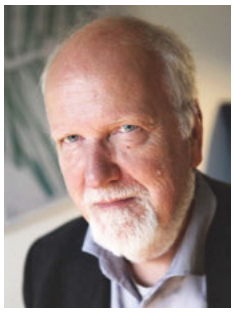
Data (with Big Data and Database Semantics)†

Birger Hjørland

University of Copenhagen, Department of Information Studies,
<Birger.Hjorland@hum.ku.dk>



Birger Hjørland holds an MA in psychology and PhD in library and information science. He is Professor in knowledge organization at the Department of Information Studies, University of Copenhagen (formerly Royal School of Library and Information Science) since 2001 and at the University College in Borås 2000-2001. He is chair of ISKO Scientific Advisory Council and a member of the editorial boards of *Knowledge Organization*, *Journal of the Association for Information Science and Technology* and *Journal of Documentation*. His h-index on 2018-12-18 is 46 in Goolge Scholar and 27 in Web of Science.

Hjørland, Birger. 2018. "Data (with Big Data and Database Semantics)." *Knowledge Organization* 45(8): 685-708. DOI:10.5771/0943-7444-2018-8-685.

Abstract: It is argued that data should be defined as information on properties of units of analysis. Epistemologically it is important to establish that what is considered data by somebody need not be data for somebody else. This article considers the nature of data and "big data" and the relation between data, information, knowledge and documents. It is common for all these concepts that they are about phenomena produced in specific contexts for specific purposes and may be represented in documents, including as representations in databases. In that process, they are taken out of their original contexts and put into new ones and thereby data loses some or all their meaning due to the principle of semantic holism. Some of this lost meaning should be reestablished in the databases and the representations of data/documents cannot be understood as a neutral activity, but as an activity supporting the overall goal implicit in establishing the database. To utilize (big) data (as it is the case with utilizing information, knowledge and documents) demands first of all the identification of the potentials of these data for relevant purposes. The most fruitful theoretical frame for knowledge organization and data science is the social epistemology suggested by Shera (1951). One important aspect about big data is that they are often unintentional traces we leave during all kinds of activities. Their potential to inform somebody about something is therefore less direct compared to data that have been produced intentionally as, for example, scientific databases.

Received:11 October 2018; Accepted: 11 November 2018

Keywords: data, information, big data, knowledge, science

† The drive to write this paper came from an invitation to give a keynote speech at the 10th International Meeting on Information, Knowledge and Action, June 22, 2018 in UNESP Marília in Brazil on the theme *Big Data: Continuity or Revolution*? Thanks to Daniel Martínez-Ávila for serving as the editor of this article as well as to three anonymous peer reviewers, who contributed substantially to its improvement. Thanks also to Lars Konzack, Volkmar Engerer and Niels Ole Finnemann from the Knowledge Organization Research Group at the Department of Information Studies for a valuable discussion of a late version of the manuscript.

1.0 Introduction

This paper seeks to define "data" and to relate this concept to other core concepts in library and information science (LIS), including facts, information and documents. It also seeks to uncover the most important theoretical principles for dealing with this phenomenon. It is a basic assumption throughout the article that these problems are of an epistemological nature and that a problematic empiricism has dominated so far, and that a proper (social) epistemological understanding is of immense importance to get established. Social epistemology and its connection to related views are briefly introduced in Section 5.0 but has been influencing the way data is understood throughout the article.

"Data" is a much-used concept in many fields, including LIS, in particular in composite terms such as "database," "data archive," "data mining," "descriptive data," "metadata," "linked data" and now "big data." These terms are common terms of the field and need proper theoretical and terminological attention. The word "data" is used as either a mass noun or plural. We will here use datum about the singular term and data about the plural. As shown in endnote 1 the term "data" is, however, also often used in the singular.[1]

Today, "data" has become a much more used concept than ever before in the history of library and information science. Formerly, concepts such as books, (graphic) records[2], documents, and information were the terms used about the core object of the field. In 1955, when the term information science was introduced, there was much talk about the information explosion[3] and how to deal with it (it seems almost a main reason for establishing information science as a field). Today, there is much talk about the data deluge, and the necessity to deal with it in various

fields including computer science and information science[4], just as there are tendencies to establish a new field, "data science."[5] The choice of concepts considered the basis for the field is in itself an important issue and should be driven by theoretical arguments rather than by importing hype words.

