação*, 15: 220-238
14
15 <https://www.scielo.br/j/pci/a/tJr4GnX9Xp7pj5pf44gK4yD/?lang=pt&format=html>.
16
17
18

19
20
21 Chierchia, Gennaro. 2003. *Semântica*. São Paulo: Ed. UNICAMP.
22

23
24
25
26 CIDOC Conceptual Reference Model Version 5.1.12. 2014. ICOM/CIDOC.
27 <http://www.cidoc-crm.org/Version/version-5.1.2>, accessed May 3, 2015.
28
29

30
31
32 Codd, Eugene. F. 1970. “A relational model of data for large shared databanks”.
33
34 *Communications of The ACM*, 13(6): 377-387.
35
36 https://dl.acm.org/doi/pdf/10.1145/362384.362685?casa_token=uOdxFTaktMAAAAAA:i_e
37
38 [wo3eO7rDNRE7VYvIBGeHn452O1VQGi69Jn13MciziUeGNMPy827WA6guuZzLkgq4D](https://dl.acm.org/doi/pdf/10.1145/362384.362685?casa_token=uOdxFTaktMAAAAAA:i_e)
39
40
41 [GI79ocfO4A](https://dl.acm.org/doi/pdf/10.1145/362384.362685?casa_token=uOdxFTaktMAAAAAA:i_e).
42
43

44
45 Dahlberg, Ingetraut. 1978. “A referent-oriented, analytical concept theory for
46 INTERCONCEPT”. *KO KNOWLEDGE ORGANIZATION*, 5(3): 142-151
47
48 https://www.ergon-verlag.de/isko_ko/downloads/ic_5_1978_3.pdf#page=20.
49
50

51
52 Dhar, Vasant. 2013. “Data science and prediction”. *Communications of the ACM*, 56(12):.
53
54 64-73. <https://dl.acm.org/doi/pdf/10.1145/2500499>.
55
56
57
58
59
60

1
2
3 Dextre Clarke, Stella G. 2019. "The Information Retrieval Thesaurus". *KNOWLEDGE*
4 *ORGANIZATION*, 46(6): 439-459. [https://www.ergon-](https://www.ergon-verlag.de/isko_ko/downloads/ko_46_2019_6_c.pdf)
5 [verlag.de/isko_ko/downloads/ko_46_2019_6_c.pdf](https://www.ergon-verlag.de/isko_ko/downloads/ko_46_2019_6_c.pdf).
6
7

8
9
10
11 Dextre Clarke, Stella G. and Zeng, Marcia Lei. 2012. "From ISO 2788 to ISO 25964: The
12 evolution of thesaurus standards towards interoperability and data modelling". *Information*
13 *Standards Quarterly (ISQ)*, 24(1).
14
15 http://eprints.rclis.org/16818/1/SP_clarke_zeng_isqv24no1.pdf.
16
17

18
19
20
21 Dierickx, Harold and Hopkinson, Alan. 1986. *Reference manual for machine-readable*
22 *bibliographic descriptions*.
23
24 http://biblio.cerist.dz/hrbdonf5214/ouvrages/0000000000000594806000000_2.pdf.
25
26

27
28
29 FAIR Compliant Biomedical Metadata Templates. 2019. CEDAR, Center for Expanded
30 Annotation and Retrieval, University of Stanford, Department of Medicine.
31
32 <https://medicine.stanford.edu/2019-report/cedar-to-the-rescue.html>.
33
34

35
36
37 Floridi, Luciano. 2019. "Semantic Conceptions of Information". In *The Stanford*
38 *Encyclopedia of Philosophy* (Winter 2019 Edition), Edward N. Zalta (ed.).
39
40 <https://plato.stanford.edu/archives/win2019/entries/information-semantic/>.
41
42

43
44
45 Freitas, C.; Carvalho, P.; Oliveira, H. G.; Mota, C. and Santos, D. 2010. "Second HAREM:
46 advancing the state of the art of named entity recognition in Portuguese". In Nicoletta
47 Calzolari et al. (eds.), *Proceedings of the International Conference on Language Resources*
48 *and Evaluation (LREC 2010)*. European Language Resources Association, pp. 3630-3637.
49
50
51 Valletta, 2010.
52
53
54
55
56
57
58
59
60

1
2
3 Giunchiglia, Fausto; Dutta, Biswanath and Maltese, Vincenzo. 2014. "From knowledge
4 organization to knowledge representation". *KNOWLEDGE ORGANIZATION*, 41(1): 44-56,
5
6 2014. <http://eprints.biblio.unitn.it/4186/1/techRep027.pdf>.
7
8
9

10
11 Guarino, Nicola. 1997. "Semantic matching: Formal ontological distinctions for information
12 organization, extraction, and integration". In *International Summer School on Information*
13 *Extraction*. Springer, Berlin, Heidelberg, 1997. 139-170.
14
15 [https://kask.eti.pg.gda.pl/redmine/projects/sova/repository/revisions/5378040326bc499e118](https://kask.eti.pg.gda.pl/redmine/projects/sova/repository/revisions/5378040326bc499e118636a1d25ad667285e005c/entry/Praca_dyplomowa/materialy/10.1.1.53.939.pdf)
16
17 [636a1d25ad667285e005c/entry/Praca_dyplomowa/materialy/10.1.1.53.939.pdf](https://kask.eti.pg.gda.pl/redmine/projects/sova/repository/revisions/5378040326bc499e118636a1d25ad667285e005c/entry/Praca_dyplomowa/materialy/10.1.1.53.939.pdf).
18
19
20
21
22

23 Guarino, Nicola; Carrara, Massimiliano an Giaretta, Pierdaniele. 1994. "Formalizing
24 ontological commitment". In *AAAI*. 1994. p. 560-567.
25
26 <https://www.aaai.org/Papers/AAAI/1994/AAAI94-085.pdf>.
27
28
29
30

31 Hey, Tony; Trefethen, Anne. 2003. "The data deluge: An e-science perspective". In *Grid*
32 *computing: Making the global infrastructure a reality*, p. 809-824.
33
34 https://eprints.soton.ac.uk/257648/1/The_Data_Deluge.pdf.
35
36
37
38
39
40

41 Hjørland, Birger. (2018). "Data (with big data and database semantics)". *Knowledge*
42 *Organization*, 45(8): 685-708.
43
44
45

46 Hjørland, Birger. (2002). "Domain analysis in information science: eleven approaches–
47 traditional as well as innovative". *Journal of Documentation*, 58(4), 422-462.
48
49
50

51 Hjørland, Birger, and Albrechtsen, Hanne. (1995). "Toward a new horizon in information
52 science: Domain-analysis". *Journal of the American society for information science*, 46(6),
53
54 400-425.
55
56
57
58
59
60

1
2
3 Hjørland, Birger and Hartel, Jenna. 2003. "Introduction to a special issue of Knowledge
4 Organization". *Knowledge Organization*, 30(3/4), 125-7.
5
6
7

8
9 International Council on Archives. Experts Group on Archival Description. 2019. Records in
10 Context: A Conceptual Model for Archival Description (Consultation Draft v0.1). ICA.
11 https://www.ica.org/sites/default/files/ric-cm-0.2_preview.pdf, accessed December 12, 2018.
12
13
14

15
16 International Federation of Library Associations and Institutions (IFLA). 1998. *Study Group*
17 *on Functional Requirements for Bibliographic Records: Final Report*. UBCIM Publications
18 New Series. München: K. G. Saur.
19
20
21
22

23
24
25
26 ISO/IEC 20546:2019(en). Information technology — Big data — Overview and vocabulary.
27 ISO, 2019.
28
29

30
31 ISO 25964-2 - Information and documentation — Thesauri and interoperability with other vocabularies
32 — Part 2: Interoperability with other vocabularies. ISO, 2013.
33
34
35

36
37
38 Lambe, Patrick. 2014. *Organising knowledge: taxonomies, knowledge and organisational*
39 *effectiveness*. Elsevier.
40
41
42

43
44
45 Marcondes, Carlos H. and Costa, Leonardo C. da. 2016. "A Model to Represent and
46 Process Scientific Knowledge in Biomedical Articles with Semantic Web Technologies".
47 *Knowledge Organization*, 43(2): 122-137. [https://www.ergon-](https://www.ergon-verlag.de/isko_ko/downloads/ko_43_2016_2_b.pdf)
48 [verlag.de/isko_ko/downloads/ko_43_2016_2_b.pdf](https://www.ergon-verlag.de/isko_ko/downloads/ko_43_2016_2_b.pdf), accessed Apr. 12, 2017.
49
50
51
52

53
54
55 Marcondes, Carlos H. and Dias, Celia. 2020. "Representing facet classification in SKOS".
56 In International ISKO Conference, Aalborg, Denmark, 16th, *Proceedings...*
57 *1. Edition*. Würzburg: Ergon Verlag. ISBN print: 978-3-95650-775-5, ISBN online: 978-3-
58
59
60

1
2
3 95650-776-2, *Series: Advances in knowledge organization* 9. Würzburg: Ergon
4 Verlag, 254–263. <https://doi.org/10.5771/9783956507762>, accessed Fev. 15, 2021.
5
6
7

8
9
10 Marcondes, Carlo. H.; Martins, Sergio. C. and Ramos Junior, Mauricio. C. 2021. The role of
11 vocabularies for the access and reuse of Big Data. *Informação & Informação*, 26(4): 146-
12 174. <https://www.uel.br/revistas/uel/index.php/informacao/article/view/44653/pdf>. Access 5
13
14
15
16
17 Jan. 2022.
18

19
20
21
22
23 De Mauro, Andrea; Greco, Marco and Grimaldi, Michele. 2015. “What is big data? A
24 consensual definition and a review of key research topics”. In *AIP conference proceedings*.
25 American Institute of Physics, 2015. p. 97-104. [http://big-data-fr.com/wp-](http://big-data-fr.com/wp-content/uploads/2015/02/aip-scitation-what-is-bigdata.pdf)
26
27
28
29
30
31
32 content/uploads/2015/02/aip-scitation-what-is-bigdata.pdf.

33 Mylopoulos, John. 1992. “Conceptual modelling and Telos”. In *Conceptual modelling,*
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
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.

61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

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

1
2
3 Prieto-Díaz, Ruben. 1990. "Domain analysis: An introduction". *ACM SIGSOFT Software*
4
5 *Engineering Notes*, 15(2): 47-54.

6
7
8 RDF semantics. W3C, 2004. <http://www.w3.org/TR/rdf-mt/>, accessed Mar, 10, 2010.

9
10
11
12
13
14 Ranganathan, S. R. and Gopinath, M. A. *Prolegomena to Library Classification*. 3 ed.
15
16 Bombay: Asia Publishing House, 1967.

17
18
19
20 RDF 1.1. PRIMER. 2014. W3C. <https://www.w3.org/TR/rdf11-primer/>, accessed 12 Dez.
21
22 2019.

23
24
25 Resource Description Framework (RDF) Model and Syntax Specification. W3C, 1998.
26
27 <https://www.w3.org/1998/10/WD-rdf-syntax-19981008/>. Accessed May 5, 2011.

28
29
30
31 Saracevic, Tefko. 2007. "Relevance: A review of the literature and a framework for thinking
32
33 on the notion in information science. Part II: Nature and manifestations of
34
35 relevance". *Journal of the american society for information science and technology*, 58(13):
36
37 1915-1933.

38
39
40
41 Shet, Amith. 2020. "Knowledge Graphs and their central role in big data processing: Past,
42
43 Present, and Future". In 7th ACM India Joint Conference on Data Science & management of
44
45 Data (COD-COMAD), Indian School of Business, Hyderabad Campus, 5-7 January 2020.
46
47 [https://www.slideshare.net/apsheth/knowledge-graphs-and-their-central-role-in-big-data-](https://www.slideshare.net/apsheth/knowledge-graphs-and-their-central-role-in-big-data-processing-past-present-and-future)
48
49 [processing-past-present-and-future](https://www.slideshare.net/apsheth/knowledge-graphs-and-their-central-role-in-big-data-processing-past-present-and-future), accessed Jun. 5, 2021.

50
51
52
53
54 Shet, Amith; Ramakrishnan, Cartic and Thomas, Christopher. 2005. "Semantics for the
55
56 semantic web: The implicit, the formal and the powerful". *International Journal on*
57
58 *Semantic Web and Information Systems (IJSWIS)*, 1(1): 1-18.

1
2
3 <http://www.ebusinessforum.gr/old/content/downloads/JSWIS.pdf#page=19>, accessed Jul 14,
4
5 2010.

6
7
8
9
10 SKOS – Simple Knowledge Organization System Namespace Document. W3C, 2012.
11
12 <https://www.w3.org/2009/08/skos-reference/skos.html#>, accessed Aug 10, 2013.

13
14
15
16
17 SPARQL 1.1 QUERY LANGUAGE, 2013. W3C. <https://www.w3.org/TR/sparql11-query/>,
18
19 accessed 12 Feb. 2010.

20
21
22
23
24 Veiga, Viviane Santo de Oliveira; Campos, Maria Luiza; Silva, Carlos Roberto Lyra;
25
26 Henning, Patricia and Moreira, João. 2021. “Vodan br: a gestão de dados no enfrentamento
27
28 da pandemia coronavirus”. *Páginas A&B, Arquivos e Bibliotecas (Portugal)*, n. Especial:
29
30 51-58. <http://hdl.handle.net/20.500.11959/brapci/157353>, accessed Oct 7, 2021.

31
32
33
34
35 Wilson, Thomas. D. 1972. “The work of the British Classification Research Group”. In
36
37 Wellish, H. (ed). *Subject retrieval in the seventies*. Westport: Greenwood Publishing Co.,
38
39 62-71.

40
41
42
43
44 Zeng, Marcia Lei. 2019. “Interoperability”. In Hjørland, Birger and Gnoli, Claudio eds. *ISKO*
45
46 *Encyclopedia of Knowledge Organization*. ISKO. <http://www.isko.org/cyclo/interoperability>,
47
48 accessed Jun 4, 2020.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

For Review Only

1
2
3 KO-2022-0003 – Answers to the Reviewers comments
4
5

6 07-Aug-2022
7

8 Dear Reviewers
9

10 Thank you for your valuable comments to our text. We tried to apply them to improve the
11 revised version the text.
12

13 The paper was rewritten with focus on digital research data within the context of big data.
14 Text excluded was highlighted in yellow, text added was highlighted red.
15
16

17 Reviewer(s)' Comments to Author:
18

19 Reviewer: 1
20

21 Comments to the Author
22

23
24 This is a very unique paper, with great connections between big data and knowledge
25 organization. Good that the authors have solid background of Linked Data and semantic
26 technologies. I have a few comments and suggestions and hope the authors will be able to
27 accommodate, turning into a better one that is well-connected to the main trends in big
28 data.
29

30
31 1. The title, "THE ROLE OF VOCABULARIES IN BIG DATA: the quest of research
32 data," should be reconsidered. Instead of a general term 'VOCABULARIES', it should be
33 more specific, especially addressing that these are knowledge organization systems (KOS),
34 not normal dictionaries, thesauri (those not used as a controlled vocabulary), or
35 vocabularies. **Answer: The text is discussing vocabularies, i.e. value, data (for example,
36 subject) and metadata, descriptive vocabularies (Zeng 2019).**
37

38 This term is also unclearly used in the text of the paper, e.g., sometime it is clear you are
39 talking about KOS, but you used this term again (1.2. Traditional use of vocabularies to
40 assign subjects to documents).
41

42
43 2. When talking about "Next, we will attempt to demonstrate how the conceptualization
44 above helps address the issues of Big Data, especially research data", it is not clear that if
45 you already considered research data as big data. Please notice the 5Vs. Big data has
46 been characterized by multiple "V"s, with the number of "V"s still increasing. Volume (data
47 quantity), Velocity (data speed), Variety (data types and nature), Variability (data
48 consistency), and Veracity data quality) (Kobielus, 2016; Zeng 1017). Any data that have
49 been processed cannot be considered as part of 'big data'. Even OCLC would not consider
50 its huge data as big data. So, do not just think the 'volume' (or amount' to be the feature of
51 big data (refer to your statement at the beginning of 1.1. The Big Data: "Big Data, the term
52 for a recent phenomenon describing the amount of data produced in digital
53 format.") **Answer: We consider, like Shiri (2013, 18), that Big Data is made up of
54 research data, open data, linked data and semantic. Today's research data has also
55 the characteristics of Big Data (Fillinger et al. 2019). The same is also stressed in the
56 National Institutes of Health Core Techniques and Technologies for Advancing Big
57 Data Science & Engineering (BIGDATA) report (Shiri 2013, 17). Consider the different
58
59
60**

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.

1
2
3 IEEE Smart Data conferences. CFPs, etc.
4

5 Kobielus, J. (2016, June). The evolution of big data to smart data [PowerPoint slides].
6 Keynote at Smart Data Online 2016. Video available at [https://www.dataversity.net/big-](https://www.dataversity.net/big-data-smart-data-big-drivers-smart-decision-making/)
7 [data-smart-data-big-drivers-smart-decision-making/](https://www.dataversity.net/big-data-smart-data-big-drivers-smart-decision-making/)
8
9

10 Schöch, C. (2013). Big? Smart? Clean? Messy? Data in the humanities. *Journal of Digital*
11 *Humanities*, 2(3), 13.
12

13 Sheth, A. (2014, March). Transforming Big Data into Smart Data: Deriving Value via
14 harnessing
15 Volume, Variety and Velocity using semantics and Semantic Web [Keynote address]. 30th
16 IEEE
17 International Conference on Data Engineering, Chicago, IL, United States.
18 DOI:10.1109/ICDE.2014.6816634
19
20

21 Zeng, M. L. (2017). Smart data for digital humanities. *Journal of Data and Information*
22 *Science*. 2(1), 1-12. DOI: 10.1515/jdis-2017-0001
23
24
25
26

27 Reviewer: 2
28

29 Comments to the Author
30

31 The topic of the paper is relevant for the journal. Unfortunately, as it stands now, it does not
32 seem to meet necessary quality criteria. CLARITY/LANGUAGE -- The text of the paper is
33 difficult to follow due to the wrong usage / choice of words and other language issues.
34 Occasionally concepts are linked to actions that they cannot perform or are given
35 properties that they cannot have. Therefore the meaning is not entirely clear or sentences
36 do not make sense. An example is in the subtitle itself "the quest of research data" [who is
37 doing the quest here?] COMPOSITION: -- The authors occasionally digress into explaining
38 common knowledge or technical or historical details which are not relevant for the topic of
39 this paper and do not contribute to the argument they are trying to make. RESEARCH
40 BACKGROUND AND CONTEXT -- The proposal outlined in this paper are not novel -
41 therefore it needs to be put into the context of the research already published on this very
42 topic rather than providing history prior to the semantic web and big data phenomena.
43
44

45 **Answer: We are interested in showing in detail how value and metadata vocabularies work**
46 **within LOD technologies to assign meaning to data, section 3.1. The history prior to the**
47 **semantic web and big data phenomena is necessary to review previous KO initiatives**
48 **concerning how to represent things within a domain, section 2.** There is a lack of
49 references to the literature discussing relationship between Big Data and KO (e.g. Ibekwe-
50 SanJuan & Geoffrey, Kwak, Hajibayova & Salaba, Shiri, Bauer, DeMauro) or KOS/LOD
51 semantic web technology (specifically Zeng, Zeng & Mayr, Busch, Stellato, Isaac, Mendez
52 & Greenberg, etc.) **Answer: Now most of such references are cited**
53
54

55 FURTHER SPECIFIC ISSUES NEED ADDRESSING:
56

57 **1-** ABSTRACT is poorly structured, incomprehensible and would need to be rewritten. The
58 section 'Methodology' does not explain methodology. It contains statements that make no
59 sense. Examples: "materials used are definitions of Big data found in... technologies used
60 in the Semantic Web and Linked Open Data" [technologies do not contain definitions] ---

1
2
3 How does the citation at the bottom of the abstract relate to the abstract or to this
4 research? The paragraph is not properly cited/referenced, it requires a proper reference
5 and page from which the citation is taken, if this is attributed to a person then it would work
6 better if incorporated somehow in the text). This citation should be placed on Page 4 (see L
7 37). **Answer: The abstract has been rewritten.**

8
9 **2-** “We present a comprehensive conceptualization of **the concept of** semantic expressivity
10 and use it to classify the different vocabularies.” --- [What does this mean, how and why
11 one uses conceptualization of semantic expressivity to classify different vocabularies?
12 Which vocabularies? What this has to do with Big Data?]. **Answer: We excluded such**
13 **claim. We used the concept of semantic expressivity by Almeida, Souza and Fonseca**
14 **(2011) to compare different vocabularies and their capacity to represent details of a**
15 **domain with greater accuracy** --- “We identify computational ontologies as a type of
16 knowledge organization system with a higher degree of semantic expressivity. It is
17 suggested that such themes should be incorporated into professional qualifications in KO.”
18 --- [How does the statement in the first sentence, which is common knowledge, become a
19 ‘theme’ and what is meant by “incorporating this into professional qualification”]?
20
21
22

23
24 **3-** PAGE 6 - The authors state (Page 6) that their objective is to show : “how KO can
25 contribute to assigning computational semantics to Big Data, especially to research data,
26 so that computers can process them, allowing their reuse on a large scale”. **Answer: We**
27 **changed the focus and objective to “The objective of this work is to discuss how**
28 **vocabularies, in the sense used within LOD Technologies i.e., value and metadata**
29 **vocabularies (Zeng 2019), can contribute to assigning computational semantics digital**
30 **research data within the context of Big Data, so that computers can process them,**
31 **allowing their reuse on large scale”**. After describing and explaining well-known
32 characteristics of semantic technology and linked data and associated ontology standards
33 the authors conclusion is that KOS should be made available in the machine-
34 understandable formats i.e. as ontologies using OWL (or some other similar formal
35 ontology language). This has been widely accepted and considered a norm in the KO
36 domain for the past two decades, need not illustration and **cannot be represented as a**
37 **finding**. Obviously, once available for machine processing KOS can be utilized for any kind
38 of automatic data processing and linking notwithstanding Big Data. Another assertion that
39 the authors wanted to demonstrate is that data cannot be understood without the context
40 i.e. semantics provided through metadata – which is again stating the obvious (data is not
41 information!). **Answer: We are not only talking about metadata but also about the entity**
42 **described by the metadata and data**. The authors should consider explaining the context of
43 their research and its focus starting from the above mentioned well known facts.
44
45
46
47

48
49 **4-** BIBLIOGRAPHIC DATA - The way the authors deal with the topic of metadata principles
50 and bibliographic control, KO and KOS shows some lack of knowledge and it would be
51 better to avoid elaborating these topics in great length. E.g. in section 3, there is an
52 irrelevant historical overview which only demonstrates the lack of understanding of the
53 typology and nature of bibliographic standards. For instance, UNISIST Reference Model is
54 placed in the same category MARC (metadata format schema) **Answer: We rewrote such**
55 **comparison, comparing the MARC format with the UNISIST reference manual for machine-**
56 **readable bibliographic descriptions (Dierickx and Hopkinson 1985) and ISBD and ACCR**
57 **ISBD, ACCR, MARC and UNISIST are collectively called “standards”** which are
58 metadata/cataloguing description standards. There is, for instance, an entire paragraph on
59 MARC data format development on page 12 – in continuation of the authors’ explanation
60

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 components **subsystems**', 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](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.

1
2
3 PAGE 3 "Although Big Data has been sparking interest in KO, contribution from the area to
4 contextualise it or to propose practical solutions are still few" --- [Unlikely - needs to be
5 supported by reference. **Interest by who? Is there any evidence that Big Data has been**
6 **sparkling interest in KO** when it comes to the domain of computer technology which is the
7 one predominantly dealing with the Big Data issue. The authors addressing this topic come
8 primarily from the KO domain] **Answer: This text has been rewritten.**

9
10
11 PAGE 4 "... we cannot address Big Data without using computers to help us. This
12 observation refers to the Semantic Web --- [Stating the obvious - Big Data is produced by
13 computers, hence the role of computers is self-evident. How does this observation relate to
14 Semantic Web?] **Answer: This text is deleted.**

15
16
17 PAGE 5 - "KO methodologies have always represented domains of knowledge when
18 building KOS like controlled /standardized vocabularies, subject headings and classification
19 schemas. The early KOS, such as thesauri --- [why KOS-like, SH and classification are
20 KOS? Not all KO-associated methodologies are dealing with domain knowledge. Thesauri
21 are not "early KOS"] **Answer: This text is deleted.**

22
23
24 PAGE 5 L48: "empower "knowledge," which is no longer just inserted into texts to
25 be interpreted by humans, but rather recorded directly in Resource Description Framework"
26 --- [knowledge is not inserted in to texts. One does not 'record' in RDF - RDF is a data
27 representation model.] **Answer: The text has been replaced by "knowledge (Soergel**
28 **2015). It is no longer just inserted into texts to be interpreted by humans, but rather**
29 **serialized in Resource Description Framework (RDF) triples".**

30
31
32 PAGE 13: " Significantly, bibliographic formats evolved separately from the also nascent
33 technology of computer databases that, from the 1970s onwards, had the relational model
34 as a paradigm (Codd, 1970). **The Text Retrieval Conferences (TREC) conference series**
35 **illustrates this separate evolution. -> Answer: This text is deleted.** --- [Irrelevant as well as
36 illogical/wrong series of statements. What do relational databases or any type of database
37 technology have to do with data schemas and formats and their evolution? Although
38 irrelevant for this argument MARC format which is a data element schema / format is used
39 in bibliographic databases that are predominantly relational - but not obligatory. The only
40 thing that is relevant to state in this context is that information retrieval in bibliographic
41 systems is metadata based (irrespective of the type of databases technology used an
42 irrespective of data exchange formats). The text retrieval approach, on the other hand, is
43 not metadata based and depends on availability of digital texts and (advanced) computer
44 based text processing/IR methods. TREC conferences and MARC format do not help in
45 explaining this fact.]

46
47
48
49 Data Integration for Research and Innovation Policy An Ontologybased Data Management
50 Approach
51
52
53
54
55
56
57
58
59
60

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 intertwined. 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

1
2
3 phenomenon “information deluge,” “data deluge,” or “tsunami of data” (Hey and Trefethen
4 2003). According to these sources, it is impacting business, government, culture, science,
5
6 and society.
7
8

9
10 Big Data reminds us of the so-called “information explosion,” a fundamental phenomenon
11 connected to the rise of Information Science and KO in the area. In response, KO created
12 knowledge organization systems (KOS) that work in conjunction as auxiliary systems with
13 information retrieval systems (IRS), which are traditionally computerized databases
14 containing representations of scientific documents. Such KOS, the “information retrieval
15 thesaurus” (Dextre Clarke 2016, 138), that control and standardize the natural language used
16 both for indexing the documents entered in the IRS and standardizing natural language the
17 keywords used in the user's queries formulated by users, in an “information retrieval
18 thesaurus” (Dextre Clarke, 2016, 138).
19
20
21
22
23
24
25
26
27
28
29
30

31
32 Most The conceptualizations of Big Data tend to repeat those originating in computer
33 science, emphasizing technological aspects such as volume, variety, velocity,
34 heterogeneity, and the need for massive computer power to process it (Gandomi and Haider
35 2015). Although Big Data has also been sparking interest in KO (Ibekwe-SanJuan and
36 Bowker 2017, 192), raising questions like its impact in KO epistemology and methodologies
37 (Hajibayova and Salaba, 2018), (Frické 2015) (Reviewer 2). However contributions from
38 the area to contextualize it and propose practical solutions are still few; Hjørland (2013, 179)
39 stressed: “But such progress is brought to us from the outside; it is not something the field of
40 KO has provided”. The availability today of huge datasets recording user interactions with
41 different systems, their interests, and preferences, gave rise to the development of data-
42 driven methodologies to guide the interactions between users and such systems, including
43 IRS, an area of application of KO. Nonetheless, methodologies and tools created on their
44 bases have been developed by private enterprises such as Google, Amazon, Netflix.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

1
2
3 way that had a precise and formal meaning or semantics and would be intelligible and
4
5 understandable by both people and machines.
6
7

8
9 In previous works, we have already discussed how to link digital representations of
10
11 objects of memory and culture through the Web (_____, 2020), and how one of the
12
13 products of science, scientific publications, could be intelligible and understandable by both
14
15 people and machines (_____ and Costa 2016), when represented with the technologies
16
17 of Linked Open Data (LOD) and the Semantic Web. Here we are interested in doing the
18
19 same for science's other great product, digital research data.
20
21
22

23 1.2. Traditional use of vocabularies to assign subjects to documents The document centered
24
25 vision of vocabularies
26
27

28
29 The technical traditions and standards developed by KO to manage the information
30
31 explosion rest on resulted in the establishment of IRS/KOS assumptions that persist to this
32
33 day. In most discourses in the area, these assumptions are so implicit that it becomes
34
35 difficult to make them explicit, consider them, and analyse their consequences. All the
36
37 theories and methodologies of KO mentioned bringing these assumptions implicitly: the IRS
38
39 represent documents in their computerized databases; MARC and the bibliographic formats
40
41 that emerged from the UNISIST Reference Manual for machine-readable bibliographic
42
43 descriptions (Dierickx and Hopkinson, 1986) (Reviewer 2) are metadata sets that represent
44
45 different descriptive properties of the documents.
46
47
48

49
50 KOS associated with IRS confirm such assumptions; they “have been designed to support
51
52 the organization of knowledge and information to make their management and retrieval
53
54 easier” (Mazzocchi 2018). They are terminological control instruments for documents'
55
56 descriptive properties, mainly subjects, among a few others (as authorities and geographical
57
58
59
60

1
2
3 names). used to standardize the records' subject and authorities fields in IRS computerized
4
5 databases, so useful for users' subject-based (Foskett 1996) retrieval.
6
7

8
9 These records represent objects that have, among others, the property of having
10
11 subjects. They are symbolic objects, documents. It is worth adding that the records
12
13 themselves, the metadata set, are also symbolic objects, representing document-type objects.
14

15
16 These implicit assumptions account for the division that occurs in the teaching and practice
17
18 of librarianship and KO between descriptive representation and thematic representation, or
19
20 of subjects, of a document.
21
22

23
24 Representing documents and their subjects has been foundational to the practices developed
25
26 by KO, Representing documents and their subjects is a practice with a long tradition in KO.
27
28 In the past such documents surrogates were especially when, unlike today, there was no
29
30 access to full-text documents in digital format and the descriptive and thematic
31
32 representation of the documents was a fundamental mechanism to provide access to
33
34 information and enable processes of relevance assessment in the intermediation, and
35
36 relevance assessment processes carried out by libraries and IRS in the retrieval of
37
38 information (Saracevic 2007). KO methodologies have always represented domains of
39
40 knowledge when building KOS like controlled/standardized vocabularies, subject headings,
41
42 and taxonomies classification schemas. The early KOS, such as thesauri, were intended to
43
44 enable subject-based retrieval in the context of IRS because their records were
45
46 representations of objects that have as one of their properties subjects. But not all objects in
47
48 a domain have subjects as one of their properties like documents. We see now that this is
49
50 just one among many cases of representing different objects in digital space.
51
52
53
54
55
56

57
58 To what extent do these assumptions hold up today, and are they sufficient to address the
59
60 challenges of the Semantic Web era, Big Data, research data, and the Internet of Things?

1
2
3 Today, it is not only the case of about retrieving documents (or their representations) but
4 also to create digital representations of anything, as demanded by such as in the “Internet of
5 Things” (IoT) (Gershenfeld, Krikorian, and Cohen 2004). If the documentation movement
6 (Otlet 2018) and then Information Science intended the empowered rment of information by
7 separating it from books, the Semantic Web proposal and Big Data did the same with the
8 knowledge to also, (Soergel 2015). which It is no longer just inserted into texts to be
9 interpreted by humans, but rather serialized recorded directly in Resource Description
10 Framework (RDF) triples (Reviewer 2) (RDF 1.1 PRIMER 2014), forming
11 representations/descriptions of “things” The Web thus becomes a large knowledge base that
12 can be consulted about the “things” thus represented (SPARQL 1.1 QUERY LANGUAGE,
13 2013).

14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29 The objective of this work is to discuss how KO vocabularies, in the sense used within LOD
30 Technologies i.e., value and metadata vocabularies (Zeng 2019), can contribute to assigning
31 computational semantics digital research data within the context of Big Data to Big Data,
32 especially to research data, so that computers can process them, allowing their reuse on large
33 scale. (Reviewer 2)

34
35
36
37
38
39
40
41
42 As a methodology, the work develops a conceptualization of data and (the few of) Big Data
43 originating in KO in an attempt to make it clearer what would be data, as an essential
44 element of the Big Data phenomenon, and in particular, digital research data. It then
45 proceeds to analyse digital research data uses and cases and their relation to semantics and
46 vocabularies its different levels of aggregation, based on the Case Report Form (CRF),
47 [WHO-COVID-CRF/WHO-2019-nCoV-Clinical_CRF-2020.3-eng.pdf](https://www.who.int/csr/don/20200323-who-clinical-crf-2020-3-eng) at master ·
48 [FAIRDataTeam/WHO-COVID-CRF · GitHub](https://github.com/FAIRDataTeam/WHO-COVID-CRF), proposed by the World Health Organization
49 (WHO) to standardize and unify the registration of cases of patients with COVID-19
50 worldwide.

1
2
3 The work is organized as follows. After this introduction, section 2 analyses data from a
4 semiotic and ontological point of view. Section 3 presents a comprehensive view of
5 vocabularies within the context of Semantic Web and LOD. Within such a context Section 4
6 definitions, their traditional use in KO, and develops a conceptualization of data that is
7 illustrated by examples of research data, research datasets, and related initiatives, and shows
8 how research data at different levels of aggregation yields semantics, relating them to the
9 representation of things in a domain and organized into vocabularies. Section 3 presents a
10 comprehensive view of vocabularies based on Semantic Web and LOD technologies and
11 discusses which functionalities vocabularies must incorporate to integrate with these
12 technologies. Section 5 draws conclusions, raises research questions to be developed and
13 presents final considerations.

2. Semiotic and ontological view of data

27
28
29
30
31
32 None of the most common Big Data definitions exclude the data component. It seems
33 reasonable, then, that to understand what Big Data is and how to operationalize solutions to
34 the problem begins by elucidating what is data. After presenting the traditional use of
35 vocabularies to represent and assign subjects to documents this section proposes a semiotic
36 and ontological analysis of data, understood as the essential component of Big Data and
37 research data. This analysis begins with the question of conceptual models and domains and
38 goes on to analyse how conceptual models of domains are expressed linguistically as
39 vocabularies. elucidating how semantics arises from data. Then data is discussed from a
40 semiotic and ontological point of view. From the elucidation of these questions, concepts of
41 research data, data concerning domains of human action, and vocabularies as representations
42 of domains, are developed.

1
2
3 2.3. Traditional use of vocabularies to assign subjects to documents, - generalized use of
4
5 vocabularies as representations of a domain
6
7

8
9 Since the onset of the information explosion, thesauri have emerged, complementary
10
11 systems to the IRS, of which the KOS were one of its components. The development of the
12
13 IRS drew on sources from other traditions of librarianship, documentation, and cataloging.
14
15 Catalog sheets and bibliographic entries served as models for computational bibliographic
16
17 formats in projects such as Machine Readable Cataloging (MARC), based on the AACR2
18
19 cataloging standard, and the UNISIST reference manual for machine-readable bibliographic
20
21 descriptions (Reviewer 2) Reference Manual (Dierickx, Hopkinson 1986), based on the ISBD
22
23 standard. These formats served as a model for library catalog databases and for indexing and
24
25 summary services. Significantly, bibliographic formats evolved separately from the also
26
27 nascent technology of computer databases that, from the 1970s onwards, had the relational
28
29 model as a paradigm (Codd, 1970). The Text Retrieval Conferences (TREC) conference
30
31 series illustrates this separate evolution. (Reviewer 2)
32
33
34
35
36

37
38 Concerning thematic representation, a whole theoretical and methodological
39
40 foundation were all developed to support the development of KOS, from classification
41
42 theories, the Faceted Classification Theory (Ranganathan and Gopinath 1967), the proposals
43
44 of the Classification Research Group (CRG) (Wilson 1972), Concept Theory (Dahlberg
45
46 1978), to Terminology (Cabr  2005). This theoretical and methodological tradition, an area
47
48 of excellence of KO, meets now, with the emergence of the Semantic Web, in subdisciplines
49
50 such as systems modeling, artificial intelligence, and computational ontologies, areas
51
52 originating from computer science. These are understood as one of the foundations of the
53
54 proposed Semantic Web proposal. Many of these new KOS vocabularies are developed by
55
56 computer professionals and scientists from different areas or specialities: biomedicine,
57
58 statistics, or from curators of digital collections in memory and culture, etc. The words of
59
60

1
2
3 Hjørland (2008, 86) highlight and warn about this approach to other areas: “(LIS) is the
4 central discipline of KO in this narrow sense (although seriously challenged by, among other
5 fields, computer science).” Will KO limit itself to developing traditional KOS and leave this
6 space to computer science specialists as Hjørland warns?
7
8
9

10
11
12
13 The technical traditions and standards developed by KO to manage the information
14 explosion resulted in the establishment of IRS/KOS assumptions that persist to this day. In
15 most discourse in the area, these assumptions are so implicit that it becomes difficult to
16 make them explicit, consider them, and analyze their consequences. All the theories and
17 methodologies of KO mentioned bring these assumptions implicitly: the IRS represent
18 documents in their computerized databases, MARC and the bibliographic formats that
19 emerged from the UNISIST Reference Manual for machine-readable bibliographic
20 descriptions (Dierickx and Hopkinson, 1986) are sets of metadata that represent different
21 (descriptive) properties of the documents, while the KOS associated with them are
22 terminological standardization instruments specifically for the subject property, the subject
23 field of the records of the IRS computerized databases. These records represent objects that
24 have, among others, the property of having subjects. They are symbolic objects, documents.
25 It is worth adding that the records themselves, the metadata set, are also symbolic objects,
26 representing document-type objects.
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44

45
46 These implicit assumptions account for the division that occurs in the teaching and
47 practice of librarianship and KO between descriptive representation and thematic
48 representation, or of subjects, of a document. To what extent do these assumptions hold up
49 today, and are they sufficient to address the challenges of the Semantic Web era, Big Data,
50 research data and the Internet of Things?
51
52
53
54
55
56
57
58
59
60

2.1. Vocabularies as representations of domains

1
2
3 In the 1980s-1990s, as a consequence of the emergence of online bibliographic catalog
4 management systems and databases, the domain of information retrieval in library
5 catalogues, so familiar to us but also so exclusive, with its diversity of objects, was **first**
6 modelled using a methodology used in computer science to plan database management
7 systems. The Functional Requirements for Bibliographic Records conceptual model (FRBR)
8 **based on Chen (1976) Entity-Relationship (E-R) model**, appeared in 1998, whose
9 development was promoted by IFLA (1998).

10
11
12 According to Mylopoulos (1992, 3) “Conceptual modeling is the activity of formally
13 describing some aspects of the physical and social world around us for purposes of
14 understanding and communication.” For Mylopoulos,

15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

1
2
3 given domain? How are they distinguished from each other? How do they relate? What are
4
5 their properties?
6
7

8
9 As a representation, a conceptual model is expressed, communicated, and externalized
10
11 through a language, or more specifically a meta-language or meta-model (Guizzardi 2007,
12
13 23), which is a language to express the vocabulary (concepts, terms) that express things in
14
15 specific domains. Examples of these meta-languages are either natural language (through a
16
17 system requirements document), which functions as the most general of all meta-languages,
18
19 or a diagrammatic meta-language, such as entity-relationship (meta) Model or the Unified
20
21 Modelling Language (UML), <https://www.uml.org/>, class diagram, in which domain-
22
23 specific ER models or class diagrams are expressed. Both an ER model and a class diagram
24
25 can define a language that designates things in a domain or a specific vocabulary to that
26
27 domain.
28
29
30
31

32
33 Within descriptive representation, once established and consolidated practical standards
34
35 such as MARC, UNISIST, AACR2 and ISDB (Reviewer 2), the question of what are the
36
37 "things" represented, were implicit in them, their conceptual models, is raised, a view with a
38
39 higher level of abstraction of a domain.
40
41

42
43 Conceptual models in the area of documentation and information have made things like
44
45 documents, authors, and subjects explicit. They evolved from the previously mentioned
46
47 standards for creating automated bibliographic records, starting with the pioneering FRBR
48
49 (Ifla 1998). FRBR, as a conceptual model of the bibliographic domain, is not intended for
50
51 describing or indexing documents, but for formalizing, identifying, agreeing, and
52
53 standardizing objects, actors, and processes and their relationships within such domain.
54
55

56
57 Modeling in documentation and information has roots in Universal bibliographic
58
59 classification systems such as the Dewey Decimal Classification (DCC) – and the Universal
60

1
2
3 Decimal Classification (UDC). The DCC and UDC can be viewed are used for thematic
4 representation, for assigning subjects – as discipline names - to books. They model the
5 universe of knowledge as a set of taxonomies, each having as a root a discipline into which
6 the universe of knowledge was classified. The use of taxonomies to organize a domain is
7 typically used today for information management within corporations and to organize the
8 content of websites (Lambe 2014). Taxonomies only organize the things in a domain in
9 class-subclass relationships. The things being organized in a universal bibliographic
10 classification are discipline names to be used as subjects to books.

11
12 However, there are more than just things or taxonomies of things in a domain. A more
13 accurate model of a domain should include not only the things within it but also their
14 properties, relationships and attributes. Things have properties, attributes, and relationships,
15 according to the ER model. The first movement within documentation and information to
16 recognize this fact was Faceted Classification (Ranganathan, Gopinath 1967). Facets are the
17 properties of a class of things of interest for information recovery (Giunchiglia et al 2014;
18 _____ and Dias, 2020). Besides things, conceptual models also embody properties of
19 things, their attributes, and relationships; Including properties of things Recognizing this fact
20 results in a more accurate representation of a domain, a conceptual model, with richer
21 semantic expressiveness (Almeida, Souza and Fonseca 2011) than a taxonomy (Reviewer 2).