2.0 Defining data

2.1 Epistemological issues

Epistemological issues involve, first of all, the question about objectivity and subjectivity of data. The term 'data' is used today in ways that is heavily loaded with epistemological problems. Rowley (2007, 107-1)[6], for example, provides examples of the view, that data are discrete, objective facts or observations; however, in this article an opposite view will be put forward. Jensen (1950, ix) wrote:

> It is an unfortunate accident of history that the term *datum* (Latin, past participle of *dare*, 'to give') rather than *captum* (Latin, past participle of *capere*, 'to take') should have come to symbolize the unit-phenomenon in science. For science deals, not with 'that which has been given' by nature to the scientist, but with 'that which has been taken' or selected from nature by the scientist in accordance with his purpose, arid to a degree even constructed out of nature by the scientist's preliminary assumptions as to which of "the things which have been given" are also to be "taken" or observed.

Drucker (2011) also found that "the concept of data as a given has to be rethought through a humanistic lens and characterized as capta, taken and constructed," and she thereby introduces an epistemological understanding related to the one mentioned by Jensen and further developed in the present article. In addition to the importance of understanding data as capta, it should be considered that Jensen understood data as being related to unit-phenomena. We shall return to this in the end of this Section 2.5 in the suggested definition of data.

In line with the understanding of data, not as "given" but as "taken" or constructed, Fritz Machlup (1984, 646-7) provided an important analysis:

> The use and misuse of the term data is due, in part, to linguistic ignorance. Many users do not know that this is a Latin word: dare means "to give"; datum, "the given" (singular); and data, "the givens" (plural). Data are the things given to the analyst, investigator, or problem-solver; they may be numbers, words, sentences, records, assumptions; just anything given, no matter in what form and of what origin. This used to be well known to scholars in most fields: Some wanted the word data to refer to facts, especially to instrument-readings; others to assumptions. Scholars with a hypothetico-deductive bent wanted data to mean the given set of assumptions; those with an empirical bent wanted data to mean the records, or protocol statements, representing the findings of observation, qualitative or quantitative ....
> One can probably find quotations supporting all possible combinations of the three terms [data, information, knowledge] or of the concepts they are supposed to denote. Each is said to be a specific type of each of the others, or an input for producing each of the others, or an output of processing each of the others.
> Now, data from the point of view of the programmers, operators, and users of the computer, need not be data in any other sense.

Machlup thus point out that data are relative to different perspectives and domains (e.g., computer programmers versus computer users) and related to different epistemological perspectives (e.g., empiricism: instrument readings or protocol statements; rationalism: assumptions). This relativity was also emphasized by Borgman (2010, 3): "The notion of "data" can vary considerably among collaborators ... and even more so between disciplines." The view that data differs among epistemological positions is here of special importance.[8] Rafael Capurro further stresses the philosophical perspective (in Zins 2007, 481):

> Data are (or datum is) an abstraction. I mean, the concept of "data" or "datum" suggests that there is something there that is purely given and that can be known as such. The last one hundred years of (late) philosophic discussion and, of course, many hundred years before, have shown that there is nothing like "the given" or "naked facts" but that every (human) experience/knowledge is biased. This is the "theory-laden" theorem that is shared today by such different philosophic schools as Popper's critical rationalism (and his followers and critics such as Kuhn or Feyerabend), analytic philosophy (Quine, for instance), hermeneutics (Gadamer), etc. Modern philosophy (Kant) is very acquainted with this question: experience ("Erfahrung") is a product of "sensory data" within the framework of perception ("Anschauung") and the categories of reason ("Verstand") ("perception without concepts is blind, concepts without perception are void"). Pure sensory data are as unknowable as "things in themselves."

2.2 Data, facts, information and documents

WordNet 3.1 defines data and datum:

- data, information (a collection of facts from which conclusions may be drawn) "statistical data"
- datum, data point (an item of factual information derived from measurement or research)

We see that WordNet considers "data" synonym with both "information" and "facts," a widely held view, which, however, will be further criticized below. Before we proceed considering definitions in the scholarly literature, an example will be given demonstrating the present author's understanding of the relations between data, facts, documents and information.