22
23 After the pioneering FRBR model (Ifla, 1998), the International Council of Museums
24 (ICOM) adopted the CIDOC Conceptual Reference Model (CIDOC 2014), IFLA released
25 the Library Reference Model (LRM) integrating the FRBR, FRAD, FRSAD models (Riva,
26 Le Boeuf, and Žumer 2017) and more recently the International Council of Archives (ICA)
27 adopted the Records in Context Conceptual Model (Ric-CM) (International Council on
28 Archives 2019). Since the publication of the FRBR model in 1998, KO has been changing
29 its representation activities and methodologies, from records describing documents and their

1
2
3 subjects to conceptual modeling, that is, representing entities, their attributes and
4 relationships (Prasad, Giunchiglia, Devika 2007). Knowledge organization and
5 representation is part of the digital research data curation effort. Such domains of application
6 also uses conceptual models to integrate heterogeneous research data sources as
7 publications, research data, patents, projects, events, funding agencies, etc. (CERIF in Brief
8 2014)
9

10
11 Conceptual models, when designating things in a domain, define a metadata vocabulary.
12 They are aligned together with different types of KOS by Almeida, Souza and Fonseca
13 (2011, 196), ordered according to their semantic expressiveness. Semantic expressiveness
14 can be understood, in the context of the previous quote, as the ability of each type of KOS to
15 distinguish and describe, that is, identify the properties and represent the different things that
16 exist in a domain of that reality.
17

18
19 Conceptual model elements - entities, attributes and relationships - are expressed
20 linguistically by a vocabulary. Most types of Vocabularies are semantic control devices,
21 formed by systematised sets of semiotic, triadic entities (PEIRCE 1994), concepts (Dahlberg
22 1978) (Dextre Clarke and Zeng 2012), units of meaning that relate something (a first: object
23 or referents), in some way (through a second: term or code), which generates or induces a
24 third: its meaning. Vocabularies are constructed to answer the basic ontological question:
25 what exists in a domain? They are representations or models of a domain of reality, pointing
26 out what things there are, what their attributes are, their relationships, and how to express
27 them linguistically, the concepts (Dahlberg 1978), and their units of meaning. Online Public
28 Access Catalogs (OPACs) or databases are used as terminological control tools with the IRS
29 used by institutions, with the function of standardising the terms used for the description and
30 indexing of scientific publications, memory and cultural objects, and other items included in
31 these systems.
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

1
2
3 about them, represented in such a way that they can be processed/intelligible by both
4
5 machines and humans. Such representations allow machines to make inferences about the
6
7 knowledge thus represented.
8
9

10
11 KO today is being called upon to model different domains of knowledge to build new
12
13 “semantic” vocabularies, *i.e., vocabularies compliant with the Semantic Web and LOD*
14
15 *technologies*. For this, it is necessary to expand the traditional notion of a domain as a
16
17 discipline or subject. In the area of software development the notion of a domain has a
18
19 broader scope: it is ‘a sphere of activity or interest: field’ [Webster]. In the context of
20
21 software engineering, it is most often understood as an application area, a field for which
22
23 software systems are developed (Prieto Díaz 1990, 50).
24
25
26
27

28 Since a vocabulary **KOS** is a terminological system that represents the “things” of interest in
29
30 a domain of action to the community of agents/users in that domain, then to create a
31
32 vocabulary **KOS** (an artifact, similar to software) several aspects and questions must be
33
34 considered: *What things are in a domain? How should they be represented? These are the*
35
36 *questions of ontology and semiotics. They must be answered to create a representation, or a*
37
38 *conceptual model, of a domain.*
39
40
41
42

43 **We must** A first step is to determine what things exist in a domain and which are relevant to
44
45 this community, what rules exist about these things or are created/approved/agreed on about
46
47 these things, and how this community uses them to act in this domain. **and, f**Finally, how the
48
49 conceptualizations *and their agreed terms* (Dahlberg 1978), **generating as** one of the by-
50
51 products of this process **a set of terms**, are to be systematised, **for example**, in a *domain*
52
53 *model to serve as bases for the construction of vocabularies such as* thesaurus or
54
55 computational ontologies.
56
57
58
59
60

1
2
3 What things are in a domain? How should they be represented? These are the
4 questions of ontology and semiotics. They must be answered to create a representation, or a
5
6 conceptual model, of a domain.
7
8
9

10 2.5. Vocabularies as representations of a domain.

11
12
13
14 As shown, a vocabulary is a representation of a domain. A domain
15 vocabulary can be used regardless of its use, either to assign subjects to documents: a)
16
17 vocabulary for indexing, which identifies what things exist in a domain (e.g. MeSH
18
19 categories describing the entities within the Healthcare domain,
20
21 <https://meshb.nlm.nih.gov/treeView>, or b) to describe objects in this domain, descriptive
22
23 metadata standards that, in addition to identify what things exist in a domain, also describe
24
25 their properties – attributes and relationships. Among the things within a domain some
26
27 vocabularies focus on specific facets for special purposes: archival science and records
28
29 management uses functional classification plans in an organization to assign the
30
31 organizational provenance or the function or organizational process that generated or used a
32
33 record.
34
35
36
37
38
39
40

41 2.1, 2.3. Data as Representations

42
43
44 What is Big Data? What is its relationship with data? What is data and how is it related to
45
46 metadata? How should semantics be assigned to data? As noted in the ISO/IEC 20546/2019
47
48 Standard, “The big data paradigm is a rapidly changing field with rapidly changing
49
50 technologies,” later suggesting a definition: “extensive datasets (3.1.11) — primarily in the
51
52 data (3.1.5) characteristics of volume, variety, velocity, and/or variability — that require a
53
54 scalable technology for efficient storage, manipulation, management, and analysis.”
55
56
57
58
59
60

1
2
3 The conceptualizations of Big Data originating from KO are few (_____, et al 2021)
4
5 and replicate those originating in computer science, define it as a phenomenon that involves
6
7 large amounts of data, the heterogeneity of that data, a continuous flow of generation and
8
9 updating, and a need for large processing capacity so that the data reveal patterns or trends
10
11 (De Mauro et al 2015). However, the same is not true for the conceptualizations of data
12
13 originating from KO. Data is mentioned frequently in the literature, along with its
14
15 relationships with information and knowledge (Buckland 1991), often called the data,
16
17 information, knowledge, wisdom (DIKW) hierarchy (Rowley 2007). In Floridi (2019),
18
19 information is related to data and semantics.
20
21
22
23
24

25 An important exception is from Hjørland (2018), who proposes a conceptualization of Big
26
27 Data arising from definitions of data, a phenomenon much better known and conceptualized
28
29 within the area KO. Data is in the essence of the Big Data phenomenon, it could not exist
30
31 without data. In this work, Hjørland lists several similar conceptualizations of data and
32
33 highlights that of Fox and Levitin:
34
35

36
37 ()
38 Within this framework, we define a datum or data item, as a triple $\langle e, a, v \rangle$,
39
40 where e is an entity in a conceptual model, a is an attribute of entity e , and v
41
42 is a value from the domain of attribute a . A datum asserts that entity and has
43
44 value v for attribute a . Data are the members of any collection of data items.
45
46

47 Such conceptualization is clarified by the following example: “2018.” What does 2018
48
49 mean? Others would say it’s a given. Let us note, however, this statement: “Giovana was
50
51 born in 2018.” In it we can identify the entity we are talking about: a child called “Giovana,”
52
53 an attribute or property of this entity, she is “born,” and the value of this attribute or
54
55 property, her birth year, “2018.” To achieve a formal representation it is very important to
56
57 clearly identify the entity being described. Although a data set usually has a title or
58
59
60

1
2
3 description identifying the entity it represents that is not always the case. A metadata set
4 may mix metadata elements of different entities as for example the MARC21 format field
5 245 – Title Statement; while MARC21 format describes a bibliographic entity, e.g., a book,
6 field 245 subfield code \$c describes another entity, the responsible for the book, and field
7 245 subfield \$f its attributes birth and death dates.
8
9

10
11
12
13
14
15
16 In the ontological scheme that goes back to Aristotle (2000), the reality is constituted of the
17 first substances, the things that have real existence in space and time, and second substances,
18 the conceptualizations we make of the first substances to think, reason, make sense of, and
19 communicate about the things in reality. Second substances are in turn subdivided into
20 essences, concepts designating things that have properties whose loss implies the non-
21 existence of that individual and have existential independence (Fonseca, Porello, Guizzardi,
22 Almeida, and Guarino 2019, 29), and accidents, concepts that designate things that are
23 existentially dependent on other substances. Things having existential independence are
24 commonly recognized in one of the most well-known ontological schemes, the entity-
25 relationships (ER) model (Chen 1976) as entities, while those that are existentially
26 dependent, as properties. Properties, in turn, are subdivided into attributes of an entity,
27 relationships between an existentially independent entity and the value of one of its
28 properties, and relationships, involving two or more individuals of the same, or of different
29 existentially independent entities, or of more than one existentially independent entity
30 (Orilia and Paoletti 2020).
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50

51
52 Classifying concepts in vocabularies as entities and their properties, attributes or
53 relationships is a practice that has become common in the specification of vocabulary
54 compliant with LOD technologies; see, for example, the DC Terms vocabulary,
55 <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>, the PROV-O ontology,
56
57
58
59
60

1
2
3 <https://www.w3.org/TR/prov-o/>, and DCAT metadata vocabualry,
4
5 <https://www.w3.org/TR/vocab-dcat-3/>.
6
7
8

9 Data is about representations of something else. A piece of data unit, a datum (Hjørland
10 2018), even in the context of Big Data, then, makes no sense without referencing the entity
11 and one of its properties, the metadata (Reviewer 2). The three concepts are inseparable and
12 cannot be understood separately. They correspond to a descriptive, representational element
13 of an entity, describing one of its properties. They correspond linguistically to a claim, a
14 basic unit of knowledge to which, according to Aristotle (2000, 39), values of truth or falsity
15 can be attributed.
16
17
18
19
20
21
22
23
24
25

26 The statements represented by triples constituted by an entity, one of its properties, and the
27 value of this property correspond to the representation of informational resources in the
28 context of LOD, using the RDF data model (RDF Primer 2014). RDF is a Semantic Web
29 standard for describing resources. Everything that is available on the Web can be accessed
30 through a link, or a Uniform Resource Identifier (URI). Today URI evolved towards IRI, the
31 Internationalised Resource Identifier, which strings incorporate characters from alphabets
32 others than the Latin alphabet. This representational model describes such a resource
33 through triples formed by a subject, the resource being described; a predicate, a property that
34 describes the resource; and an object, the value of this property for this resource. The RDF
35 model assumes a minimum semantics, that is, three elements with specific roles, the subject,
36 the predicate, and the object that form the triple are identified and appear in this order.
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51

52 Semiotic and ontological analysis identifies a piece of data as an artificial and intentional
53 artefact that represents something. The foundational types of the things that exist are entities
54 - existentially independent things - and their properties: relationships between two
55 existentially independent individuals, and attributes of an individual, its qualities and
56
57
58
59
60

1
2
3 quantities. Ontological Analysis of things in a domain, classifying and assigning types to
4 these things makes the terms in a domain vocabulary consistent, as they inherit the
5 ontological nature of their types and enable their representations to be machine-processable.
6
7
8
9

10 11 **3. A Comprehensive view of vocabularies**

12
13
14 In this section, a comprehensive view of vocabularies based on the previous discussion in
15 section 2 and on contributions by Hjørland (2018) and Zeng (2019) was compiled and
16 developed.
17
18
19

20 21 22 3.1. Vocabularies, Web of Data, Linked Open Data, and Big Data

23
24
25 LOD technologies are an integral part of the Web of Data project. Although this is its best-
26 known name, the project is also known as Web of Data, a name that describes it better, since
27 semantics concerns meanings (Chierchia, 2003), and the ability of the Web of Data to
28 convey meanings is quite limited and different from the sense in our understanding of
29 expressions in natural language.
30
31
32
33
34
35

36
37 The project was initially formulated by computer scientist Tim Berners-Lee, the creator of
38 the Web, among others. According to its formulators, the Semantic Web aims to propose “A
39 new form of Web content that is meaningful to computers will unleash a revolution of new
40 possibilities” (Berners-Lee et al 2001). To its authors, “Most of the Web’s content today is
41 designed for humans to read, not for computer programs to manipulate meaningfully.” The
42 Semantic Web then “will bring structure to the meaningful content of Web pages, creating
43 an environment where software agents roaming from page to page can readily carry out
44 sophisticated tasks for users.”
45
46
47
48
49
50
51
52
53
54

55
56 The Web of Data then refers to content represented in such a way that it can be understood
57 by both machines and people. The current Web is made up of pages, such as
58
59
60

1
2
3 <http://www.uff.br>, formatted in Hypertext Markup Language (HTML), accessible and
4
5 interconnected with each other through links. Navigating these pages through these links is
6
7 done by browsers, such as Internet Explorer, Google Chrome, or Mozilla Firefox. HTML is
8
9 a content markup language; it formats the content of a text of a page through a predefined
10
11 set of markups, which instruct browsers to display them on computer screens for human
12
13 users. The content of HTML pages is interpreted by browsers to make it readable and
14
15 visually pleasing to people.
16
17

18
19
20 The proposed Web of Data is quite different. The Web will no longer be constituted of pages
21
22 to be read by people, but of content, called informational resources, digital representations
23
24 of things: concrete, like me, you, an industrial product, a monument, a geographical
25
26 accident; abstract, like a musical genre, a scientific discipline; or just has a digital existence,
27
28 such as a photo in a JPG file or a scientific article in a PDF file. These are the entities in the
29
30 proposal by Hjørland (2018). Each of these resources is uniquely identified by a link, or a
31
32 URI. A resource, identified/accessed by its URI, is described in a structured way through
33
34 triples, each one formed by the URI of the resource, by each of its properties, and by the
35
36 corresponding values of each of these properties. An example of how this representational
37
38 model works is the Leonardo Da Vinci resource on Wikidata,
39
40 <https://www.wikidata.org/wiki/Q762>.
41
42
43
44
45

46 This model of structuring data through the description of resources formed by one or more
47
48 linguistic claims made up of triples <Subject> <Predicate> <Object> is RDF (RDF Primer,
49
50 2004). From an ontological point of view, subject, predicate, and object can be understood
51
52 as an entity, a property, and the value of this property.
53
54

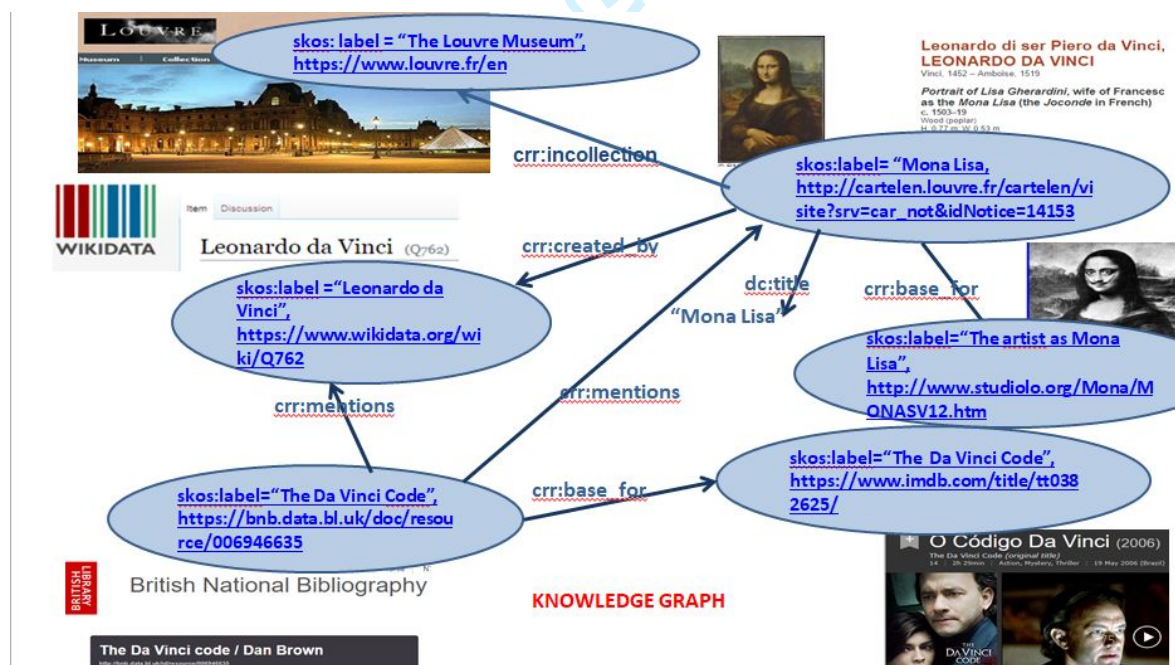
55
56 Looking in more detail at structuring a triple; for example,
57

58
59 “The page <http://www.uff.br> is authored by _____.”
60

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

1
2
3 they can be processed by machines (including their user-friendly visualisation), thus helping to
4
5 organise, retrieve, and make these resources accessible.
6
7

8
9 Naturally, given a triple like `< http://www.uff.br > < is authored > <`
10
11 `>`, a machine cannot do much more than identify the subject, the predicate or the object of
12
13 the triple. In this example, predicate and object are names, strings of characters
14
15 understandable only by people, and holders of a set of contextual and cultural information,
16
17 accumulated throughout their life histories. RDF Semantics is limited to its model of
18
19 structuring triples as subject, predicate, and object.
20
21
22

23
24 The way to extend these semantics beyond the limits of the RDF model is also to make
25
26 predicates and/or objects into URI and that these URI refer to concepts of vocabularies with
27
28 specific semantics. According to RDF Semantics (2004) “There are several aspects of
29
30 meaning in RDF which are ignored by these semantics; in particular, it treats URI references
31
32 as simple names, ignoring aspects of meaning encoded in particular URI forms.” A URI in
33
34 the RDF model is just a name, an identifier. The advantage of a URI over a natural language
35
36 identifier such as the linguistic term “author”, is its uniqueness, its validity, since a URI is
37
38 valid and unique throughout the web space, and its persistence, that is, the commitment of
39
40 whoever assigns it. a URI to never change it (Berners-Lee 1998).
41
42
43
44

45
46 The previous example can be extended by using URI for the subject, the predicate, and the
47
48 object of the triple.
49

50
51 `<http://www.uff.br> <http://purl.org/dc/elements/1.1/creator> <https://orcid.org/0000-0003-`
52
53 `0929-8475>`
54

55
56
57 In this example, the original predicate “author” is replaced by the URI referenced by the
58
59 “creator” element of the well-known Dublin Core (DC) metadata standard. In its context,
60

1
2
3 dc:creator has specific semantics. It is defined as “An entity responsible for making the
4 resource.” The triple’s object, the value or content of dc:creator, has been replaced by the
5 Open Researcher and Contributor ID (ORCID), <https://orcid.org>, of the page’s author.
6
7

8
9
10
11 It is with the semantics in specific vocabularies that the limited semantic expressiveness of
12 the RDF model can be expanded, as seen in the example of the CRF. Once specified in
13 elements of a vocabulary, the semantics can be processed by programs. While the features
14 provided in the Web of Data, represented in markup languages such as XML, RDF, HTML,
15 etc. are contents, programs are procedures or algorithms according to the data processing
16 model. Programs only know how to process content, For this, they need to be clearly
17 instructed (programmed) on what to do with certain content in a certain situation. Specially
18 formatted vocabularies, the LOV (Mendez and Greenberg 2012) used to assign semantics to
19 LOD (Zeng 2019) must clearly define, restrict, and specify the semantics of their concepts.
20 For example, the DC metadata vocabulary clearly defines the semantics of each of its
21 concepts (called elements in the DC initiative); for example, dc:creator, is the creator/author
22 or responsible for a resource, e.g., a digital scientific paper. Furthermore, the dc:creator
23 element has itself, a unique persistent identifier, a link, a URI:
24 <http://purl.org/dc/elements/1.1/creator>. This persistent identifier, unique throughout the Web
25 space, works as a guarantee of the metadata element semantics, allowing a developer to
26 create a specific program to process this element of the DC vocabulary unambiguously,
27 using the semantics specified and standardized in the DC vocabulary to the dc:creator
28 element.
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51

52
53 Here is another example of what was just explained. Let the following RDF triples be:
54
55

56
57 TABLE 1. Two triples with the same predicates
58
59
60

<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

1
2
3
4
5 Through unique and persistent identifiers, metadata and data vocabularies can be used to
6
7 assign machine-understandable semantics to predicates and objects in triples RDF. Many old
8
9 vocabularies are being restructured to be compatible with LOD technologies (Soergel, 2004;
10
11 Dos Santos Maculan, 2015), such as the UNESCO Thesaurus,
12
13 <http://vocabularies.unesco.org/browser/thesaurus/en/>, the FAO Thesaurus,
14
15 http://aims.fao.org/aos/agrovoc/c_8003.html, the AGROVOC Thesaurus,
16
17 <https://agrovoc.fao.org/browse/agrovoc/en/>, the Paul Getty Foundation Vocabularies,
18
19 <https://www.getty.edu/research/tools/vocabularies/lod/>, the Art and Architecture Thesaurus,
20
21 the Union List of Artists Names, the Cultural Objects Name Authority, the Getty Thesaurus
22
23 of Geographic Names, the DeCS/MeSH, Health Science Descriptors,
24
25 <https://decs.bvsalud.org/th/>, the Library of Congress Subject Headings (LCSH),
26
27 <https://id.loc.gov/authorities/subjects.html>, in addition to many others.
28
29
30
31
32
33
34

35 Vocabularies used with LOD need to meet requirements such as having their concepts
36
37 persistently and uniquely identified through valid URIs on the internet, being represented in
38
39 machine-readable formats such as RDF, containing precise definitions of the semantics of
40
41 their concepts, and generally, being multilingual. Many of these vocabularies that meet the
42
43 principles of LOD can be found in the aforementioned LOV vocabulary registry service. By
44
45 meeting the requirements for use with LOD as described above, vocabularies, an area of
46
47 study, research, and practical use of KO, can contribute to addressing the issues brought
48
49 about by Big Data.
50
51
52
53
54
55

56 Elements of data or metadata vocabularies referenced by URI account for the semantics of
57
58 an individual “datum” according to (Hjørland 2018), an element of a triple. An example is
59
60

1
2
3 **the fields of the CRF form**. These vocabularies use different approaches to semantics, as
4
5 pointed out in Almeida et al (2011, 195), ranging from semantics for humans, which is
6
7 implicit, informal or formal, to semantics for machines, which is informal, formal, or even
8
9 “powerful semantics” (Shet, 2020). In any case, used in the context of the RDF model these
10
11 vocabularies allow the processing of RDF triples by machines.
12
13
14
15
16

17 3.34. Ontologies as domain models, **definitions, specifications**

18
19
20 Since 1993 Gruber (1993, 199) coined a definition of ontology which is used until nowadays
21
22 as “An ontology an explicit specification of a conceptualization”. Borst (1997, 12)
23
24 developed Gruber’s definition as “Ontologies are defined as a formal specification of a
25
26 shared conceptualization”. Two concepts in this last definition are of importance to the
27
28 present discussion, - formal, i.e. computers’ readable, and – shared, i.e., agreed by a
29
30 community of agents, being them humans or computers.
31
32
33

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

35
36
37 OWL can be used to explicitly represent the meaning of terms in
38
39 vocabularies and the relationships between those terms. This representation
40
41 of terms and their interrelationships is called an ontology. OWL has more
42
43 facilities for expressing meaning and semantics than XML, RDF, and RDF-
44
45 S, and thus OWL goes beyond these languages in its ability to represent
46
47 machine interpretable content on the Web.
48
49
50

51 OWL is a standard language (meta-language in the aforementioned sense) of the W3C for
52
53 representing ontologies, that is, vocabularies that specify the things existing in a domain and
54
55 their interrelationships. Further on, the same specification compares the semantic
56
57 expressiveness of OWL with that of other languages to represent machine-interpretable
58
59
60

1
2
3 content such as XML, XML Schema, RDF, and RDFS (ONTOLOGY WEB LANGUAGE
4 OVERVIEW, 2004). It can thus be concluded that, with current technologies, a
5
6 computational ontology developed in OWL is the most expressive type of KOS, because the
7
8 “facilities” provided by OWL allow restricting, specifying, and expressing the intended
9
10 meaning (Guarino 1994, 560) of the conceptual model of a domain **obtained by the**
11
12 **modeling process.**
13
14
15
16
17

18 **Each concept of an ontology vocabulary is typed; it is a class, or a property of a class or an**
19 **instance, an individual of a class.** Among these facilities are the possibility of specifying
20 data properties (attributes, in Chen's ER model), object properties (relationships in Chen's
21 ER model), domain and scope of the two types of properties, and cardinality constraints of
22 each class involved in an object property, transitivity and reflexivity of properties, the
23 disjunction between individuals of different classes, axioms for restricting the inclusion of
24 instances in a class (ONTOLOGY WEB LANGUAGE OVERVIEW 2004), etc. These
25 facilities can make conceptual models implicit in a computational OWL ontology more
26 faithful to reality. **Ontologies also do not distinguish thematic versus descriptive**
27 **representation; every concept is described by its properties, whether thematic or descriptive.**
28
29
30
31
32
33
34
35
36
37
38
39
40

41 As seen earlier, the Web of Data project, the large-scale reuse of Big Data **and research data**
42 available in increasing amounts on the Web, depends on the one hand on the most
43 expressive vocabularies that describe them, and on the other hand, on programs capable of
44 making inferences, or at least algorithmic processing, on these representations. In this
45 context, specific domain models, intelligible by machines and represented with the
46 maximum possible semantic expressiveness **such as computational ontologies** gain
47 importance, **which, in the current stage of technology, are computational ontologies.**
48
49
50
51
52
53
54
55
56
57
58
59
60

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-

1
2
3 19, the CRF, which was proposed by the WHO. The GO FAIR initiative, [https://www.go-](https://www.go-fair.org/)
4 fair.org/, addresses the WHO proposal by creating a worldwide network of catalogs
5
6 referencing research data collected through the CRF and deposited in repositories and
7
8 available according to the FAIR principles, <https://www.go-fair.org/fair-principles/>, the
9
10 “FAIR Data Points.” Brazil participates in this initiative through the VODAN-Br Virus
11
12 Outbreak Data Network initiative (Veiga et al 2021).
13
14
15

16
17
18 The VODAN initiative is expected to collect huge datasets worldwide. The CRF
19
20 standardized a set of fields of interest to COVID-19 epidemic research. Such fields must be
21
22 filled with metadata and data associated with vocabularies largely agreed and standardized
23
24 within the health sciences domain. This allows the interoperability of different datasets and,
25
26 without which their processing by computers in order would not be possible, and
27
28 consequently, neither drawing conclusions and insights from the data would have the ability
29
30 to extract conclusions and insights. VODAN and FAIR Data Points are efforts to provide
31
32 smart data (Kobielus 2016) to be used to control COVID-19 outbreak.
33
34
35

36
37 Within the RDF model, instead of the subject, predicate, and object of a triple being
38
39 represented in natural language, which is ambiguous and difficult for programs to process,
40
41 each can be identified by a URI. These URIs identify specific terms, both from metadata
42
43 vocabularies—descriptive properties of things in a domain—and data vocabularies—values
44
45 assumed by these properties for specific descriptive metadata. This unified characterization
46
47 of vocabularies as, that is, sets of systematized terms that identify either the descriptive
48
49 properties (metadata) of objects in a domain, or the values - the data - assumed by these
50
51 properties for instance, is due to Marcia Zeng (2019).
52
53
54

55
56 Another important feature of using vocabularies with LOD technologies is that different
57
58 vocabularies can be used simultaneously in the form fields. In Figure 1 shows we see an
59
60

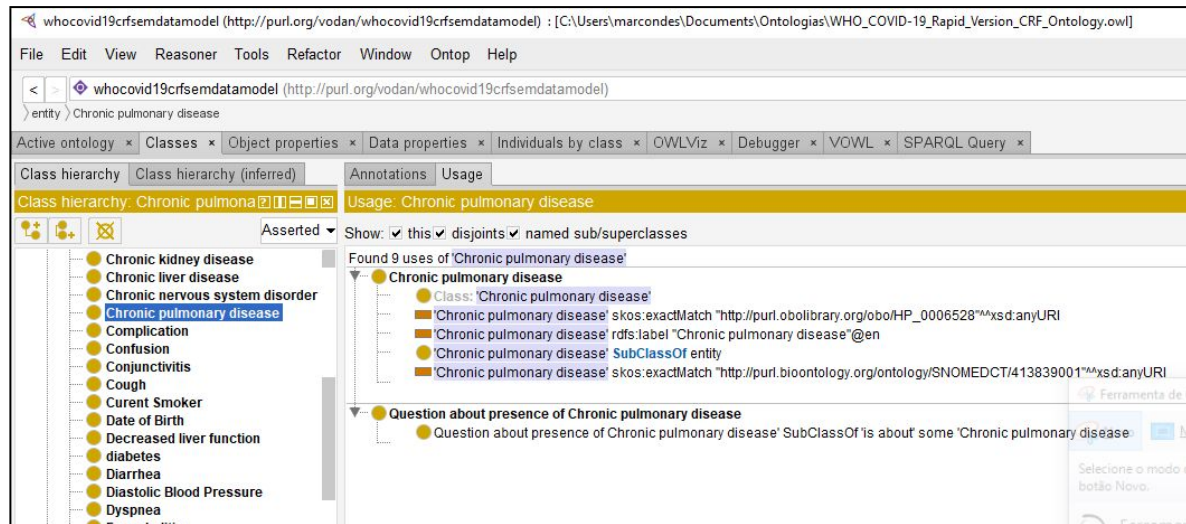
1
2
3 excerpt from the CRF. As co-morbidity data, “CO-MORBIDITIES,” of a patient (the entity)
4 are recorded, concepts such as chronic cardiac disease (the attribute or metadata) are taken
5 from specific biomedical ontologies or vocabularies: Yes, No, Unk (the value or data). the
6 co-morbidity data, “CO-MORBIDITIES,” of a patient (the entity); they are recorded as
7 follows: concepts such as chronic cardiac disease (the attribute or metadata, the co-
8 morbidity presented by the patient) are taken from specific biomedical ontologies or
9 vocabularies that describe specific co-morbidity types; if a specific one applies, it is
10 recorded as data as follows: Yes, No, Unk. These data have to be processed by programs so
11 that the immense amount of records collected through the CRF around the world can serve
12 as inputs for the planning and control of the pandemic. The question about co-morbidities
13 has several answer options, each of which indicates a type of disease. For it to be processed
14 by machines, each type of co-morbidity expressed in natural language must reference a
15 concept in a vocabulary or ontology, such as SNOMED-CT,
16 <https://www.nlm.nih.gov/healthit/snomedct/index.html>, for example. Another question on
17 the CRF, such as the one related to “PRE-ADMISSION AND CHRONIC MEDICATION,”
18 has as one of its answer options “Angiotensin converting enzyme inhibitors (ACE
19 inhibitors)?”, which may be referenced in another vocabulary such as MeSH,
20 <https://meshb.nlm.nih.gov/search>, the term with identifier
21 <http://id.nlm.nih.gov/mesh/D000806>.

22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

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**

1
2
3 research data must be accessible through a URI, represented in RDF, constituting the Linked
4
5 Open Vocabularies.

6
7
8
9 The FAIR principles allow research data to be processed by machines. The M4M
10
11 principle—metadata for machines—states that “There is no FAIR data without machine-
12
13 actionable metadata. The overall goal of Metadata for Machines workshops (M4M) is to
14
15 make routine use of machine-actionable metadata in a broad range of fields.” The CRF
16
17 described above is an example of the importance of research data standardization and the
18
19 adoption of principles that allow its wide dissemination and reuse is the CRF form described
20
21 above. Without standardization, its processing by machines would be impossible.

22
23
24
25
26 Applying the FAIR principles to research data causes data to be represented as RDF triples.
27
28 Such a process is named “FAIRification”, see [https://www.go-fair.org/fair-](https://www.go-fair.org/fair-principles/fairification-process/)
29
30 [principles/fairification-process/](https://www.go-fair.org/fair-principles/fairification-process/). FAIR compliant data is generally derived data from
31
32 datasets. A distributed network of FAIR Data Points provides access to different FAIR data.
33
34 That raises the question of using vocabularies to describe both the original datasets and their
35
36 FAIR compliant datasets versions generated.

37
38
39
40
41 RDA – Research Data Alliance, <https://www.rd-alliance.org/>

42
43
44 Other vocabularies also have emerged, not to describe or provide standardized values for
45
46 each piece of data, but to provide descriptive and value metadata of the datasets as a whole.
47
48 Digital curation of research data is an emerging field of activity for KO professionals; one of
49
50 its activities is to apply metadata to research datasets, see <https://www.dcc.ac.uk/>. For the
51
52 curation of these datasets, metadata standards such as Data Catalog Vocabulary (DCAT)
53
54 <https://www.w3.org/TR/vocab-dcat-2/>, or the Provenance Ontology (PROV-O)
55
56 <https://www.w3.org/TR/prov-o/>, have been adopted to describe the provenance of the
57
58 dataset. As datasets have been made available as informational resources on the Web,
59
60

1
2
3 information on their provenance and the record of the processing carried out on them, the
4
5 extract, transform, load (ETL) and the FAIRrification processes of such data, see
6
7 https://en.wikipedia.org/wiki/Extract,_transform,_load) are essential elements for research
8
9 data reliability to enable sharing and reuse (See <https://www.go-fair.org/fair-principles/>).

10
11
12
13
14 The amount of research data being available every day on Coronavirus epidemic – the
15
16 “V”ariety” of Big Data - makes the integration of such sources essential to control the
17
18 epidemic. The Coronavirus Infectious Disease Ontology (CIDO) (He et al. 2020) stresses
19
20 the essential role computational ontologies in the integration of different and heterogeneous
21
22 research data sources, promoting interoperability between such sources.
23
24
25

26
27
28 These datasets, in addition to the metadata that describe their fields, are themselves of
29
30 interest for the research data exploration which describe the entity represented by the
31
32 dataset. They need additional metadata provided by vocabularies such as DCAT, PROV-O,
33
34 as the type of licence under which data can be reused, the dataset creator, its Publisher, its
35
36 format, its update date, etc, all of which are metadata for the dataset as a whole. Such
37
38 metadata is provided by vocabularies as DCAT, PROV-O. They contain metadata such as
39
40 the format of the dataset, the number of records, the last update date, licences to use this
41
42 dataset, etc. (from DCAT), or metadata such as the entity who generate it (in this case, the
43
44 dataset for which the provenance is to be registered), the agent that created the dataset, and
45
46 the process that generated it (from PROV-O). Standards such as these have been used in
47
48 several research data repositories to index the datasets deposited there. Indeed, digital
49
50 curation is an increasingly common application by KO professionals (Poole 2013).
51
52
53
54
55
56
57
58
59
60

1
2
3 Digital Humanities is another growing area of application of digital research data. It grew
4 from the wide availability of data from social activities (search and social media activity
5 every minute, see [https://www.smartinsights.com/internet-marketing-statistics/happens-](https://www.smartinsights.com/internet-marketing-statistics/happens-online-60-seconds/)
6 [online-60-seconds/](https://www.smartinsights.com/internet-marketing-statistics/happens-online-60-seconds/)) and culture, including science. Scientific articles have long been
7 recognized as a privilege knowledge source (Swanson, 2008), see PubMed Citations per
8 year, https://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html). Significant examples
9 of research projects in Digital Humanities using a variety of such sources can be found in
10 the Digging into Data Challenge program ([https:// diggingintodata.org/](https://diggingintodata.org/)) mentioned by Zeng
11 (2017); in this article, the author describes in details how Digital Humanities is related to
12 Big Data and the challenges to process such data and turn it into Smart Data (Reviewer 1).

13
14
15 A huge amount of such data is textual, resulting from posts on social media, emails,
16 newspaper articles, scientific articles, and text in encyclopaedias such as Wikipedia, among
17 others. This data is unstructured or semi-structured.

18
19
20 The exploitation of such potential information sources may lay on the development of
21 vocabularies for special purposes. Their processing using techniques such as information
22 extraction, named-entity recognition, natural language processing, text mining, machine
23 learning, text annotation, aim at transforming such non-structured or semistructured textual
24 data into structured.

25
26
27 Examples of such techniques in biomedical sciences are the National Library of Medicine
28 Natural Language Processing tools, <https://lhncbc.nlm.nih.gov/LHC-research/nlp.html>,
29 which lay on dictionaries and KOS like MeSH, the Medical Subject Headings, and UMLS,
30 the Unified Medical Language System (Bodenreider 2004), (Aronson and Lang 2010).

31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58 **3.3.** 4.2. Semantics beyond the data.
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

1
2
3 representation of a domain while mapping properties (SKOS 2012), (ISO 25964-2 2013) and
4 also OWL property “sameAs” (Ontology Web Language Overview (2004) enable the
5 mapping of concepts in a data resource to concepts in another.
6
7
8
9

10
11
12 **User-generated content, folksonomies (Hajibayova and Salaba 2018, 145)**
13

14
15 The Web of Data technologies bring a significant advance by incorporating more
16 semantic expressiveness and program independence to data published on the Web. **Big Data**
17 **and research data also poses several issues related to the semantic of data.** This article
18 sought to demonstrate that data, which have a semiotic and ontological character and are
19 artificial and intentional representations, cannot be understood apart from the entity to which
20 they refer and from the metadata—the properties of this entity—that describe it. **Unspecified**
21 **data is also an imprecise concept.** It is necessary also to distinguish one piece of datum as
22 referred to by Hjørland (2018), **which is** a unit that represents the value of one (of the)
23 properties of an entity, from a record, a set of several datum describing different properties
24 of an entity, from datasets, representing the various entities and their properties, and from
25 databases, bringing together different datasets representing different interrelated entities.
26 **Such are different data aggregation levels, Datum as a unit is incomplete, meaningless**
27 **without knowing the entity to which it refers and the specific property of that entity referred.**
28 **has poor semantics, but such data aggregates** having higher levels of semantics in the
29 computational environment. **Vocabularies can play an important role in addressing**
30 **semantics to data at those different levels of aggregation.**
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52

53 The Web of Data technologies bring a significant advance by incorporating more
54 semantic expressiveness and program independence to data published on the Web according
55 to the RDF model. In this model, vocabularies can play a significant role, as has been
56 suggested. **Big data and research data poses several issues related to semantic of data: the**
57
58
59
60

1
2
3 semantic of each piece of datum, the semantic of and RDF triple, of a knowledge graph, the
4
5 “powerful” semantics of the different datasets (the semantic expressiveness of the
6
7 aggregated datasets of other data), the semantic expressiveness embedded in textual Big
8
9 Data which needs to be processed for the identification of entity names, named-entity
10
11 recognition (NER) (Freitas et al 2010), for aggregating annotations and making this data
12
13 structured, the semantic expressiveness given by programs. **Vocabularies can play an**
14
15 **important role in addressing them.** There are however several levels of semantics in the
16
17 variety and heterogeneity of data published on the Web **according to its levels of**
18
19 **aggregation:** the “powerful” semantics of the different datasets (the semantic expressiveness
20
21 of the aggregated datasets of other data), the semantic expressiveness embedded in textual
22
23 Big Data which needs to be processed for the identification of entity names, named-entity
24
25 recognition (NER) (Freitas et al 2010), for aggregating annotations and making this data
26
27 structured, the semantic expressiveness given by programs according to the data processing
28
29 model (for data being processed one way and not another), etc. A systematisation of these
30
31 issues should be included in the KO research agenda.
32
33
34
35
36
37
38

39 Acknowledgments: This work was carried out with the support of the Brazilian agencies
40
41 CAPES - Financing Code 001, and CNPq, grant number 305253/2017-4. **We are also**
42
43 **grateful to the anonymous reviewers of this work for their suggestions to improve this text.**
44
45

46 **References**

47
48
49 Almeida, Mauricio; Souza, Renato and Fonseca, Fred. 2011. “Semantics in the Semantic
50
51 Web: A Critical Evaluation”. *Knowledge Organization* 38 no. 3: 187-203.
52
53 <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1041.7976&rep=rep1&type=pdf>,
54
55 accessed 25 Mar 2021.
56
57
58
59
60

1
2
3 Aristóteles. *Categorias*. Porto: Porto Editora Ltda, 1995.

4
5
6 Aronson, Alan R., and François-Michel Lang. 2010. "An overview of MetaMap: historical perspective
7 and recent advances." *Journal of the American Medical Informatics Association* 17 no. 3: 229-236.

8
9
10
11 Barbosa, Nilson. T. and; Campos, Maria. L. de Almeida. 2017. "A questão da interoperabilidade em
12 repositórios institucionais e sistemas de informação de pesquisas correntes (cris): uma abordagem
13 preliminary". In *Encontro Nacional de Pesquisa em Ciência da Informação, n. XVIII ENANCIB,*
14 *2017*. <http://hdl.handle.net/20.500.11959/brapci/104600>, accessed 25 Dez. 2021.

15
16
17
18
19
20
21
22 Bergman, Mike. 2011. "Ontology-Driven Apps Using Generic Applications". *AI3 blog*.
23 <https://www.mkbergman.com/948/ontology-driven-apps-using-generic-applications/>.

24
25
26
27
28
29 Berners-Lee, Tim. 1998. "Cool URIs don't change".
30 <https://www.w3.org/Provider/Style/URI>.

31
32
33
34
35 Bodenreider, Olivier. (2004). "The unified medical language system (UMLS): integrating
36 biomedical terminology." *Nucleic acids research* 32 no. suppl_1: D267-D270.

37
38
39
40
41 Borst, Willem N. 1997. *Construction of Engineering ontologies*. Centre for Telematica and
42 Information Technology. University of Twente, Enschede, The Nederalands.

43
44
45
46 Cabré, María Teresa. 2005. A Terminologia, uma disciplina em evolução: passado, presente
47 e alguns elementos de futuro. *Debate Terminológico*. ISSN: 1813-1867, v1.
48 <https://www.seer.ufrgs.br/riterm/article/download/21286/15349>, accessed 21 Set. 2020.

49
50
51
52
53
54
55 Campos, Maria Luiza de Almeida. 2010. "O papel das definições na pesquisa em
56 ontologia". *Perspectivas em Ciência da Informação*, 15: 220-238
57 <https://www.scielo.br/j/pci/a/tJr4GnX9Xp7pj5pf44gK4yD/?lang=pt&format=html>.

1
2
3
4
5 Capurro, R. 2000. "Angeletics—A message theory. In H.H. Diebner & L. Ramsay (Eds.)
6 (2003), Hierarchies of communication". *In inter-institutional and international symposium*
7 *on aspects of communication on different scales and levels*. Karlsruhe, Germany: ZKM.
8 Retrieved July 25, 2005, from http://www.capurro.de/angeletics_zkm.html.
9
10
11
12
13
14
15

16
17 CERIF in Brief. (2014). [https://eurocris.org/eurocris_archive/cerifsupport.org/cerif-in-](https://eurocris.org/eurocris_archive/cerifsupport.org/cerif-in-brief/index.html)
18 [brief/index.html](https://eurocris.org/eurocris_archive/cerifsupport.org/cerif-in-brief/index.html)
19
20

21
22
23
24 Chen, Peter Pin-Shan. 1976. "The Entity-Relationship Model-Toward a Unified View of
25 Data". *ACM Transactions on Database Systems* 1 no.1: 9-36.
26
27
28
29

30
31 Chierchia, Gennaro. 2003. *Semântica*. São Paulo: Ed. UNICAMP.
32
33
34

35
36 CIDOC Conceptual Reference Model Version 5.1.12. 2014. ICOM/CIDOC.
37 <http://www.cidoc-crm.org/Version/version-5.1.2>, accessed May 3, 2015.
38
39
40
41
42

43 Codd, Eugene. F. 1970. "A relational model of data for large shared databanks".
44 *Communications of The ACM*, 13(6): 377-387.
45 https://dl.acm.org/doi/pdf/10.1145/362384.362685?casa_token=uOdxFTaktMAAAAAA:i_e
46 [wo3eO7rDNRE7VYvIBGeHn452O1VQGi69Jn13MciziUeGNMPy827WA6guuZzLkgq4D](https://dl.acm.org/doi/pdf/10.1145/362384.362685?casa_token=uOdxFTaktMAAAAAA:i_e)
47 [GI79ocfO4A.](https://dl.acm.org/doi/pdf/10.1145/362384.362685?casa_token=uOdxFTaktMAAAAAA:i_e)
48
49
50
51
52
53
54

55
56 Dahlberg, Ingetraut. 1978. "A referent-oriented, analytical concept theory for
57 INTERCONCEPT". *Knowledge Organization* 5 no. 3: 142-151 [https://www.ergon-](https://www.ergon-verlag.de/isko_ko/downloads/ic_5_1978_3.pdf#page=20)
58 [verlag.de/isko_ko/downloads/ic_5_1978_3.pdf#page=20](https://www.ergon-verlag.de/isko_ko/downloads/ic_5_1978_3.pdf#page=20).
59
60

1
2
3 Dhar, Vasant. 2013. "Data science and prediction". *Communications of the ACM* 56 no. 12:
4 64-73. <https://dl.acm.org/doi/pdf/10.1145/2500499>.