Consider genealogists, who are constructing family trees by putting up family relations between individuals. Important data are here, for example, date of birth of a given person (X), sex of X, father of X and mother of X. First drafts of such a tree may be based on interviews and recollections from close relatives. Because such information is often unreliable, improved documentation, such as birth certificates are looked for. A birth certificate is a document that contains data about name, birth dates and parents of an individual. Normally, such a documentation is considered proof of a fact, e.g., that X is child of Y. A birth certificate provides information[9] about family relations, i.e., it answers real or hypothetical questions, about, among other things, family relations. Birth certificates are made reliable by the institutions and procedures that have the responsibility to issue the certificates, and these institutions and procedures are established by the authorities based on a perceived balance between expenditure, needed security and conservatism of administrative institutions (depending on the jurisdiction, a record of birth might or might not, for example, contain verification of the event by a midwife or doctor). Today genealogists have got a new kind of document: Reports from DNA analysis. They are highly reliable, for example, one report said "The probability of Mr. Jim Doe being the biological father of John Test is > 99.9999 %."[10] However, nothing can be considered facts in an absolute sense (due to the view of fallibilism, further described below). Birth certificates and DNA-reports are documents reporting data about relationships between individuals, but as they are sometimes conflicting neither of them can be understood as reporting pure facts. They may, however, be understood as claims that are normally accepted as facts by a given community. Here, as in science, great efforts may be made to secure the truth of the claims, but established procedures and findings are normally not questioned until there seem good reasons for doing so, or if very much is at stake. In other words, documented data are considered as being facts for the tasks they are produced to serve, they represent sufficient facts in a given social context. If they were not, it would be impossible to act on the basis of data; for example, it would be impossible to construct family trees, and there would be no reason to issue such documents (e.g., birth certificates). To summarize: The datum "X is child of Y" may, for example, be obtained from tree kinds of documents:

(1) an interview with Y;
(2) the birth certificate of X;
(3) and DNA analysis based on DNA from X and Y.

Normally all tree documents will be considered reporting this datum as a fact. The DNA report is the most reliable source today, but all three contain the datum.

We have now presented an example in which data, facts, documents and information are clearly distinguished as different concepts and will from this basis continue considering conceptual discussions in the scholarly literature.

In information science, Henning Spang-Hanssen (2001) related the data concept to the issue of fact retrieval versus document retrieval:

> Information about some physical property of a material is actually incomplete without information about the precision of the data and about the conditions under which these data were obtained. Moreover, various investigations of a property have often led to different results that cannot be compared and evaluated apart from information about their background. An empirical fact has always a history and a perhaps not too certain future. This history and future can be known only through information from particular documents, i.e., by document retrieval. The so-called fact retrieval centers seem to be just information centers that keep their information sources—i.e. their documents—exclusively to themselves.

Spang-Hanssen here provided an argument against a view that has been strongly represented in information science: That what should be retrieved or found is not documents, but "information," a view implicit in the term "information retrieval" (as opposed to "document retrieval").[11] In connection to the understanding of the concept "data," Spang-Hanssen addressed its fallible nature and the necessity to consider information about how data has been obtained as well as about their contexts in general. The quote by Spang-Hanssen provides therefore a strong argument for preferring the term "document retrieval" and to consider documents, rather than information the core object of the field. In the context of the present article the pri-

mary function of the quote is, however, to introduce fallibilism in relation to data. Fallibilism is a term coined by the founder of pragmatic philosophy, Charles Sanders Peirce (Peirce 1955, 59), for the view that "people cannot attain absolute certainty concerning questions of fact." From the point of hermeneutics, Caputo (2018, 4) reached the same conclusion:

> In hermeneutics, we defend the idea that there are no pure facts. Behind every interpretation lies another interpretation. We never reach an understanding of anything that is not an interpretation. We can never peel away the layers to get to some pure, uninterpreted, naked fact of the matter. No matter how loudly you proclaim you are just sticking to the facts, you are only raising the volume of your own interpretation. In hermeneutics, I like to say, interpretation goes all the way down.

Data has often been considered the bottom or basis of the so-called Data-Information-Knowledge-Wisdom hierarchy (DIKW), where information is understood as processed or interpreted data, knowledge as processed or interpreted information and wisdom as interpreted knowledge. This model has been discussed and criticized by, among others, Frické (2018), who found that this model is based on the problematic epistemology of inductivism. The main problem in the DIKW-model is that data cannot be seen as the given objective facts, on which interpretation is made to produce the higher levels. Data themselves are also based on interpretations as it was argued in the quotes above.