7
8 Dextre Clarke, Stella G. 2019. "The Information Retrieval Thesaurus". *Knowledge*
9 *Organization* 46 no. 6: 439-459. [https://www.ergon-](https://www.ergon-verlag.de/isko_ko/downloads/ko_46_2019_6_c.pdf)
10 [verlag.de/isko_ko/downloads/ko_46_2019_6_c.pdf](https://www.ergon-verlag.de/isko_ko/downloads/ko_46_2019_6_c.pdf).

13
14 Dextre Clarke, Stella G. and Zeng, Marcia Lei. 2012. "From ISO 2788 to ISO 25964: The
15 evolution of thesaurus standards towards interoperability and data modelling". *Information*
16 *Standards Quarterly (ISQ)* 24 no. 1.
17 http://eprints.rclis.org/16818/1/SP_clarke_zeng_isqv24no1.pdf.

18
19 Dierickx, Harold and Hopkinson, Alan. 1986. *Reference manual for machine-readable*
20 *bibliographic descriptions*.
21 http://biblio.cerist.dz/hrbdonf5214/ouvrages/00000000000000594806000000_2.pdf.

22
23 FAIR Compliant Biomedical Metadata Templates. 2019. CEDAR, Center for Expanded
24 Annotation and Retrieval, University of Stanford, Department of Medicine.
25 <https://medicine.stanford.edu/2019-report/cedar-to-the-rescue.html>.

26
27 Floridi, Luciano. 2019. "Semantic Conceptions of Information". In *The Stanford*
28 *Encyclopedia of Philosophy* (Winter 2019 Edition), Edward N. Zalta (ed.).
29 <https://plato.stanford.edu/archives/win2019/entries/information-semantic/>.

30
31 Foskett, A. C. (1996). *"The subject approach to information"*. Facet Publishing.

32
33 Fonseca, Claudenir M., Porello, Daniele, Guizzardi, Giancarlo, Almeida, João Paulo A. and
34 Guarino, Nicola. (2019). *Relations in Ontology-Driven Conceptual Modeling*. In Laender,
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 A., Pernici, B., Lim, EP., de Oliveira, J. (eds) *Conceptual Modeling. ER 2019. Lecture*
4
5 *Notes in Computer Science* 11788. Springer, Cham. [https://doi.org/10.1007/978-3-030-](https://doi.org/10.1007/978-3-030-33223-5_4)
6
7 [33223-5_4](https://doi.org/10.1007/978-3-030-33223-5_4).
8

9
10
11 Fillinger, Sven et al. 2019. "Challenges of big data integration in the life
12 sciences." *Analytical and bioanalytical chemistry* 411 no. 26: 6791-6800.
13
14 doi:10.1007/s00216-019-02074-9
15
16

17
18
19
20
21 Freitas, C.; Carvalho, P.; Oliveira, H. G.; Mota, C. and Santos, D. 2010. "Second HAREM:
22
23 advancing the state of the art of named entity recognition in Portuguese". In Nicoletta
24 Calzolari et al. (eds.), *Proceedings of the International Conference on Language Resources*
25 *and Evaluation (LREC 2010)*. European Language Resources Association, pp. 3630-3637.
26
27 Valletta, 2010.
28
29
30

31
32
33 Frické, Martin. 2015. "Big Data and Its Epistemology". *Journal of the Association for*
34 *Information Science and Technology* 66 no. 4: 651-61.
35
36
37

38
39
40 Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and
41
42 analytics." *International journal of information management* 35.2 (2015): 137-144.
43
44
45

46
47
48 Gershenfeld, Nel, Krikorian, Raffi, and Cohen, Danny. 2004. "The Internet of Things".
49 *Scientific American*, October: 76-81. Available: <http://cba.mit.edu/docs/papers/04.10.i0.pdf>.
50
51
52 Accessed May 5 2021.
53
54
55
56
57
58
59
60

1
2
3 Giunchiglia, Fausto; Dutta, Biswanath and Maltese, Vincenzo. 2014. "From knowledge
4 organization to knowledge representation". *Knowledge Organization* 41 no. 1: 44-56.
5
6 <http://eprints.biblio.unitn.it/4186/1/techRep027.pdf>.

7
8
9
10
11 Gray, Jim. 2009. "eScience: A Transformed Scientific Method". In *The Fourth Paradigm,*
12 *Data-intensive Scientific Discovery*, ed. Tony Hey, Stewart Tansley and Kristin Tolle.
13 Redmond, Wash.: Microsoft Research, 19-33. available at:
14
15 <http://itre.cis.upenn.edu/myl/JimGrayOnE-Science.pdf>.

16
17
18
19
20
21 Guarino, Nicola. 1997. "Semantic matching: Formal ontological distinctions for information
22 organization, extraction, and integration". In *International Summer School on Information*
23 *Extraction*. Springer, Berlin, Heidelberg, 1997. 139-170.
24
25 [https://kask.eti.pg.gda.pl/redmine/projects/sova/repository/revisions/5378040326bc499e118](https://kask.eti.pg.gda.pl/redmine/projects/sova/repository/revisions/5378040326bc499e118636a1d25ad667285e005c/entry/Praca_dyplomowa/materialy/10.1.1.53.939.pdf)
26
27 [636a1d25ad667285e005c/entry/Praca_dyplomowa/materialy/10.1.1.53.939.pdf](https://kask.eti.pg.gda.pl/redmine/projects/sova/repository/revisions/5378040326bc499e118636a1d25ad667285e005c/entry/Praca_dyplomowa/materialy/10.1.1.53.939.pdf).

28
29
30
31
32
33 Guarino, Nicola; Carrara, Massimiliano and Giaretta, Pierdaniele. 1994. "Formalizing
34 ontological commitment". In *AAAI*. 1994. p. 560-567.
35
36 <https://www.aaai.org/Papers/AAAI/1994/AAAI94-085.pdf>.

37
38
39
40
41 Gruber, Thomas R. 1993. "A translation approach to portable ontology
42 specifications." *Knowledge acquisition* 5 no. 2: 199-220.

43
44
45
46
47 Hajibayova, Lala, and Athena Salaba. 2018. "Critical questions for big data approach in
48 knowledge representation and organization." *Challenges and Opportunities for Knowledge*
49 *Organization in the Digital Age: Proceedings of the Fifteenth International ISKO*
50 *Conference 9-11 July 2018 Porto, Portugal*, Vol. 16. Ergon Verlag.

51
52
53
54
55
56
57 He, Yongqun, et al. 2020. "CIDO, a community-based ontology for coronavirus disease
58 knowledge and data integration, sharing, and analysis." *Scientific data* 7 no. 1: 1-5.

1
2
3 Hey, Tony; Trefethen, Anne. 2003. "The data deluge: An e-science perspective". In *Grid*
4 *computing: Making the global infrastructure a reality*, p. 809-824.
5
6 https://eprints.soton.ac.uk/257648/1/The_Data_Deluge.pdf.
7
8
9

10
11
12
13 Hjørland, Birger. (2018). "Data (with big data and database semantics)". *Knowledge*
14 *Organization* 45 no. 8: 685-708.
15
16

17
18 Hjørland, Birger. (2002). "Domain analysis in information science: eleven approaches—
19 traditional as well as innovative". *Journal of Documentation*, 58 no. 4: 422-462.
20
21

22
23
24 Hjørland, Birger. 2013. "Theories of knowledge organization — theories of knowledge.",
25 *Knowledge Organization* 40: 169–181.
26
27

28
29
30 Hjørland, Birger, and Albrechtsen, Hanne. (1995). "Toward a new horizon in information
31 science: Domain-analysis". *Journal of the American society for information science* 46 no.
32 6: 400-425.
33
34
35

36
37
38 Hjørland, Birger and Hartel, Jenna. 2003. "Introduction to a special issue of Knowledge
39 Organization". *Knowledge Organization* 30 no. 3/4: 125-7.
40
41

42
43 Iafate, Fernando. (2015). *From Big Data to Smart Data*. London: ISTE Ltd., and Hoboken,
44 NJ: John Wiley & Sons, Inc.
45
46
47

48
49 Ibekwe-SanJuan, Fidelia and Geoffrey C. Bowker. 2017. "Implications of Big Data for
50 Knowledge Organization". *Knowledge Organization* 44, no. 3: 187-98.
51
52

53
54 International Council on Archives. Experts Group on Archival Description. 2019. Records in
55 Context: A Conceptual Model for Archival Description (Consultation Draft v0.1). ICA.
56
57 https://www.ica.org/sites/default/files/ric-cm-0.2_preview.pdf, accessed December 12, 2018.
58
59
60

1
2
3 International Federation of Library Associations and Institutions (IFLA). 1998. *Study Group*
4 *on Functional Requirements for Bibliographic Records: Final Report*. UBCIM Publications
5
6 New Series. München: K. G. Saur.

7
8
9
10
11
12 ISO/IEC 20546:2019(en). 2019. *Information technology — Big data — Overview and*
13 *vocabulary*. ISO.

14
15
16
17
18 ISO 25964-2 (2013). *Information and documentation — Thesauri and interoperability with*
19 *other vocabularies — Part 2: Interoperability with other vocabularies*. ISO, 2013.

20
21
22
23
24 Kahn, Robert; Wilensky, Robert. 2006. “A Framework for Distributed Digital Objects
25 Services”. *International Journal on Digital Libraries* 6 no. 2: 115–123.
26
27 https://www.doi.org/topics/2006_05_02_Kahn_Framework.pdf. Access: June 28, 2022.
28
29

30
31
32
33 Lambe, Patrick. 2014. *Organising knowledge: taxonomies, knowledge and organizational*
34 *effectiveness*. Elsevier.

35
36
37
38
39 Leonelli, Sabina. 2012. “Classificatory Theory in Data-intensive Science: The Case of Open
40 Biomedical Ontologies”. *International Studies in the Philosophy of Science* 26 no. 1: 47–65.

41
42
43
44
45 _____ and Costa, Leonardo C. da. 2016. “A Model to Represent and
46 Process Scientific Knowledge in Biomedical Articles with Semantic Web Technologies”.
47 *Knowledge Organization*, 43(2): 122-137. [https://www.ergon-](https://www.ergon-verlag.de/isko_ko/downloads/ko_43_2016_2_b.pdf)
48
49 [verlag.de/isko_ko/downloads/ko_43_2016_2_b.pdf](https://www.ergon-verlag.de/isko_ko/downloads/ko_43_2016_2_b.pdf), accessed Apr. 12, 2017.
50
51

52
53
54 _____ and Dias, Celia. 2020. “Representing facet classification in SKOS”.
55 In International ISKO Conference, Aalborg, Denmark, 16th, *Proceedings...*
56
57 *1. Edition*. Würzburg: Ergon Verlag. ISBN print: 978-3-95650-775-5, ISBN online: 978-3-
58
59
60

1
2
3 95650-776-2, *Series: Advances in knowledge organization* 9. Würzburg: Ergon
4 Verlag, 254–263. <https://doi.org/10.5771/9783956507762>, accessed Feb. 15, 2021.
5
6
7
8
9

10 _____; Martins, Sergio. C. and Ramos Junior, Mauricio. C. 2021. The role
11 of vocabularies for the access and reuse of Big Data. *Informação & Informação*, 26 no. 4:
12 146-174. <https://www.uel.br/revistas/uel/index.php/informacao/article/view/44653/pdf>.
13
14
15
16
17 Access 5 Jan. 2022.
18
19

20
21
22 De Mauro, Andrea; Greco, Marco and Grimaldi I, Michele. 2015. “What is big data? A
23 consensual definition and a review of key research topics”. In *AIP conference proceedings*.
24 American Institute of Physics, 2015. p. 97-104. [http://big-data-fr.com/wp-](http://big-data-fr.com/wp-content/uploads/2015/02/aip-scitation-what-is-bigdata.pdf)
25
26
27
28
29
30
31 content/uploads/2015/02/aip-scitation-what-is-bigdata.pdf.

32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
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.

1
2
3 Orilia, Francesco and Paoletti, Michele Paolini. 2020. "Properties", *The Stanford*
4 *Encyclopedia of Philosophy* (Winter 2020 Edition), Edward N. Zalta (ed.).
5
6 <https://plato.stanford.edu/archives/win2020/entries/properties/>, accessed 20 Sept. 2021.
7
8

9
10
11 Otlet, Paul. (2018). *Tratado de Documentação: o livro sobre o livro, teoria e prática*.
12
13 Brasília: Briquet de Lemos Livros.

14
15
16
17
18 Peirce, Charles. S. 1869 . "On a new list of categories". *Proceedings of the American*
19 *Academy of Arts and Sciences*, v. 7, p. 287-298, 1868. Disponível em: <
20 <http://www.bocc.ubi.pt/pag/peirce--charles-list-categories.pdf>>. Access July 28, 2021.
21
22
23

24
25
26 Poole, Alex H. "Now is the Future Now? 2013. "The Urgency of Digital Curation in the Digital
27
28 Humanities." *DHQ: Digital Humanities Quarterly* 7 no. 2.

29
30
31 Prasad, A. R. D., Giunchiglia, Fausto; Devika, P. Madalli. 21017. "DERA: from document
32
33 centric to entity centric knowledge modelling". In: *Proceedings of the International UDC*
34 *seminar 2017. Faceted classification today*. London: September, 2017. p. 169-179.
35
36 <http://seminar.udcc.org>.
37
38

39
40
41 Prieto-Díaz, Ruben. 1990. "Domain analysis: An introduction". *ACM SIGSOFT Software*
42
43 *Engineering Notes*, 15 no. 2: 47-54.

44
45
46
47 Ranganathan, S. R. and Gopinath, M. A. *Prolegomena to Library Classification*. 3 ed.
48
49 Bombay: Asia Publishing House, 1967.

50
51
52 RDF semantics. W3C, 2004. <http://www.w3.org/TR/rdf-mt/>, accessed Mar, 10, 2010.
53

54
55
56 RDF 1.1. PRIMER. 2014. W3C. <https://www.w3.org/TR/rdf11-primer/>, accessed 12 Dez.
57
58 2019.
59
60

1
2
3 Resource Description Framework (RDF) Model and Syntax Specification. W3C, 1998.
4
5 <https://www.w3.org/1998/10/WD-rdf-syntax-19981008/>. Accessed May 5, 2011.
6
7

8 Riva, Pat, Le Boeuf, Patrick, and Žumer, Maja. 2017 “*IFLA Library Reference Model: A*
9
10 *Conceptual Model for Bibliographic Information*”. IFLA. [online]
11
12 <https://www.ifla.org/publications/node/11412> (Accessed 23 March 2019)
13
14

15
16 Rowley, Jennifer. 2007. “The wisdom hierarchy: representations of the DIKW hierarchy”.
17
18 *Journal of information science* 33 no. 2: 163-180.
19
20 <http://web.dfc.unibo.it/buzzetti/IUcorso2007-08/mdidattici/rowleydikw.pdf>, access Jul 14
21
22 2013.
23
24

25
26 Saracevic, Tefko. 2007. “Relevance: A review of the literature and a framework for thinking
27
28 on the notion in information science. Part II: Nature and manifestations of
29
30 relevance”. *Journal of the american society for information science and technology* 58 no.
31
32 13: 1915-1933.
33
34

35
36 Shet, Amith. 2020. “Knowledge Graphs and their central role in big data processing: Past,
37
38 Present, and Future”. In 7th ACM India Joint Conference on Data Science & management of
39
40 Data (COD-COMAD), Indian School of Business, Hyderabad Campus, 5-7 January 2020.
41
42 [https://www.slideshare.net/apsheth/knowledge-graphs-and-their-central-role-in-big-data-](https://www.slideshare.net/apsheth/knowledge-graphs-and-their-central-role-in-big-data-processing-past-present-and-future)
43
44 [processing-past-present-and-future](https://www.slideshare.net/apsheth/knowledge-graphs-and-their-central-role-in-big-data-processing-past-present-and-future), accessed Jun. 5, 2021.
45
46
47
48

49
50 Shet, Amith; Ramakrishnan, Cartic and Thomas, Christopher. 2005. “Semantics for the
51
52 semantic web: The implicit, the formal and the powerful”. *International Journal on*
53
54 *Semantic Web and Information Systems (IJSWIS)*, no. 1 vol. 1:1-18.
55
56 <http://www.ebusinessforum.gr/old/content/downloads/JSWIS.pdf#page=19>, accessed Jul 14,
57
58 2010.
59
60

1
2
3
4
5 Shiri, Ali. 2013. "Linked data meets big data: A knowledge organization systems
6 perspective." *Advances in Classification Research Online* 24 no. 1: 16-20.
7
8

9
10
11
12 SKOS – Simple Knowledge Organization System Namespace Document. W3C, 2012.
13
14 <https://www.w3.org/2009/08/skos-reference/skos.html#>, accessed Aug 10, 2013.
15
16

17
18
19 Soergel, Dagobert. 2015. "Unleashing the Power of Data Through Organization: Structure
20 and Connections for Meaning, Learning and Discovery." *Knowledge Organization* 42 no. 6:
21 401-427.
22
23
24

25
26
27
28 SPARQL 1.1 QUERY LANGUAGE, 2013. W3C. <https://www.w3.org/TR/sparql11-query/>,
29 accessed 12 Fev. 2010.
30
31

32
33
34 Strecker, Dorothea et al. 2021. *Metadata Schema for the Description of Research Data*
35 *Repositories*. Re3data, 2021. Available at: <https://doi.org/10.48440/re3.010>. Access 08 Jul.
36
37
38
39 2022.
40

41
42 Swanson, Don R. (2008). "Literature-based discovery? The very idea." In *Literature-based*
43 *discovery*. Springer, Berlin, Heidelberg. 3-11.
44
45
46

47
48
49
50 Veiga, Viviane Santos de Oliveira; Campos, Maria Luiza; Silva, Carlos Roberto Lyra;
51
52 Henning, Patricia and Moreira, João. 2021. "Vodan br: a gestão de dados no enfrentamento
53 da pandemia coronavirus". *Páginas A&B, Arquivos e Bibliotecas (Portugal)*, n. Especial:
54 51-58. <http://hdl.handle.net/20.500.11959/brapci/157353>, accessed Out 7, 2021.
55
56
57
58
59
60

1
2
3 Wilson, Thomas. D. 1972. "The work of the British Classification Research Group". In
4
5 Wellish, H. (ed). Subject retrieval in the seventies. Westport: Greenword Publishing Co.,
6
7
8 62-71.
9

10
11
12
13 Zeng, Marcia Lei. 2019. "Interoperability". *Knowledge Organization* 46, no. 2: 122-146.
14
15 Also available in Hjørland, Birger and Gnoli, Claudio eds. *ISKO Encyclopedia of*
16
17 *Knowledge Organization*, <http://www.isko.org/cyclo/interoperability>.
18
19

20
21 Zeng, Marcia. L. (2017). "Smart data for digital humanities". *Journal of Data and*
22
23 *Information Science* 2 no. 1: 1-12. DOI: 10.1515/jdis-2017-0001.
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 intertwined. 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.

1
2
3 Big Data reminds the so-called “information explosion,” a phenomenon connected to the rise
4 of Information Science and KO. In response, KO created knowledge organization systems
5 (KOS) that work in conjunction with information retrieval systems (IRS), computerized
6 databases containing representations of scientific documents. Such KOS, the “information
7 retrieval thesaurus” (Dextre Clarke 2016, 138), control and standardize the natural language
8 used both for indexing the documents entered in the IRS and the keywords used in the user's
9 queries.
10
11

12
13
14
15
16
17
18
19
20 Most conceptualizations of Big Data tend to emphasize technological aspects such as volume,
21 variety, velocity, heterogeneity, and the need for massive computer power to process it
22 (Gandomi and Haider 2015). Big Data has also been sparking interest in KO (Ibekwe-SanJuan
23 and Bowker 2017, 192), raising questions like its impact in KO epistemology and
24 methodologies (Hajibayova and Salaba, 2018), (Frické 2015). However contributions from
25 the area to propose practical solutions are still few; Hjørland (2013, 179) stressed: “But such
26 progress is brought to us from the outside; it is not something the field of KO has provided”.
27 The availability today of huge datasets recording user interactions with different systems, their
28 interests, and preferences, gave rise to the development of data-driven methodologies to guide
29 the interactions between users and such systems, including IRS, an area of application of KO.
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

1
2
3 of information technologies. Examples of this huge amount of digital survey data are those
4 generated by the Hubble Space Telescope,
5 https://www.nasa.gov/mission_pages/hubble/main/index.html, the Human Genome research
6 project, <https://www.genome.gov/human-genome-project>, or the Large Hadron Collider,
7 <https://home.cern/science/accelerators/large-hadron-collider>, the largest and most powerful
8 particle accelerator in the world. A large amount of digital research data now available has
9 even raised debates concerning scientific methodology (Gray 2009), (Leonelli 2012), (Frické
10 2015).