2.3 Data and documents confused

Borgman (2007, 41-2 and 119-21 and 2015, 17-29) presents and discusses definitions of data and refers in this regard to Meadows (2001)[12], Hey and Trefethen (2003), Consultative Committee for Space Data Systems (2002, Section 1:9), National Research Council (1999)[13], Hodge and Frangakis (2004) and National Science Board (2005). The definitions from some of these sources and Borgman's conclusions will be discussed below.

One of the definitions of data is from Consultative Committee for Space Data Systems (2002, Section 1:9):

> Data: A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen.[14]

Borgman (2007, 119) wrote that this definition of data is widely accepted in this context. Concerning the distinction between data and documents, she wrote (Borgman 2007, 41-2):

> Discussions of digital scholarship tend to distinguish (implicitly or explicitly) between data and documents. Some view data and documents as a continuum rather than as a dichotomy (Hey and Trefethen 2003). In this sense, data such as numbers, images, and observations are the initial products of research, and publications are the final products that set research findings in context.

Besides considering data and documents as being related to different stages of the research process, the two quotes does not provide criteria to distinguish the two concepts. Borgman does not consider why the definitions of data given above do not *also* qualify as definitions of documents. (As we shall see below she also consider the situation about defining data unsatisfactory). In particular, many of the specific examples of data provided (e.g., recording of sounds or a moon rock specimen) are explicitly considered documents in the literature. Suzanne Briet (1951, 7; 2006, 9-10) defined: "A document is evidence in support of a fact and could be any physical or symbolic sign, preserved or recorded, intended to represent, to reconstruct, or to demonstrate a physical or conceptual phenomenon" (see also Buckland (2018)).

The same confusion is repeated by Bugaje and Chowdhury (2018, 253), who wrote (quoting Borgman 2015): "Among the more comprehensive definitions of research data is that they are "entities used as evidence of phenomena for the purposes of research or scholarship," which may range in form from digital records (e.g. text, audio, video, spreadsheets, etc.) to physical objects (e.g. laboratory specimens, historical artefacts, soil samples, etc.)" This definition of data matches the definition of document as entities used for documenting something and it therefore also fails to distinguish data from documents. The distinction between data and documents was also illustrated in the example about genealogy above and will be made clearer below when a suggested definition by Kaase is considered.

The confusion of distinguishing the two concepts increases when Borgman (2007, 120; emphasis original) with reference to Hodge and Frangakis 2005 and National Science Board 2005 writes: "*Records* of government, business, and public and private life also yield useful data for scientific, social scientific, and humanistic research." Of course, they *yield* data, but that does not *make* them data, but rather kinds of documents. Borgman herself called for more a precise definition of data (2010, 2-3):

> All too rarely do those promoting the sharing and curation of data define 'data' explicitly or acknowledge the diversity of forms that data may take. The definition established in a *National Research Council* report suggests the complexity of the concept: "Data are facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors." [Serafin et al. 1999, 15]. The notion of "data" can vary considerably among collaborators [Wallis et al. 2008], and even more so between disciplines.

This is, however, just a list of examples of what falls under the concept "data" and not a proper definition. Therefore, Borgman (2007 and 2010) and Bugaje and Chowdhury have not clearly analyzed if "data" and "document" should be considered synonymous concepts – and if they should not, then identify what distinguishes them. Leonelli (2016), which includes a deep discussion of data from the perspective of the philosophy of science, also seems to blur documents and data.[15] Borgman (2015, 28-9) acknowledges that the situation is not satisfactory:

> Lists of entities that could be considered data are unsatisfactory as definitions, yet such "definitions" abound in the scholarly literature and in policy documents. The inability to anchor the concept in ways that clarify what are and are not data in a given situation contributes mightily to the confusion about matters such as data management plans, open data policies, and data curation. Concrete, bounded definitions are most often found in operational contexts.

Borgman then concludes with the former quoted definition from Bugaje and Chowdhury (2018, 253), which – as it has been shown –- confuses data and documents: "entities used as evidence of phenomena for the purposes of research or scholarship."[16]

2.4 Recorded data

It seems important to point out that data can only be managed if they are somehow recorded, that is appear in documents. The recent interest in data in information science and in data science cannot be understood as establishing a new category besides documents, only as a growing interest in new kinds of documents (cf., Furner 2016, 297). In the abstract, Furner clearly expressed this relation: "it is not in fact the case that documents are made up of data, nor that the document is a species of dataset: rather it is the other way round, in both respects. A dataset is made up of documents; and the dataset is a species of document."

In past years, the movement of data sharing has been enjoying great popularity within LIS. In this context, Thomson Reuters (now Clarivate Analytics) launched at the end of 2012 the *Data Citation Index* as a new product in the *Web of Knowledge* family. This index covers alternative document types compared to the traditional *Web of Science* documents: data repositories, data studies and data sets. These are all explicitly referred to as document types (cf., Nielsen and Hjørland 2014, 230; See also Clarivate Analytics 2018). Compared to textual documents, however, such data documents are different and require new kinds of search systems, cf., Bugaje and Chowdhury (2018). This points to the question on how data are communicated and published. The original UNISIST model had a special communication channel for data including data centers but it seems to be an open issue whether this is actually the case (cf., Fjordback Søndergaard, Andersen and Hjørland 2003); and the more theoretical study of data publishing and its metaphors: Parsons and Fox 2013). There exist specific document types for data, for example data journals[17] and data handbooks[18], but datasets are probably more often published on the WWW or in data repositories[19] in connection with articles in scientific journals (see Nielsen and Hjørland 2014). It is also an issue to what degree specific data documents are being cited.[20] In this way data can be researched and understood in relation to classifications of kinds of documents and communication units in scholarly communication.[21]

2.5 Further definitions and conclusion

How then, should we define data? It was demonstrated above that the concept is relative (e.g. in the Borgman 2010, 3) quote: "The notion of "data" can vary considerably among collaborators … and even more so between disciplines." Does this mean that no general definition can be interdisciplinarily valid?[22] Suggestions for a general definition of data from five more researchers or group of researchers shall briefly be considered.

Floridi (2008) identifies four analyses of the concept of data[23]:

(1) The epistemic interpretation (data = facts),
(2) The informational interpretation (data are information or bits of information),
(3) The computational interpretation (data are collections of binary elements), and
(4) The diaphoric interpretation (Floridi's own preferred interpretation) which says that a datum is a lack of uniformity in some domain. Or, formally, a "datum = x being distinct from y, where x and y are two uninterpreted variables and the domain is left open to further interpretation" (Floridi 2008, 235). This interpretation was criticized by

> Lyon (2016, 743; emphasis original), "the diaphoric interpretation of data lets too many things count as data. Indeed, *any thing* x that is distinct from some other thing y will count as a datum."

Based on Floridi's criticism of interpretation 1-3 and on Lyon's criticism on Floridi's preferred definition another definition of data clearly seems to be needed.

Furner (2017, 66 emphasis original) suggested the following definition: "data are *concrete instantiations of symbolic representations of descriptive propositions, informed by empirical observation, about the quantitative and qualitative properties of real-world phenomena*." Undoubtedly, this definition covers most of what should be termed 'data'. But need data to be observational? Could they not be mathematical or theoretical? And need they be about real-world phenomena? For example, does the expression π= 3.1415926535 not contain a datum which is not observational? Or could a book about fictional figures not containing data, although they are not real-world phenomena? Again, based on such considerations, another definition seems to be needed.

Redman, Fox and Levitin (2017, 1173) (based on Fox, Levitin and Redman 1994) presented "the ordered triple view of data":

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

The authors also wrote (1172-3) "Many definitions of data have been suggested in the literature; the one that best meets the criteria discussed in the previous section is based on the classic definition of the term accepted in the database community."

Wikipedia has for a long time used the following definition: "Data is a set of values of qualitative or quantitative variables" [23a]. This definition seems to be rather widely cited (e.g., by Roy and Zeng 2015, p. 19) but without indication of authorship and without arguments or comparison with other definitions. (An alternative formulation using data as a plural seems also to be widely used: "Data are values of qualitative or quantitative variables, belonging to a set of items"). Referring to our former genealogical example where a datum could be "date of birth of a given person (X)," is found here to be better termed "a unit of analysis" rather than a variable, but we cannot here open an analysis of the concept "variable" (and this should be done by those suggesting the definition, which they have not).

Kaase (2001, 3251) provided the following definition: "Data is [are] information on properties of units of analysis." Example: A datum about a book could be that it contains 225 pages. The unit of analysis is here "page" and 225 is a datum that informs on the property of pages in the book that is their number (not about other properties of pages, for example, their size).[24] Other units could have been chosen, and different ways of what to include in counting pages can be chosen. (Whether or not a given metadata standard choses to include number of pages, and the way they are counted are, of course, "social constructs" or "captus," but still a real phenomenon).