11
12
13
14
15
16
17
18
19
20
21
22 Research data is defined as “factual records (numerical scores, textual records, images, and
23 sounds) used as primary sources for scientific research, and that are commonly accepted in
24 the scientific community as necessary to validate research findings.” (OECD, 2007, 13). Share
25 and reuse of research data presupposes its openness but not only that. As quoted by researcher
26 the Mark Musen at the beginning of this work: “the metadata — the machine-readable
27 descriptions that provide data about the data”, has been gaining increasing importance.
28 Vocabularies, i.e., data vocabularies or metadata vocabularies (Zeng 2019), is an important
29 research area in KO. Musen’s observation refers to the Semantic Web project (Berners_Lee
30 et al 2001), the proposal for a Web whose resources would be represented in a way that had a
31 precise and formal meaning or semantics and would be intelligible and understandable by both
32 people and machines.
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

48 1.2. The document centered vision of vocabularies

49
50
51
52 The technical traditions and standards developed by KO to manage the information explosion
53 rest on assumptions that persist to this day. In most discourses in the area, these assumptions
54 are so implicit that it becomes difficult to make them explicit, consider them, and analyse their
55 consequences. All the theories and methodologies of KO mentioned bringing these
56
57
58
59
60

1
2
3 assumptions implicitly: the IRS represent documents in their computerized databases; MARC
4 and the bibliographic formats that emerged from the UNISIST Reference Manual for
5 machine-readable bibliographic descriptions (Dierickx and Hopkinson, 1986) (Reviewer 2)
6 are metadata sets that represent different descriptive properties of the documents.
7
8
9

10
11
12
13 KOS associated with IRS confirm such assumptions; they “have been designed to support the
14 organization of knowledge and information to make their management and retrieval easier”
15 (Mazzocchi 2018). They are terminological control instruments used to standardize the
16 records’ subject and authorities fields in IRS computerized databases, so useful for users’
17 subject-based (Foskett 1996) retrieval.
18
19
20
21
22
23
24
25

26 Representing documents and their subjects is a practice with a long tradition in KO. In the
27 past such documents surrogates were a fundamental mechanism to provide access to
28 information and enable processes of relevance assessment carried out by libraries and IRS
29 (Saracevic 2007). KO methodologies have always represented domains of knowledge when
30 building KOS like controlled/standardized vocabularies, subject headings, and taxonomies
31 KOS, such as thesauri, were intended to enable subject-based retrieval in the context of IRS
32 because their records were representations of objects that have as one of their properties
33 subjects. But not all objects in a domain have subjects as one of their properties like
34 documents. We see now that this is just one among many cases of representing different
35 objects in digital space.
36
37
38
39
40
41
42
43
44
45
46
47
48
49

50 To what extent do these assumptions hold up today, and are they sufficient to address the
51 challenges of the Semantic Web era, Big Data, research data, and the Internet of Things?
52 Today, it is not only the case of retrieving documents (or their representations) but also to
53 create digital representations of anything, as demanded by the “Internet of Things” (IoT)
54 (Gershenfeld, Krikorian, and Cohen 2004). If the documentation movement (Otlet 2018) and
55
56
57
58
59
60

1
2
3 then Information Science empowered information by separating it from books, the Semantic
4 Web proposal and Big Data did the same with the knowledge (Soergel 2015). It is no longer
5 just inserted into texts to be interpreted by humans, but rather serialized in Resource
6 Description Framework (RDF) triples (RDF 1.1 PRIMER 2014), forming
7 representations/descriptions of “things”.
8
9

10
11
12
13
14
15 The objective of this work is to discuss how vocabularies, in the sense used within LOD
16 Technologies i.e., value vocabularies, or KOS, and metadata vocabularies (Zeng 2019), can
17 contribute to assigning computational semantics to digital research data within the context of
18 Big Data, so that computers can process them, allowing their reuse on large scale. Descriptive
19 metadata sets represent specific entities, or resources in the Web context; value vocabularies
20 assign standardized data values to specific descriptive items of entity instances described by
21 metadata vocabularies.
22
23
24
25
26
27
28
29
30
31

32
33 As a methodology, the work develops a conceptualization of data in an attempt to make it
34 clearer what would be data, as an essential element of the Big Data phenomenon, and in
35 particular, digital research data. It then proceeds to analyse digital research data uses and cases
36 and their relation to semantics and vocabularies.
37
38
39
40
41

42
43 The work is organized as follows. After this introduction, section 2 analyses data from a
44 semiotic and ontological point of view. Section 3 presents a comprehensive view of
45 vocabularies within the context of Semantic Web and LOD. Within such a context Section 4
46 develops a conceptualization of data that is illustrated by examples of research data, research
47 datasets, and related initiatives, and shows how research data at different levels of aggregation
48 yields semantics. Section 5 draws conclusions, raises research questions to be developed and
49 presents final considerations.
50
51
52
53
54
55
56
57
58
59

60 **2. Semiotic and ontological view of data**

1
2
3 None of the most common Big Data definitions exclude the data component. It seems
4 reasonable, then, that to understand what Big Data is and how to operationalize solutions to
5 the problem begins by elucidating what is data. After presenting the traditional use of
6 vocabularies to represent and assign subjects to documents this section proposes a semiotic
7 and ontological analysis of data, understood as the essential component of Big Data and
8 research data. This analysis begins with the question of conceptual models and domains and
9 goes on to analyse how conceptual models of domains are expressed linguistically as
10 vocabularies. Then data is discussed from a semiotic and ontological point of view.
11
12
13
14
15
16
17
18
19
20
21

22 2.1. Vocabularies as representations of domains

23
24
25 In the 1980s-1990s, as a consequence of the emergence of online bibliographic catalog
26 management systems and databases, the domain of information retrieval in library catalogues,
27 so familiar to us but also so exclusive, with its diversity of objects, was first modelled using a
28 methodology used in computer science to plan database management systems. The Functional
29 Requirements for Bibliographic Records conceptual model (FRBR) based on Chen (1976)
30 Entity-Relationship (E-R) model, appeared in 1998, whose development was promoted by
31 IFLA (1998).
32
33
34
35
36
37
38
39
40
41
42

43 According to Mylopoulos (1992, 3) “Conceptual modeling is the activity of formally
44 describing some aspects of the physical and social world around us for purposes of
45 understanding and communication.” For Mylopoulos,
46
47
48
49

50 the descriptions that arise from conceptual modeling activities are intended to
51 be used by humans, not machines. . . [and] The adequacy of a conceptual
52 modeling notation rests on its contribution to the construction of models of
53 reality that promote a common understanding of that reality among their
54 human users.
55
56
57
58
59
60

1
2
3 A conceptual model sets an agreement between users of a system on what kinds of things exist
4 and will be represented in the system, or entities (also called classes) in a given domain of
5 reality, e.g. documents of historical value, the properties of these entities and how they relate
6 to each other (relationships). Thus, a conceptual model is a representation, in the form of an
7 abstract and generic description, independent of computational implementations (hardware,
8 operating systems, languages, database management systems) of a given domain of reality. It
9 aims at understand this reality, reason about it, and establish a common view of this reality; a
10 conceptual model answers questions such as: What different things exist in a given domain?
11 How are they distinguished from each other? How do they relate? What are their properties?
12
13

14 As a representation, a conceptual model is expressed, communicated, and externalized
15 through a language, or more specifically a meta-language or meta-model (Guizzardi 2007,
16 23), which is a language to express the vocabulary (concepts, terms) that express things in
17 specific domains. Examples of these meta-languages are either natural language (through a
18 system requirements document), which functions as the most general of all meta-languages,
19 or a diagrammatic meta-language, such as entity-relationship (meta) Model or the Unified
20 Modelling Language (UML), <https://www.uml.org/>, class diagram, in which domain-specific
21 ER models or class diagrams are expressed.
22
23

24 Within descriptive representation, once established and consolidated practical standards such
25 as MARC, UNISIST, AACR2 and ISDB, the question of what are the "things" represented is
26 raised, a view with a higher level of abstraction of a domain.
27
28

29 Conceptual models in the area of documentation and information have made things like
30 documents, authors, and subjects explicit. They evolved from the previously mentioned
31 standards for creating automated bibliographic records, starting with the pioneering FRBR
32 (Ifla 1998). FRBR, as a conceptual model of the bibliographic domain, is not intended for
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 describing or indexing documents, but for formalizing, identifying, agreeing, and
4
5 standardizing objects, actors, and processes and their relationships within such domain.
6
7

8
9 Universal bibliographic classification systems such as the Dewey Decimal Classification
10 (DCC) – and the Universal Decimal Classification (UDC) are used for thematic
11 representation, for assigning subjects – as discipline names - to books. They model the
12 universe of knowledge as a set of taxonomies, each having as a root a discipline. The use of
13 taxonomies to organize a domain is typically used today for information management within
14 corporations and to organize the content of websites (Lambe 2014). Taxonomies only
15 organize the things in a domain in class-subclass relationships. The things being organized in
16 a universal bibliographic classification are discipline names to be used as subjects to books.
17
18
19
20
21
22
23
24
25
26

27
28 However, there are more than just things or taxonomies of things in a domain. A more accurate
29 model of a domain should include also their properties, relationships and attributes, according
30 to the ER model. The first movement within documentation and information to recognize this
31 fact was Faceted Classification (Ranganathan, Gopinath 1967). Facets are the properties of a
32 class of things of interest for information recovery (Giunchiglia et al 2014; _____ and
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

1
2
3 activities and methodologies, from records describing documents and their subjects to
4
5 conceptual modeling, that is, representing entities, their attributes and relationships (Prasad,
6
7 Giunchiglia, Devika 2007). Knowledge organization and representation is part of the digital
8
9 research data curation effort. Such domains of application also uses conceptual models to
10
11 integrate heterogeneous research data sources as publications, research data, patents, projects,
12
13 events, funding agencies, etc. (CERIF in Brief 2014)
14
15

16
17
18 Conceptual models are aligned together with different types of KOS by Almeida, Souza and
19
20 Fonseca (2011, 196), ordered according to their semantic expressiveness. Semantic
21
22 expressiveness can be understood, in the context of the previous quote, as the ability of each
23
24 type of KOS to distinguish and describe, that is, identify the properties and represent the
25
26 different things that exist in a domain of that reality.
27
28

29
30 Conceptual model elements - entities, attributes and relationships - are expressed linguistically
31
32 by a vocabulary. Vocabularies are semantic control devices, formed by systematised sets of
33
34 semiotic, triadic entities (PEIRCE 1994), concepts (Dahlberg 1978), units of meaning that
35
36 relate something (a first: object or referents), in some way (through a second: term or code),
37
38 which generates or induces a third: its meaning.
39
40

41 42 43 2.2. Domains 44

45
46 Aside from the general library classification systems such as the CDD and the CDU, KOS are
47
48 developed and used concerning specific domains. The domain notion commonly used in KO
49
50 is that of a specialized knowledge area.
51

52
53 Hjørland and Albrechtsen (1995, 400), in the text in which they propose the analysis of
54
55 domains as the foundation of KO, define domains as: “thought or discourse communities,
56
57
58
59
60

1
2
3 which are parts of society's division of labour." They also label a domain as a
4
5 "specialty/discipline/domain/environment" (Hjørland and Albrechtsen 1995, 401).
6
7

8
9 Hjørland (2002, 422) conceptualizes domains associated with specialized libraries,
10
11 questioning what knowledge would be necessary for information professionals to work in "in
12
13 a specific subject field like medicine, sociology or music?" In Hjørland and Hartel (2003,
14
15 239), this view of domains as systems of thought, theories, is reaffirmed.
16
17

18
19 Domains are basically of three kinds of theories and concepts: (1) ontological
20
21 theories and concepts about the objects of human activity; (2) epistemological
22
23 theories and concepts about knowledge and the ways to obtain knowledge,
24
25 implying methodological principles about the ways objects are investigated;
26
27 and (3) sociological concepts about the groups of people concerned with the
28
29 objects.
30
31

32
33 The oldest thesaurus were intended to enable subject-based retrieval in the context of IRS
34
35 because their records were representations of objects that had subjects as one of their
36
37 properties, that is, documents. Today, it is not just about retrieving documents (or their
38
39 representations) but digital representations of anything, as exemplified in the IoT. These
40
41 representations are no longer just access points for documents, but also information resources
42
43 themselves, complex descriptions of these objects, and sources of knowledge about them,
44
45 represented in such a way that they can be processed/intelligible by both machines and
46
47 humans. Such representations allow machines to make inferences about the knowledge thus
48
49 represented.
50
51

52
53
54 KO today is being called upon to model different domains of knowledge to build new
55
56 "semantic" vocabularies, i.e, vocabularies compliant with the Semantic Web and LOD
57
58 technologies. For this, it is necessary to expand the traditional notion of a domain as a
59
60

1
2
3 discipline or subject. In the area of software development the notion of a domain has a broader
4 scope: it is ‘a sphere of activity or interest: field’ [Webster]. In the context of software
5 engineering, it is most often understood as an application area, a field for which software
6 systems are developed (Prieto Díaz 1990, 50).
7
8
9
10

11
12
13 Since a vocabulary is a terminological system that represents the “things” of interest in a
14 domain of action to the community of agents/users in that domain, then to create a vocabulary
15 (an artifact, similar to software) several aspects and questions must be considered: What things
16 are in a domain? How should they be represented? These are the questions of ontology and
17 semiotics. They must be answered to create a representation, or a conceptual model, of a
18 domain.
19
20
21
22
23
24
25
26

27
28 A first step is to determine what things exist in a domain and which are relevant to this
29 community, what rules exist about these things or are created/approved/agreed on about these
30 things, and how this community uses them to act in this domain. Finally, how the
31 conceptualizations and their agreed terms (Dahlberg 1978), one of the by-products of this
32 process, are to be systematised in a domain model to serve as bases for the construction of
33 vocabularies such as thesaurus or computational ontologies.
34
35
36
37
38
39
40
41
42
43
44
45

46 As shown, vocabularies can be representations of domains. A domain vocabulary can be used
47 either to assign subjects to documents: a) (e.g. MeSH categories describing the entities within
48 the Healthcare domain, <https://meshb.nlm.nih.gov/treeView>, or b) to describe objects in this
49 domain, descriptive metadata standards that, in addition to identify what things exist in a
50 domain, also describe their properties – attributes and relationships. Among the things within
51 a domain some vocabularies focus on specific facets for special purposes: archival science
52 and records management uses functional classification plans in an organization to assign the
53
54
55
56
57
58
59
60

1
2
3 organizational provenance or the function or organizational process that generated or used a
4
5 record.
6
7

8 9 2.3. Data as Representations 10

11
12 What is Big Data? What is its relationship with data? What is data and how is it related to
13
14 metadata? How should semantics be assigned to data? As noted in the ISO/IEC 20546/2019
15
16 Standard, “The big data paradigm is a rapidly changing field with rapidly changing
17
18 technologies,” later suggesting a definition: “extensive datasets (3.1.11) — primarily in the
19
20 data (3.1.5) characteristics of volume, variety, velocity, and/or variability — that require a
21
22 scalable technology for efficient storage, manipulation, management, and analysis.”
23
24
25

26
27 The conceptualizations of Big Data define it as a phenomenon that involves large amounts of
28
29 data, the heterogeneity of that data, a continuous flow of generation and updating, and a need
30
31 for large processing capacity so that the data reveal patterns or trends (De Mauro et al 2015).
32
33 However, the same is not true for the conceptualizations of data originating from KO. Data is
34
35 mentioned frequently in the literature, along with its relationships with information and
36
37 knowledge (Buckland 1991), often called the data, information, knowledge, wisdom (DIKW)
38
39 hierarchy (Rowley 2007). In Floridi (2019), information is related to data and semantics.
40
41
42

43
44 An important exception is from Hjørland (2018), who proposes a conceptualization of Big
45
46 Data arising from definitions of data, a phenomenon much better known and conceptualized
47
48 within KO. Data is in the essence of the Big Data phenomenon, it could not exist without data.
49
50 In this work, Hjørland lists several similar conceptualizations of data and highlights that of
51
52 Fox and Levitin:
53
54

55
56
57 Within this framework, we define a datum or data item, as a triple $\langle e, a, v \rangle$,
58
59 where e is an entity in a conceptual model, a is an attribute of entity e , and v
60

1
2
3 is a value from the domain of attribute a. A datum asserts that entity and has
4
5 value v for attribute a. Data are the members of any collection of data items.
6
7

8
9 Such conceptualization is clarified by the following example: “2018.” What does 2018 mean?
10
11 Others would say it’s a given. Let us note, however, this statement: “Giovana was born in
12
13 2018.” In it we can identify the entity we are talking about: a child called “Giovana,” an
14
15 attribute or property of this entity, she is “born,” and the value of this attribute or property,
16
17 her birth year, “2018.” To achieve a formal representation it is very important to clearly
18
19 identify the entity being described. Although a data set usually has a title or description
20
21 identifying the entity it represents that is not always the case. A metadata set may mix
22
23 metadata elements of different entities as for example the MARC21 format field 245 – Title
24
25 Statement; while MARC21 format describes a bibliographic entity, e.g., a book, field 245
26
27 subfield code \$c describes another entity, the responsible for the book, and field 245 subfield
28
29 \$f its attributes birth and death dates.
30
31
32
33

34
35 In the ontological scheme that goes back to Aristotle (2000), the reality is constituted of the
36
37 first substances, the things that have real existence in space and time, and second substances,
38
39 the conceptualizations we make of the first substances to think, reason, make sense of, and
40
41 communicate about the things in reality. Second substances are in turn subdivided into
42
43 essences, concepts designating things that have properties whose loss implies the non-
44
45 existence of that individual and have existential independence (Fonseca, Porello, Guizzardi,
46
47 Almeida, and Guarino 2019, 29), and accidents, concepts that designate things that are
48
49 existentially dependent on other substances. Things having existential independence are
50
51 commonly recognized in one of the most well-known ontological schemes, the entity-
52
53 relationships (ER) model (Chen 1976) as entities, while those that are existentially dependent,
54
55 as properties. Properties, in turn, are subdivided into attributes of an entity, relationships
56
57 between an existentially independent entity and the value of one of its properties, and
58
59
60

1
2
3 relationships, involving two or more individuals of the same, or of different existentially
4 independent entities (Orilia and Paoletti 2020).
5
6
7

8
9 Classifying concepts in vocabularies as entities and their properties, attributes or relationships
10 is a practice that has become common in the specification of vocabulary compliant with LOD
11 technologies; see, for example, the DC Terms vocabulary,
12 <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>, the PROV-O ontology,
13 <https://www.w3.org/TR/prov-o/>, and DCAT metadata vocabualry,
14 <https://www.w3.org/TR/vocab-dcat-3/>.
15
16
17
18
19
20
21
22

23
24 Data is about representations of something else. A data unit, a datum (Hjørland 2018), even
25 in the context of Big Data, then, makes no sense without referencing the entity and one of its
26 properties, the metadata. The three concepts are inseparable and cannot be understood
27 separately. They correspond to a descriptive, representational element of an entity, describing
28 one of its properties. They correspond linguistically to a claim, a basic unit of knowledge to
29 which, according to Aristotle (2000, 39), values of truth or falsity can be attributed.
30
31
32
33
34
35
36
37

38
39 The statements represented by triples constituted by an entity, one of its properties, and the
40 value of this property correspond to the representation of informational resources in the
41 context of LOD, using the RDF data model (RDF Primer 2014). RDF is a Semantic Web
42 standard for describing resources. Everything that is available on the Web can be accessed
43 through a link, or a Uniform Resource Identifier (URI). Today URI evolved towards IRI, the
44 Internationalised Resource Identifier, which strings incorporate characters from alphabets
45 others than the Latin alphabet. This representational model describes such a resource through
46 triples formed by a subject, the resource being described; a predicate, a property that describes
47 the resource; and an object, the value of this property for this resource. The RDF model
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 assumes a minimum semantics, that is, three elements with specific roles, the subject, the
4
5 predicate, and the object that form the triple and appear in this order.
6
7

8
9 Semiotic and ontological analysis identifies a piece of data as an artificial and intentional
10
11 artefact that represents something. The foundational types of the things that exist are entities
12
13 - existentially independent things - and their properties: relationships between two
14
15 existentially independent individuals, and attributes of an individual, its qualities and
16
17 quantities. Ontological Analysis of things in a domain, classifying and assigning types to these
18
19 things makes the terms in a domain vocabulary consistent, as they inherit the ontological
20
21 nature of their types and enable their representations to be machine-processable.
22
23
24
25

26 **3. A Comprehensive view of vocabularies**

27
28

29 In this section, a comprehensive view of vocabularies based on the previous discussion in
30
31 section 2 and on contributions by Hjørland (2018) and Zeng (2019) was compiled and
32
33 developed.
34
35

36 **3.1. Vocabularies, Web of Data, Linked Open Data, and Big Data**

37
38
39

40 LOD technologies are an integral part of the Web of Data project. Although this is its best-
41
42 known name, the project is also known as Web of Data, a name that describes it better, since
43
44 semantics concerns meanings (Chierchia, 2003), and the ability of the Web of Data to convey
45
46 meanings is quite limited and different from the sense in our understanding of expressions in
47
48 natural language.
49
50

51
52 The project was initially formulated by computer scientist Tim Berners-Lee, the creator of the
53
54 Web, among others. According to its formulators, the Semantic Web aims to propose “A new
55
56 form of Web content that is meaningful to computers will unleash a revolution of new
57
58 possibilities” (Berners-Lee et al 2001). To its authors, “Most of the Web’s content today is
59
60

1
2
3 designed for humans to read, not for computer programs to manipulate meaningfully.” The
4
5 Semantic Web then “will bring structure to the meaningful content of Web pages, creating an
6
7 environment where software agents roaming from page to page can readily carry out
8
9 sophisticated tasks for users.”
10