The strength of Kaase's definition is:

- It is clear and specific and does not confuse data and documents;
- It refers to unit-phenomena, which is a general characteristic for all kinds of data (cf. Jensen above in Section 2.1).[25] (A unit can be defined as a single thing or phenomenon).
- It seems to be able to avoid the view that data are facts or that "data speaks for themselves."

It is possible to avoid the view that "data speaks for themselves" because Kaase's definition allows different interpretations about both (a) "units of analysis" and (b) the information about these units. There is no contradiction between claiming that the abstract concept "data" is precisely defined by Kaase but that different units of analysis will be chosen, understood and described differently in different contexts. Considering our former genealogy example, parent-child relationship (e.g. X is a child of Y) is a unit of analysis in genealogy, on which birth certificates, DNA-analysis reports, data from family interviews and recollections, among many other documents may provide information. Any document reporting such a child-parent relationship contains this datum, whether it is true or not. The word "information" in Kaase's definition is, as formerly explained, always relative to a question. For example, a birth certificate informs somebody about the parents of an individual. But birth certificates contain many other kinds of information. For example, somebody may ask how such certificates have developed historically or compare certificates from different jurisdictions. For the expert interested in printing and paper technology, a certificate may answer different questions. But of course, the intended purpose of birth certificates normally is their main informational function, and thus they primarily inform about birth date and place, name and parents of individuals.

Different societies, cultures, communities and sectors (state and municipalities, banks, churches, hospitals, sciences etc.) establish institutions, rules and processes to produce data and documentation that serve, as formerly explained, as sufficient facts for their main activities. In science

and scholarship observational data from measurements and experiments including computational data are good examples. In the humanities they may be data about documents (bibliographical data), about works by artists, about individuals, about dates, events and historical periods.

If we consider medicine as an example, researchers do investigations to determine, for example, the effects and side effects of different kinds of treatments. Within this domain, evidence-based medicine (EBM) has become important, with an attempt to base decision on the best available evidence. Data are here evaluated in relation to their perceived value as evidence (with data from randomized controlled trials normally given the highest status, while the opinions of patients have the lowest status). Such norms, although they are strongly influential, are debatable, cf. Hjørland 2011. This example shows, that data are not "raw", but they are carefully evaluated and selected in relation to how they have been obtained. With the words of Geoffrey Bowker (2005, 184): "Raw data is both an oxymoron[26] and a bad idea; to the contrary, data should be cooked with care." The literature about EBM shows that the norms associated with using and evaluating data are basically rooted in different epistemological positions (see, e.g., Cohen, Starvi and Hersh 2004).

Data have been difficult to define in the humanities. Keralis (2012) said:

> What counts as humanities data? The term data is unsettling for many scholars in part because it connotes something definitive and unproblematic. Where humanities scholarship often thrives on complication and constructivism, data seeks repeatability and finality. Datasets are construed as a kind of incontestable bedrock which, to some, make them not only a little boring, but dangerously and deceptively boring. Is there a way for humanities researchers to have our constructivist cake and eat it, too? Can we, in other words, productively question the constructedness of datasets even as we assemble them? And can we expand the kinds of information that constitute data?"

However, we use Kaase's definition there seems to be no problem at all. Humanists are recording information about units of analysis, for example, about works of art, artists, historical events, theoretical positions and archaeological finds.[27] Therefore, our conclusion of this Section is that Kaase's (2001, 3251) definition seems the most fruitful one suggested thus far:

> Data are information on properties of units of analysis.

3.0 Big data

3.1 Etymology

The term "big data" has been traced back to the 1990s (Ibekwe-SanJuan and Bowker 2017, 192):

> The statistician Francis Diebold is generally credited with coining the term "big data" in a paper that he presented in 2000 entitled "Big Data Dynamic. Factor Models for Macroeconomic Measurement and Forecasting." Diebold himself noted that the term was already used earlier in a non-academic context in advertisements run by Silicon Graphics International (SGI) between late 1990 and 1998. A slide deck prepared by the former Chief Scientist at SGI, John Mashey was entitled "Big Data and the Next Wave of InfraStress."[28] Another occurrence of the term was found in a 1998 computer science paper by Weiss and Indurkhya [1997]. However, it was the data analyst Douglas Laney who in 2001 made a decisive contribution towards the current characterisation of the big data by coining the popular and catchy "three V's" of big data (volume, variety and velocity) in an unpublished 2001 research note at META Group.[29]

Ekbia et al. (2015, 1524) wrote "In academia, the number of dedicated venues (journals, workshops, and conferences), initiatives, and publications on this topic reveal a continuous and consistent growing trend." They documented this with data about the occurrence of the term big data in five databases. Most occurrences appeared in ABI/Inform (Abstracted Business Information) database rapidly growing to more than 5,000 in 2013. No doubt, this has become a hot term. But what does it mean, and what kind of research programs are involved in it? How does this term affect LIS and knowledge organization? What is its real importance, and what is just hype and vogue?