11
12
13 The Web of Data then refers to content represented in such a way that it can be understood by
14
15 both machines and people. The current Web is made up of pages, such as <http://www.uff.br>,
16
17 formatted in Hypertext Markup Language (HTML), accessible and interconnected with each
18
19 other through links. Navigating these pages through these links is done by browsers, such as
20
21 Internet Explorer, Google Chrome, or Mozilla Firefox. HTML is a content markup language;
22
23 it formats the content of a text of a page through a predefined set of markups, which instruct
24
25 browsers to display them on computer screens for human users. The content of HTML pages
26
27 is interpreted by browsers to make it readable and visually pleasing to people.
28
29
30

31
32 The proposed Web of Data is quite different. The Web will no longer be constituted of pages
33
34 to be read by people, but of content, called informational resources, digital representations of
35
36 things: concrete, like me, you, an industrial product, a monument, a geographical accident;
37
38 abstract, like a musical genre, a scientific discipline; or just has a digital existence, such as a
39
40 photo in a JPG file or a scientific article in a PDF file. These are the entities in the proposal
41
42 by Hjørland (2018). Each of these resources is uniquely identified by a link, or a URI. A
43
44 resource, identified/accessed by its URI, is described in a structured way through triples, each
45
46 one formed by the URI of the resource, by each of its properties, and by the corresponding
47
48 values of each of these properties. An example of how this representational model works is
49
50 the Leonardo Da Vinci resource on Wikidata, <https://www.wikidata.org/wiki/Q762>.
51
52
53
54
55

56 This model of structuring data through the description of resources formed by one or more
57
58 linguistic claims made up of triples <Subject> <Predicate> <Object> is RDF (RDF Primer,
59
60

1
2
3 2004). From an ontological point of view, subject, predicate, and object can be understood as
4 an entity, a property, and the value of this property.
5
6

7
8 Looking in more detail at structuring a triple; for example,
9

10
11 “The page <http://www.uff.br> is authored by _____.”
12
13

14
15 Such a claim consists of three elements: the subject, “<http://www.uff.br>,” the predicate, “has
16 as author” and the object, “_____”
17
18

19
20 The RDF model presupposes a minimum semantics, derived from its corresponding linguistic
21 claim. That is, they are identified and appear in this order: the subject, the predicate and the
22 object of the claim that form the triple (Resource Description Framework (RDF) Model and
23 Syntax Specification 1998). A triple describes a specific piece of data from the resource
24 description (what Hjørland calls a “datum:” a unit of data). Sets of triples with the same subject
25 describe the same resource. Sets of interlinked triples describing a resource form a graph.
26
27
28
29
30
31
32

33
34
35 SPARQL is the query language that allows users to query sets of RDF triples (SPARQL 1.1
36 QUERY LANGUAGE 2013), navigating through the graphs formed by them and performing
37 inferences. It is the materialization of the Web of Data proposal of a Web that can be queried
38 as if it were a database.
39
40
41
42
43
44

45
46 RDF can be serialized in several formats, such as RDF/XML, N Triples, JSON, or TURTLE
47 (RDF Primer, 2004). Of course, RDF triples coded in these formats are not as human-friendly
48 or as clearly readable as HTML pages when viewed by browsers. But they contain elements
49 that allow browsers to understand these formats and display them in a human-friendly manner,
50 if applicable. The main objective of the resources described in RDF is that they can be
51 processed by machines (including their user-friendly visualisation), thus helping to organise,
52 retrieve, and make these resources accessible.
53
54
55
56
57
58
59
60

1
2
3 The way to extend these semantics beyond the limits of the RDF model is also to make
4 predicates and/or objects into URI and that these URI refer to concepts of vocabularies with
5 specific semantics. According to RDF Semantics (2004) “There are several aspects of
6 meaning in RDF which are ignored by these semantics; in particular, it treats URI references
7 as simple names, ignoring aspects of meaning encoded in particular URI forms.” A URI in
8 the RDF model is just a name, an identifier. The advantage of a URI over a natural language
9 identifier such as the linguistic term “author”, is its uniqueness, its validity, since a URI is
10 valid and unique throughout the web space, and its persistence, that is, the commitment of
11 whoever assigns it. a URI to never change it (Berners-Lee 1998).
12
13
14
15
16
17
18
19
20
21
22
23
24

25 The previous example can be extended by using URI for the subject, the predicate, and the
26 object of the triple.
27
28
29

30 <http://www.uff.br> <http://purl.org/dc/elements/1.1/creator> <https://orcid.org/0000-0003-
31 0929-8475>
32
33
34
35

36 In this example, the original predicate “author” is replaced by the URI referenced by the
37 “creator” element of the well-known Dublin Core (DC) metadata standard. In its context,
38 dc:creator has specific semantics. It is defined as “An entity responsible for making the
39 resource.” The triple’s object, the value or content of dc:creator, has been replaced by the
40 Open Researcher and Contributor ID (ORCID), <https://orcid.org>, of the page’s author.
41
42
43
44
45
46
47
48

49 It is with the semantics in specific vocabularies that the limited semantic expressiveness of
50 the RDF model can be expanded. Once specified in elements of a vocabulary, the semantics
51 can be processed by programs. While the features provided in the Web of Data, represented
52 in markup languages such as XML, RDF, HTML, etc. are contents, programs are procedures.
53
54
55
56
57
58 Programs only know how to process content, they need to be clearly instructed (programmed)
59
60

1
2
3 on what to do with certain content in a certain situation. Specially formatted vocabularies, the
4
5 LOV (Mendez and Greenberg 2012) used to assign semantics to LOD (Zeng 2019) must
6
7 clearly define, restrict, and specify the semantics of their concepts. For example, the DC
8
9 metadata vocabulary clearly defines the semantics of each of its concepts (called elements in
10
11 the DC initiative); for example, `dc:creator`, is the creator/author or responsible for a resource,
12
13 e.g., a digital scientific paper. Furthermore, the `dc:creator` element has itself, a unique
14
15 persistent identifier, a link, a URI: <http://purl.org/dc/elements/1.1/creator>. This persistent
16
17 identifier, unique throughout the Web space, works as a guarantee of the metadata element
18
19 semantics, allowing a developer to create a specific program to process this element of the DC
20
21 vocabulary unambiguously, using the semantics specified and standardized in the DC
22
23 vocabulary to the `dc:creator` element.
24
25
26
27
28
29
30
31

32 3.2. Functionalities for vocabularies to be used within the context of the Web of Data and 33 34 LOD 35

36
37
38 Through unique and persistent identifiers, metadata and data vocabularies can be used to
39
40 assign machine-understandable semantics to predicates and objects in triples RDF. Many old
41
42 vocabularies are being restructured to be compatible with LOD technologies (Soergel, 2004;
43
44 Dos Santos Maculan, 2015), such as the UNESCO Thesaurus,
45
46 <http://vocabularies.unesco.org/browser/thesaurus/en/>, the FAO Thesaurus,
47
48 http://aims.fao.org/aos/agrovoc/c_8003.html, the AGROVOC Thesaurus,
49
50 <https://agrovoc.fao.org/browse/agrovoc/en/>, the Paul Getty Foundation Vocabularies,
51
52 <https://www.getty.edu/research/tools/vocabularies/lod/>, the Art and Architecture Thesaurus,
53
54 the Union List of Artists Names, the Cultural Objects Name Authority, the Getty Thesaurus
55
56 of Geographic Names, the DeCS/MeSH, Health Science Descriptors,
57
58
59
60

1
2
3 <https://decs.bvsalud.org/this/>, the Library of Congress Subject Headings (LCSH),
4
5 <https://id.loc.gov/authorities/subjects.html>, in addition to many others.
6
7
8
9

10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

1
2
3 discussion, - formal, i.e. computers' readable, and – shared, i.e., agreed by a community of
4
5 agents, being them humans or computers.
6
7

8 The language specification OWL – Ontology Web Language Overview (2004) states that:
9

10
11 OWL can be used to explicitly represent the meaning of terms in vocabularies
12
13 and the relationships between those terms. This representation of terms and
14
15 their interrelationships is called an ontology. OWL has more facilities for
16
17 expressing meaning and semantics than XML, RDF, and RDF-S, and thus
18
19 OWL goes beyond these languages in its ability to represent machine
20
21 interpretable content on the Web.
22
23
24

25 OWL is a standard language (meta-language in the aforementioned sense) of the W3C for
26
27 representing ontologies, that is, vocabularies that specify the things existing in a domain and
28
29 their interrelationships. Further on, the same specification compares the semantic
30
31 expressiveness of OWL with that of other languages to represent machine-interpretable
32
33 content such as XML, XML Schema, RDF, and RDFS (ONTOLOGY WEB LANGUAGE
34
35 OVERVIEW, 2004). It can thus be concluded that, with current technologies, a computational
36
37 ontology developed in OWL is the most expressive type of KOS, because the “facilities”
38
39 provided by OWL allow restricting, specifying, and expressing the intended meaning
40
41 (Guarino 1994, 560) of the conceptual model of a domain.
42
43
44
45

46 Each concept of an ontology vocabulary is typed; it is a class, or a property of a class or an
47
48 instance, an individual of a class. Among these facilities are the possibility of specifying data
49
50 properties (attributes, in Chen's ER model), object properties (relationships in Chen's ER
51
52 model), domain and scope of the two types of properties, and cardinality constraints of each
53
54 class involved in an object property, transitivity and reflexivity of properties, the disjunction
55
56 between individuals of different classes, axioms for restricting the inclusion of instances in a
57
58
59
60

1
2
3 class (ONTOLOGY WEB LANGUAGE OVERVIEW 2004), etc. These facilities can make
4
5 conceptual models implicit in a computational OWL ontology more faithful to reality.
6
7 Ontologies also do not distinguish thematic versus descriptive representation; every concept
8
9 is described by its properties, whether thematic or descriptive.
10
11

12
13 As seen earlier, the Web of Data project, the large-scale reuse of Big Data and research data
14
15 available in increasing amounts on the Web, depends on the one hand on the most expressive
16
17 vocabularies that describe them, and on the other hand, on programs capable of making
18
19 inferences, or at least algorithmic processing, on these representations. In this context, specific
20
21 domain models, intelligible by machines and represented with the maximum possible
22
23 semantic expressiveness such as computational ontologies gain importance.
24
25

26
27 Another important aspect related to this issue; Bergman (2011) discusses ODapps: The
28
29 Ontology-Driven Application Approach, an automatic program development methodology
30
31 based heavily on ontologies, a set of them, from high-level ontologies, task ontologies, domain
32
33 ontologies, to specific application ontologies (Guarino 1997, 145). In the context of ODapps,
34
35 domain computational ontologies, with a high degree of semantic expressiveness, are an
36
37 essential component for developing generic application programs, capable of processing,
38
39 making inferences, discovering, and reusing the knowledge contained in the domain
40
41 representation. It is therefore necessary to advance in the creation of domain-specific
42
43 computational ontologies domains that are increasingly semantically expressive to equip
44
45 programs capable of processing these representations to make inferences about them and
46
47 extract and reuse the knowledge contained therein.
48
49
50
51

52 53 **4. Results**

54
55 In the sequel the previous conceptualizations are applied to cases of research data and
56
57 discussed.
58
59
60

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

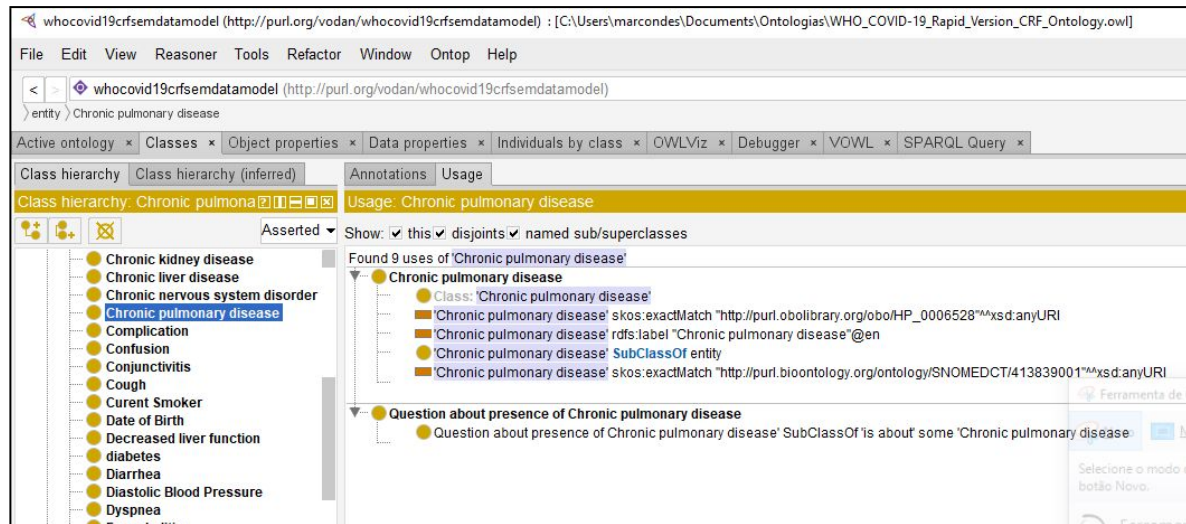
1
2
3 co-morbidity presented by the patient) are taken from specific biomedical ontologies or
4 vocabularies that describe specific co-morbidity types; if a specific one applies, it is recorded
5 as data as follows: Yes, No, Unk. These data have to be processed by programs so that the
6 immense amount of records collected through the CRF around the world can serve as inputs
7 for the planning and control of the pandemic. The question about co-morbidities has several
8 answer options, each of which indicates a type of disease. For it to be processed by machines,
9 each type of co-morbidity expressed in natural language must reference a concept in a
10 vocabulary or ontology, such as SNOMED-CT,
11 <https://www.nlm.nih.gov/healthit/snomedct/index.html>. Another question on the CRF, such
12 as the one related to “PRE-ADMISSION AND CHRONIC MEDICATION,” has as one of its
13 answer options “Angiotensin converting enzyme inhibitors (ACE inhibitors)?”, which may be
14 referenced in another vocabulary such as MeSH, <https://meshb.nlm.nih.gov/search>, the term
15 with identifier <http://id.nlm.nih.gov/mesh/D000806>.

16
17
18 In order to have precise meaning, concepts such as those shown in the CRF must refer to
19 specific, standardized ontologies or biomedical vocabularies to enable the processing of these
20 data.

21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 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.

1
2
3 The FAIR principles allow research data to be processed by machines. The M4M principle—
4 metadata for machines—states that “There is no FAIR data without machine-actionable
5 metadata. The overall goal of Metadata for Machines workshops (M4M) is to make routine
6 use of machine-actionable metadata in a broad range of fields.” The CRF described above is
7 an example of the importance of research data standardization and the adoption of principles
8 that allow its wide dissemination and reuse.
9

10 Applying the FAIR principles to research data causes data to be represented as RDF triples.
11 Such a process is named “FAIRification”, see [https://www.go-fair.org/fair-](https://www.go-fair.org/fair-principles/fairification-process/)
12 [principles/fairification-process/](https://www.go-fair.org/fair-principles/fairification-process/). FAIR compliant data is generally derived data from datasets.
13 A distributed network of FAIR Data Points provides access to different FAIR data. That raises
14 the question of using vocabularies to describe both the original datasets and their FAIR
15 compliant datasets versions generated.
16

17 Other vocabularies also have emerged, not to describe or provide standardized values for each
18 piece of data, but to provide descriptive and value metadata of the datasets as a whole. Digital
19 curation of research data is an emerging field of activity for KO professionals; one of its
20 activities is to apply metadata to research datasets, see <https://www.dcc.ac.uk/>. For the
21 curation of these datasets, metadata standards such as Data Catalog Vocabulary (DCAT)
22 <https://www.w3.org/TR/vocab-dcat-2/>, or the Provenance Ontology (PROV-O)
23 <https://www.w3.org/TR/prov-o/>, have been adopted to describe the provenance of the dataset.
24 As datasets have been made available as informational resources on the Web, information on
25 their provenance and the record of the processing carried out on them, the extract, transform,
26 load (ETL), see https://en.wikipedia.org/wiki/Extract,_transform,_load, and the
27 FAIRification processes of such data, are essential elements for research data reliability to
28 enable sharing and reuse.
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 The amount of research data being available every day on Coronavirus epidemic – the
4 “V”ariety” of Big Data - makes the integration of such sources essential to control the
5 epidemic. The Coronavirus Infectious Disease Ontology (CIDO) (He et al. 2020) stresses the
6 essential role computational ontologies in the integration of different and heterogeneous
7 research data sources, promoting interoperability between such sources.
8
9
10
11
12
13
14
15
16

17 These datasets, in addition to the metadata that describe their fields, are themselves of interest
18 for the research data exploration. They need additional metadata as the type of licence under
19 which data can be reused, the dataset creator, its publisher, its format, its update date, etc, all
20 of which are metadata for the dataset as a whole. They contain metadata such as the format of
21 the dataset, the number of records, the last update date, licences to use this dataset, etc. (from
22 DCAT), or metadata such as the agent that created the dataset, and the process that generated
23 it (from PROV-O). Standards such as these have been used in several research data
24 repositories to index the datasets deposited there. Indeed, digital curation is an increasingly
25 common application by KO professionals (Poole 2013).
26
27
28
29
30
31
32
33
34
35
36
37
38
39

40 Digital Humanities is another growing area of application of digital research data. It grew
41 from the wide availability of data from social activities (search and social media activity every
42 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
43 privilege knowledge source (Swanson, 2008), see PubMed Citations per year,
44 https://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html). Significant examples of
45 research projects in Digital Humanities using a variety of such sources can be found in the
46 Digging into Data Challenge program ([https:// diggingintodata.org/](https://diggingintodata.org/)) mentioned by Zeng
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 (2017); in this article, the author describes in details how Digital Humanities is related to Big
4
5 Data and the challenges to process such data and turn it into Smart Data.
6
7

8
9 A huge amount of such data is textual, resulting from posts on social media, emails, newspaper
10
11 articles, scientific articles, and text in encyclopaedias such as Wikipedia, among others. This
12
13 data is unstructured or semi-structured.
14
15

16
17 The exploitation of such potential information sources may lay on the development of
18
19 vocabularies for special purposes. Their processing using techniques such as information
20
21 extraction, named-entity recognition, natural language processing, text mining, machine
22
23 learning, text annotation, aim at transforming such non-structured or semistructured textual
24
25 data into structured.
26
27

28
29 Examples of such techniques in biomedical sciences are the National Library of Medicine
30
31 Natural Language Processing tools, <https://lhncbc.nlm.nih.gov/LHC-research/nlp.html>, which
32
33 lay on dictionaries and KOS like MeSH, the Medical Subject Headings, and UMLS, the
34
35 Unified Medical Language System (Bodenreider 2004), (Aronson and Lang 2010).
36
37
38

39 40 4.2. Semantics beyond the data. 41 42 43

44
45 Semantics is a very general concept. An operational concept of semantics applied to messages
46
47 – data: in digital environment is the inference made by an agent based on a message that
48
49 enables such agent to make decisions and, possibly, to act accordingly.
50
51

52
53 The concept of “powerful semantics,” originally devised by Shet, Ramakrishnan, and Thomas
54
55 (2005) and developed in Shet (2020, slide 42), is defined as “statistical analysis [that] allows
56
57 the exploration of relationships that are not stated.” Semantics may be obtained from statistical
58
59
60

1
2
3 patterns, not from individual datum referenced by metadata describing an entity, but rather
4
5 from data sets as a whole, or Big Data. To identify this semantics, Big Data, whether structured
6
7 or unstructured, has to be processed by programs. This is so-called data science (Dhar 2013).
8
9

10
11
12 Entities are the units to be represented by digital metadata and data within a domain, even if
13
14 an entity is represented by only one of its properties. As so they are the units of meaning and
15
16 correspond to what has been called a digital object. The concept of a digital object was first
17
18 proposed in 1995 by Kahn and Wilensky (2006) as a set of bits that has a special interest in
19
20 applications or software agents; it is related to the concept of data as a representation of an
21
22 entity or phenomenon (Hjørland 2018). Digital objects of interest to research data are also just
23
24 now (see <https://www.fdo2022.org/>) being conceptualized by initiatives such as FAIR Digital
25
26 Object Framework: “In the FDOF, a digital object is a bit sequence located in a digital memory
27
28 or storage that has, on its own, an informational value, i.e., the bit sequence represents an
29
30 informational unit such as a document, a dataset, a photo, a service, etc”, see
31
32 <https://fairdigitalobjectframework.org/>.
33
34
35
36
37
38
39

40 Within the Web of Data context vocabularies are meaning control and standardization
41
42 artefacts aimed at making knowledge records meaningful. The previous discussion poses the
43
44 question of levels of meaning related to levels of data aggregation. Table 1 sketch the
45
46 relationships between data aggregation levels to digital units of meaning.
47
48
49
50

51 DATA AGGREGATION LEVELS	52 DIGITAL UNITS OF MEANING
53 Level 1 - a datum (Hjørland 2018), the 54 basic element of data	55 the value of a database field, the content or an 56 excel cell

57
58
59
60

<p>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.</p>	<p>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</p>
<p>Level 3 - A data structure, a conceptualization, a message (CAPURRO, 2000) a row in a specific database table, a digital object, a named graph</p>	<p>a row in a specific database table, a digital object, a named graph A data structure, a conceptualization, a message (CAPURRO, 2000)</p>
<p>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</p>	<p>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).</p>
<p>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</p>	<p>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</p>

<p>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.</p>	<p>semantics that enable the integration and interoperability between such different and heterogeneous research data sources.</p>
--	---

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

1
2
3 play a central role in the semantic integration of different and heterogeneous research data
4
5 sources, promoting interoperability between such sources. In practical terms ontologies hold
6
7 representation of a domain while mapping properties (SKOS 2012), (ISO 25964-2 2013) and
8
9 also OWL property “sameAs” (Ontology Web Language Overview (2004) enable the
10
11 mapping of concepts in a data resource to concepts in another.
12
13
14
15
16
17

18 It is necessary also to distinguish one piece of datum as referred to by Hjørland (2018),
19
20 a unit that represents the value of one (of the) properties of an entity, from a record, a set of
21
22 several datum describing different properties of an entity, from datasets, representing the
23
24 various entities and their properties, and from databases, bringing together different datasets
25
26 representing different interrelated entities. Such are different data aggregation levels, having
27
28 higher levels of semantics in the computational environment. Vocabularies can play an
29
30 important role in addressing semantics to data at those different levels of aggregation.
31
32
33
34

35 Acknowledgments: This work was carried out with the support of the Brazilian agencies
36
37 CAPES - Financing Code 001, and CNPq, grant number 305253/2017-4. We are also grateful
38
39 to the anonymous reviewers of this work for their suggestions to improve this text.
40
41
42

43 **References**

44
45
46 Almeida, Mauricio; Souza, Renato and Fonseca, Fred. 2011. “Semantics in the Semantic Web:
47
48 A Critical Evaluation”. *Knowledge Organization* 38 no. 3: 187-203.
49
50 <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1041.7976&rep=rep1&type=pdf>,
51
52 accessed 25 Mar 2021.
53
54
55

56
57 Aristóteles. 1995. *Categorias*. Porto: Porto Editora Ltda.
58
59
60

1
2
3 Aronson, Alan R., and François-Michel Lang. 2010. "An overview of MetaMap: historical
4 perspective and recent advances." *Journal of the American Medical Informatics*
5
6 *Association* 17 no. 3: 229-236.
7
8

9
10
11 4- Bergman, Mike. 2011. "Ontology-Driven Apps Using Generic Applications". *AI3 blog*.
12
13 <https://www.mkbergman.com/948/ontology-driven-apps-using-generic-applications/>.
14
15

16
17
18 Berners-Lee, Tim. 1998. "Cool URIs don't change". <https://www.w3.org/Provider/Style/URI>.
19

20
21 Bodenreider, Olivier. 2004. "The unified medical language system (UMLS): integrating
22
23 biomedical terminology." *Nucleic acids research* 32 no. suppl_1: D267-D270.
24
25

26
27 Borst, Willem N. 1997. *Construction of Engineering ontologies*. Centre for Telematica and
28
29 Information Technology. University of Twente, Enschede, The Netherlands.
30
31

32
33 Capurro, R. 2000. "Angeletics—A message theory". In H.H. Diebner & L. Ramsay (Eds.),
34
35 *Hierarchies of communication*. Karlsruhe, Germany: ZKM.
36
37 http://www.capurro.de/angeletics_zkm.html.
38
39

40
41
42 CERIF in Brief. 2014. [https://eurocris.org/eurocris_archive/cerifsupport.org/cerif-in-](https://eurocris.org/eurocris_archive/cerifsupport.org/cerif-in-brief/index.html)
43
44 [brief/index.html](https://eurocris.org/eurocris_archive/cerifsupport.org/cerif-in-brief/index.html)
45
46

47
48
49 Chen, Peter Pin-Shan. 1976. "The Entity-Relationship Model-Toward a Unified View of
50
51 Data". *ACM Transactions on Database Systems* 1 no.1: 9-36.
52
53

54
55
56 Chierchia, Gennaro. 2003. *Semântica*. São Paulo, Ed. UNICAMP.
57
58
59
60

1
2
3 CIDOC Conceptual Reference Model Version 5.1.12. 2014. ICOM/CIDOC.

4
5 <http://www.cidoc-crm.org/Version/version-5.1.2>.

6
7
8
9 Dahlberg, Ingetraut. 1978. "A referent-oriented, analytical concept theory for
10 INTERCONCEPT". *Knowledge Organization* 5 no. 3: 142-151. [https://www.ergon-](https://www.ergon-verlag.de/isko_ko/downloads/ic_5_1978_3.pdf#page=20)
11
12
13
14
15 [verlag.de/isko_ko/downloads/ic_5_1978_3.pdf#page=20](https://www.ergon-verlag.de/isko_ko/downloads/ic_5_1978_3.pdf#page=20).

16
17 Dhar, Vasant. 2013. "Data science and prediction". *Communications of the ACM* 56 no. 12:
18
19
20
21 64-73. <https://dl.acm.org/doi/pdf/10.1145/2500499>.

22
23 Dextre Clarke, Stella G. 2019. "The Information Retrieval Thesaurus". *Knowledge*
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
Organization 46 no. 6: 439-459. [https://www.ergon-](https://www.ergon-verlag.de/isko_ko/downloads/ko_46_2019_6_c.pdf)
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

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>.

1
2
3 Floridi, Luciano. 2019. "Semantic Conceptions of Information". In *The Stanford*
4 *Encyclopedia of Philosophy* (Winter 2019 Edition), Edward N. Zalta (ed.).
5
6 <https://plato.stanford.edu/archives/win2019/entries/information-semantic/>.
7
8
9

10
11
12 Foskett, A. C. 1996. "*The subject approach to information*". Facet Publishing.
13
14

15
16
17 Fonseca, Claudenir M., Porello, Daniele, Guizzardi, Giancarlo, Almeida, João Paulo A. and
18
19 Guarino, Nicola. 2019. "Relations in Ontology-Driven Conceptual Modeling". In Laender,
20
21 A., Pernici, B., Lim, EP., de Oliveira, J. (eds) *Conceptual Modeling*. ER 2019. Lecture Notes
22
23 in Computer Science 11788. Springer, Cham. https://doi.org/10.1007/978-3-030-33223-5_4.
24
25

26
27 Fillinger, Sven et al. 2019. "Challenges of big data integration in the life sciences." *Analytical*
28
29 *and bioanalytical chemistry* 411 no. 26: 6791-6800. doi:10.1007/s00216-019-02074-9
30
31

32
33
34
35 Freitas, C.; Carvalho, P.; Oliveira, H. G.; Mota, C. and Santos, D. 2010. "Second HAREM:
36
37 advancing the state of the art of named entity recognition in Portuguese". In Nicoletta
38
39 Calzolari et al. (eds.), *Proceedings of the International Conference on Language Resources*
40
41 *and Evaluation (LREC 2010)*. European Language Resources Association, Valletta, 2010. pp.
42
43 3630-3637.
44
45

46
47 Frické, Martin. 2015. "Big Data and Its Epistemology". *Journal of the Association for*
48
49 *Information Science and Technology* 66 no. 4: 651-61.
50
51

52
53
54 Gandomi, Amir, and Murtaza Haider. 2015. "Beyond the hype: Big data concepts, methods,
55
56 and analytics." *International journal of information management* 35 no.2: 137-144.
57
58
59
60

1
2
3 Gershenfeld, Nel, Krikorian, Raffi, and Cohen, Danny. 2004. "The Internet of Things".
4
5 *Scientific American*, October: 76-81. <http://cba.mit.edu/docs/papers/04.10.i0.pdf>.
6
7

8
9 Giunchiglia, Fausto; Dutta, Biswanath and Maltese, Vincenzo. 2014. "From knowledge
10
11 organization to knowledge representation". *Knowledge Organization* 41 no. 1: 44-56.
12
13 <http://eprints.biblio.unitn.it/4186/1/techRep027.pdf>.
14
15

16
17 Gray, Jim. 2009. "eScience: A Transformed Scientific Method". In *The Fourth Paradigm,*
18
19 *Data-intensive Scientific Discovery*, ed. Tony Hey, Stewart Tansley and Kristin Tolle.
20
21 Redmond, Washington, Microsoft Research, 19-33.
22
23 <http://itre.cis.upenn.edu/myl/JimGrayOnE-Science.pdf>.
24
25

26
27 Guarino, Nicola. 1997. "Semantic matching: Formal ontological distinctions for information
28
29 organization, extraction, and integration". In *International Summer School on Information*
30
31 *Extraction*. Springer, Berlin, Heidelberg, 1997. 139-170.
32
33 [https://kask.eti.pg.gda.pl/redmine/projects/sova/repository/revisions/5378040326bc499e118](https://kask.eti.pg.gda.pl/redmine/projects/sova/repository/revisions/5378040326bc499e118636a1d25ad667285e005c/entry/Praca_dyplomowa/materialy/10.1.1.53.939.pdf)
34
35 [636a1d25ad667285e005c/entry/Praca_dyplomowa/materialy/10.1.1.53.939.pdf](https://kask.eti.pg.gda.pl/redmine/projects/sova/repository/revisions/5378040326bc499e118636a1d25ad667285e005c/entry/Praca_dyplomowa/materialy/10.1.1.53.939.pdf).
36
37
38

39
40 Guarino, Nicola; Carrara, Massimiliano an Giaretta, Pierdaniele. 1994. "Formalizing
41
42 ontological commitment". In *AAAI*. 1994 : 560-567.
43
44 <https://www.aaai.org/Papers/AAAI/1994/AAAI94-085.pdf>.
45
46

47
48 Gruber, Thomas R. 1993. "A translation approach to portable ontology
49
50 specifications." *Knowledge acquisition* 5 no. 2: 199-220.
51
52

53
54 Hajibayova, Lala, and Athena Salaba. 2018. "Critical questions for big data approach in
55
56 knowledge representation and organization." *Challenges and Opportunities for Knowledge*
57
58 *Organization in the Digital Age: Proceedings of the Fifteenth International ISKO Conference*
59
60 *9-11 July 2018 Porto, Portugal*, Vol. 16. Ergon Verlag.

1
2
3 He, Yongqun, et al. 2020. "CIDO, a community-based ontology for coronavirus disease
4 knowledge and data integration, sharing, and analysis." *Scientific data* 7 no. 1: 1-5.
5
6
7

8 Hey, Tony; Trefethen, Anne. 2003. "The data deluge: An e-science perspective". In *Grid*
9 *computing: Making the global infrastructure a reality*, p. 809-824.
10
11 https://eprints.soton.ac.uk/257648/1/The_Data_Deluge.pdf.
12
13
14
15
16
17

18 Hjørland, Birger. 2018. "Data (with big data and database semantics)". *Knowledge*
19 *Organization* 45 no. 8: 685-708.
20
21
22

23 Hjørland, Birger. 2002. "Domain analysis in information science: eleven approaches—
24 traditional as well as innovative". *Journal of Documentation*, 58 no. 4: 422-462.
25
26
27

28 Hjørland, Birger. 2013. "Theories of knowledge organization — theories of knowledge.",
29 *Knowledge Organization* 40: 169–181.
30
31
32
33

34 Hjørland, Birger, and Albrechtsen, Hanne. 1995. "Toward a new horizon in information
35 science: Domain-analysis". *Journal of the American Society for Information Science* 46 no.
36 6: 400-425.
37
38
39
40
41
42

43 Hjørland, Birger and Hartel, Jenna. 2003. "Introduction to a special issue of Knowledge
44 Organization". *Knowledge Organization* 30 no. 3/4: 125-7.
45
46
47

48 Iafate, Fernando. 2015. *From Big Data to Smart Data*. London: ISTE Ltd., and Hoboken,
49 NJ: John Wiley & Sons, Inc.
50
51
52
53

54 Ibekwe-SanJuan, Fidelia and Geoffrey C. Bowker. 2017. "Implications of Big Data for
55 Knowledge Organization". *Knowledge Organization* 44, no. 3: 187-98.
56
57
58
59
60

1
2
3 International Council on Archives. Experts Group on Archival Description. 2019. Records in
4 Context: A Conceptual Model for Archival Description (Consultation Draft v0.1). ICA.
5
6 https://www.ica.org/sites/default/files/ric-cm-0.2_preview.pdf.
7
8
9

10
11 International Federation of Library Associations and Institutions (IFLA). 1998. *Study Group*
12 *on Functional Requirements for Bibliographic Records: Final Report*. UBCIM Publications
13
14 New Series. München: K. G. Saur.
15
16
17
18
19

20
21 ISO 25964-2. 2013. *Information and documentation — Thesauri and interoperability with*
22 *other vocabularies — Part 2: Interoperability with other vocabularies*. ISO, 2013.
23
24
25

26
27 ISO/IEC 20546:2019(en). 2019. *Information technology — Big data — Overview and*
28 *vocabulary*. ISO.
29
30
31

32
33 Kahn, Robert; Wilensky, Robert. 2006. “A Framework for Distributed Digital Objects
34 Services”. *International Journal on Digital Libraries* 6 no. 2: 115–123.
35
36 https://www.doi.org/topics/2006_05_02_Kahn_Framework.pdf.
37
38
39

40
41
42 Lambe, Patrick. 2014. *Organising knowledge: taxonomies, knowledge and organizational*
43 *effectiveness*. Elsevier.
44
45
46

47
48 Leonelli, Sabina. 2012. “Classificatory Theory in Data-intensive Science: The Case of Open
49 Biomedical Ontologies”. *International Studies in the Philosophy of Science* 26 no. 1: 47–65.
50
51
52

53 _____ and Dias, Celia. 2020. “Representing facet classification in SKOS”. In
54 International ISKO Conference, Aalborg, Denmark, 16th, *Proceedings...*
55
56
57
58 *1. Edition*. Würzburg: Ergon Verlag. *ISBN print: 978-3-95650-775-5, ISBN online: 978-3-*
59
60

1
2
3 95650-776-2, *Series: Advances in knowledge organization* 9. Würzburg: Ergon
4 Verlag, 254–263. <https://doi.org/10.5771/9783956507762>.
5
6
7
8
9

10 De Mauro, Andrea; Greco, Marco and Grimaldi, Michele. 2015. “What is big data? A
11 consensual definition and a review of key research topics”. In *AIP conference proceedings*.
12 American Institute of Physics, 2015. p. 97-104. [http://big-data-fr.com/wp-](http://big-data-fr.com/wp-content/uploads/2015/02/aip-scitation-what-is-bigdata.pdf)
13 content/uploads/2015/02/aip-scitation-what-is-bigdata.pdf.
14
15
16
17
18
19

20
21 Mazzocchi, Fulvio. 2018. “Knowledge organization system (KOS)”. *Knowledge*
22 *Organization* 45, no.1: 54-78. Also available in ISKO Encyclopaedia of Knowledge
23 Organization, eds. Birger Hjørland and Claudio Gnoli. <http://www.isko.org/cyclo/kos>.
24
25
26
27

28 Méndez, Eva; Greenberg, Jane. 2012. “Linked data for open vocabularies and HIVE’s global
29 framework”. *El profesional de la información* 21 no. 3: 236-244.
30
31
32
33

34 Mylopoulos, John. 1992. “Conceptual modelling and Telos”. In *Conceptual modelling,*
35 *databases, and CASE: An integrated view of information system development*, p. 49-68.
36
37
38
39 <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.3647&rep=rep1&type=pdf>.
40
41

42 Ontology Web Language Overview. 2004. W3C. <https://www.w3.org/TR/owl-features/>.
43
44

45 Orilia, Francesco and Paoletti, Michele Paolini. 2020. "Properties", *The Stanford*
46 *Encyclopedia of Philosophy* (Winter 2020 Edition), Edward N. Zalta (ed.).
47
48
49 <https://plato.stanford.edu/archives/win2020/entries/properties/>.
50
51

52 Otlet, Paul. 2018. *Tratado de Documentação: o livro sobre o livro, teoria e prática*. Brasília:
53 Briquet de Lemos Livros.
54
55
56
57
58
59
60

1
2
3 Peirce, Charles. S. 1869. "On a new list of categories". *Proceedings of the American Academy*
4 *of Arts and Sciences*, v. 7, p. 287-298, 1868. [http://www.bocc.ubi.pt/pag/peirce--charles-list-](http://www.bocc.ubi.pt/pag/peirce--charles-list-categories.pdf)
5 [categories.pdf](http://www.bocc.ubi.pt/pag/peirce--charles-list-categories.pdf).
6
7
8
9

10
11 Poole, Alex H. "Now is the Future Now? 2013. "The Urgency of Digital Curation in the
12 Digital Humanities." *DHQ: Digital Humanities Quarterly* 7 no. 2.
13
14

15
16 Prasad, A. R. D., Giunchiglia, Fausto; Devika, P. Madalli. 21017. "DERA: from document
17 centric to entity centric knowledge modelling". In: Proceedings of the International UDC
18 seminar 2017. Faceted classification today. London: September, 2017. p. 169-179.
19
20
21 <http://seminar.udcc.org>.
22
23
24
25

26
27 Prieto-Díaz, Ruben. 1990. "Domain analysis: An introduction". *ACM SIGSOFT Software*
28 *Engineering Notes*, 15 no. 2: 47-54.
29
30

31
32 Ranganathan, S. R. and Gopinath, M. A. 1967. *Prolegomena to Library Classification*. 3 ed.
33
34 Bombay: Asia Publishing House.
35
36

37
38 RDF semantics. W3C, 2004. <http://www.w3.org/TR/rdf-mt/>.
39
40

41
42 RDF 1.1. PRIMER. 2014. W3C. <https://www.w3.org/TR/rdf11-primer/>. 2019.
43
44

45 Resource Description Framework (RDF) Model and Syntax Specification. 1998. W3C.
46 <https://www.w3.org/1998/10/WD-rdf-syntax-19981008/>.
47
48
49

50 Riva, Pat, Le Boeuf, Patrick, and Žumer, Maja. 2017 "*IFLA Library Reference Model: A*
51 *Conceptual Model for Bibliographic Information*". IFLA. [online]
52
53 <https://www.ifla.org/publications/node/11412>.
54
55
56
57
58
59
60

1
2
3 Rowley, Jennifer. 2007. "The wisdom hierarchy: representations of the DIKW hierarchy".
4
5 *Journal of information science* 33 no. 2: 163-180.
6
7 <http://web.dfc.unibo.it/buzzetti/IUcorso2007-08/mdidattici/rowleydikw.pdf>.
8
9

10
11 Saracevic, Tefko. 2007. "Relevance: A review of the literature and a framework for thinking
12
13 on the notion in information science. Part II: Nature and manifestations of relevance". *Journal*
14
15 *of the american society for information science and technology* 58 no. 13: 1915-1933.
16
17

18
19 Shet, Amith. 2020. "Knowledge Graphs and their central role in big data processing: Past,
20
21 Present, and Future". In 7th ACM India Joint Conference on Data Science & management of
22
23 Data (COD-COMAD), Indian School of Business, Hyderabad Campus, 5-7 January 2020.
24
25 [https://www.slideshare.net/apsheth/knowledge-graphs-and-their-central-role-in-big-data-](https://www.slideshare.net/apsheth/knowledge-graphs-and-their-central-role-in-big-data-processing-past-present-and-future)
26
27 [processing-past-present-and-future](https://www.slideshare.net/apsheth/knowledge-graphs-and-their-central-role-in-big-data-processing-past-present-and-future), accessed Jun. 5, 2021.
28
29

30
31
32 Shet, Amith; Ramakrishnan, Cartic and Thomas, Christopher. 2005. "Semantics for the
33
34 semantic web: The implicit, the formal and the powerful". *International Journal on Semantic*
35
36 *Web and Information Systems (IJSWIS)*, no. 1 vol. 1:1-18.
37
38 <http://www.ebusinessforum.gr/old/content/downloads/IJSWIS.pdf#page=19>.
39
40
41

42
43
44 Shiri, Ali. 2013. "Linked data meets big data: A knowledge organization systems
45
46 perspective." *Advances in Classification Research Online* 24 no. 1: 16-20.
47
48

49
50
51 SKOS – Simple Knowledge Organization System Namespace Document. 2012. W3C.
52
53 <https://www.w3.org/2009/08/skos-reference/skos.html#>.
54
55
56
57
58
59
60

1
2
3 Soergel, Dagobert. 2015. "Unleashing the Power of Data Through Organization: Structure
4 and Connections for Meaning, Learning and Discovery." *Knowledge Organization* 42 no. 6:
5
6 401-427.
7
8

9
10
11
12 SPARQL 1.1 QUERY LANGUAGE, 2013. W3C. <https://www.w3.org/TR/sparql11-query/>.
13

14
15 Strecker, Dorothea et al. 2021. *Metadata Schema for the Description of Research Data*
16 *Repositories*. Re3data, 2021. Available at: <https://doi.org/10.48440/re3.010>. Access 08 Jul.
17
18 2022.
19

20
21
22
23 Swanson, Don R. 2008. "Literature-based discovery? The very idea." In *Literature-based*
24 *discovery*. Springer, Berlin, Heidelberg. 3-11.
25
26

27
28
29
30
31
32 Veiga, Viviane Santos de Oliveira; Campos, Maria Luiza; Silva, Carlos Roberto Lyra;
33
34 Henning, Patricia and Moreira, João. 2021. "Vodan br: a gestão de dados no enfrentamento
35 da pandemia coronavirus". *Páginas A&B, Arquivos e Bibliotecas (Portugal)*, n. Especial: 51-
36
37 58. <http://hdl.handle.net/20.500.11959/brapci/157353>.
38
39

40
41
42
43
44 Zeng, Marcia Lei. 2019. "Interoperability". *Knowledge Organization* 46, no. 2: 122-146. Also
45 available in Hjørland, Birger and Gnoli, Claudio eds. *ISKO Encyclopedia of Knowledge*
46
47 *Organization*, <http://www.isko.org/cyclo/interoperability>.
48
49

50
51
52 Zeng, Marcia. L. 2017. "Smart data for digital humanities". *Journal of Data and*
53
54 *Information Science* 2 no. 1: 1-12. DOI: 10.1515/jdis-2017-0001.
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Review Only

1
2
3 KO-2022-0003 - revised-v3 - Answers to the Reviewers comments
4
5

6 October 1, 2022
7

8 Dear Reviewers
9

10 Thank you for your valuable comments to our text.
11

12 The text was edited removing the revised v2 version markup in red and yellow made to
13 attend to the reviewer comments. Text added to this version was highlighted blue.
14

15 Explanation of value vocabularies and metadata vocabularies were expanded according to
16 the reviewer's suggestion in section 4.1.
17

18 The reference list was also checked and revised.
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Review Only