3.2 Defining big data

Ekbia et al. (2015, 1525) wrote "A preliminary examination of the debates, discussions, and writings on Big Data demonstrates a pronounced lack of consensus about the definition, scope, and character of what falls within the purview of Big Data;" and Floridi (2012, 435) concluded that "it is unclear what exactly the term 'Big Data' means and hence refers to."

Most discussions of the term conclude that big data is not just a certain quantity of data, for example, more than 1 terabyte[30]. It is also just not regarded as big in relation to standard computers (e.g. amounts of data which need

main-frame computers in order to be processed). A very common way to define big data is by referring to the 3 Vs: volume, velocity and variety (Laney 2001; Zikopoulos et al. 2012[31]), all of which should be fulfilled for a set of data to be considered big.

- Volume: The quantity of generated and stored data.
- Velocity: The speed at which the data is generated and processed, often real-time
- Variety: Different kinds of data, e.g. drawn from text, images and audio

More characteristics have been introduced and discussed in the literature (cf., endnote 29). Ibekwe-SanJuan and Bowker (2017, 192) wrote:

> Having retraced the origins of the term, the question about what it is remains open. There is a consensus, at least from a physical standpoint, that big data represents volumes of data such that traditional database algorithms are unable to cope with it and that it requires more robust and distributed computer infrastructures and algorithms such as hadoop clusters, grid infrastructure and cloud clusters.

However, it is not theoretically satisfactory to define big data relative to present-day information technology. It is not "a physical standpoint" but a conceptual and theoretical standpoint that is important defining the term: A term should only be part of the terminology of knowledge organization, if it contributes developing the field.

In the USA, the National Science Foundation (2012) have specified:

> The phrase 'big data' in this solicitation refers to large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future.

Which, however, in National Science Foundation (2014) was changed to:

> The phrase 'big data' refers to data that challenge existing methods due to size, complexity, or rate of availability.

Note that the last definition does not make the claim that all the characteristics should be fulfilled. In other words, there is a retreat from tree defining V's and thereby a specific definition of the term.

Kitchin (2014, 80-99) defines the term implicitly by arguing that the production of big data has been facilitated by the confluence of five technological innovations from the start of the twenty-first century onwards:

- Growing computational power
- Dense internetworking
- Pervasive and ubiquitous computing
- Indexical and machine-readable identification
- Massive distributed storage

In combination, Kitchin argues, these developments have led to a diverse set of social-technical systems that produce what is today considered big data.

Boyd and Crawford (2012, 663) suggested the following definition:

> We define Big Data[32] as a cultural, technological, and scholarly phenomenon that rests on the interplay of:
> (1) Technology: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.
> (2) Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
> (3) Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy.

These authors here suggest a new dimension: mythology, which is here understood as critique of many assumptions associated with the term. There is no doubt that there is much hype, buzz word and vogue associated with the term – and from a scholarly point of view it is important to try to identify what is sound and theoretically important and what is not.[33]

All these characteristics (3 or 5 Vs among others) seem differently represented in different kinds of big data and thus not necessary in order to define big data. On the other hand, large data sets are not always associated with the term. Bibliometrics is a field which has for a long term used large, well organized databases. Altmetrics is a newer field, an expansion of bibliometrics to cover measures of impact based on data from social media such as blogs, tweets, bookmarks, and downloads. It comes closer to the above understanding of big data because of the heterogeneity of the data (which poses greater challenges in interpreting altmetric studies, cf., Haustein, 2016). There seems, however, to be some arbitrariness whether or not something falls under the label 'big data' and Golub and Hans-

son (2017, 1098) stated: "Data is nothing new to Library and Information Science (LIS) and Big Data presents a quantitative expansion of an already well-known object of study."

3.3 Big data in different contexts

Big data are produced in different domains. Physics and astronomy are the most data-intensive sciences. In CERN (*Conseil Européen pour la Recherche Nucléaire*), for example, which is Europe's particle-physics laboratory near Geneva in Switzerland, the particle-collision events in the Large Hadron Collider generate around 15 petabytes of data annually—the equivalent of about 4 million high-definition feature-length films (Marx 2013, 255). And the construction of the Large Synoptic Survey Telescope (LSST) in Chile is designed to produce about 15 terabytes of raw data per night and 30 petabytes over its 10-year survey life. These raw data are being refined, but even after processing, still 15 PB remains (Murray 2017). Compared to such amounts of data, Marx (2013, 257) found biology to have arrived later in the big science field, and having relatively smaller amounts of data and relatively more diverse data:

> Biology data mining has challenges all of its own, says Birney. Biological data are much more heterogeneous than those in physics. They stem from a wide range of experiments that spit out many types of information, such as genetic sequences, interactions of proteins or findings in medical records. The complexity is daunting, says Lawrence Hunter, a computational biologist at the University of Colorado Denver. "Getting the most from the data requires interpreting them in light of all the relevant prior knowledge," he says.

A third context represent the social media, like Facebook, Google[34], Twitter and YouTube collecting enormous amounts of information from us and about us, which may be used commercially, scholarly, and for other purposes. Here big data are often unintended traces that people leave during all kinds of activities. Such traces are often very vague indications and their informational value are therefore lesser compared to carefully constructed data. However, for many companies, such data have nonetheless turned out to be highly valuable because they large quantity may draw a rather detailed picture of our habits and preferences.

A basic assumption in the present article is that variety (= diversity, heterogeneity, messiness or semantic ambiguity) in both production and use of data is the main theoretical issue.

1. If the purposes for which the data are going to be used are ill-defined, then it is difficult to optimize the system providing the data. For example, because web archives are always flawed (see Finnemann, in press http://www.isko.org/cyclo/web_archive#5) and even if they were not, their metadata and search facilities could not satisfy all needs. Social, pragmatic and critical epistemology acknowledge that a given knowledge organization system cannot be neutral, but always tends to make some perspectives more visible at the cost of other perspectives. It follows that well-defined needs can be better served than ill-defined needs.
2. If the data themselves represent an ill-defined set, their merging imply that it is difficult to distinguish the relevant data from the non-relevant, and to identify how the dataset is related to different ideological positions (more about this follows in Section: 5.3 database semantics).

By implication we find that the term big data should include large data sets whether or not they are fulfilling the other criteria mentioned above. Big data should not be considered a (new) theoretical term, but just a synonym for large data sets, which may reflect greater or smaller semantic ambiguity, which is the issue of greatest theoretical importance for knowledge organization.[35]

4.0 Big data and undiscovered public knowledge

"Undiscovered public knowledge" is an important idea in LIS and for LIS. It is mostly associated with the pioneering research done by Swanson's (1986a, 1986b) work. His idea was that there is important knowledge in libraries and the literature, that is published and thereby public, yet "undiscovered" in the sense that new scientific knowledge can be generated from it, but nobody is aware of the relevance of the published work in this connection. Exemplifying with fish oil and Raynaud's syndrome, Swanson (1986a, 7-8) wrote:

> Dietary fish oil has been shown in many experiments, human and animal, to lead to reductions in blood lipids, platelet aggregability, blood viscosity, and vascular reactivity—changes that are likely to improve blood circulation. Raynaud's syndrome is a peripheral circulatory disorder associated with and exacerbated by high platelet aggregability, high blood viscosity, and vasoconstriction. These two ideas—the fish oil/blood connection and the Raynaud/blood connection—are each supported by a substantial body of scientific evidence and literature; each idea separately represents knowledge that is publicly available. What is notable about the two

> ideas is that, apparently, they have not heretofore been brought together in print. Together they obviously suggest the hypothesis that dietary fish oil might ameliorate or prevent Raynaud's syndrome. So far as I have been able to determine, that hypothesis also has never appeared in print. In some sense it has existed implicitly for years simply because the above two premises that lead to it have existed in published form for years. We can presume that the hypothesis has remained hidden because the separate literatures on fish oil and on Raynaud's syndrome have never been brought together in a way that would reveal their logical connection and so reveal the hypothesis.

Based on the study of the medical literature alone, Swanson was able to suggest a new, important medical treatment and thereby to make undiscovered public knowledge to discovered public knowledge. This idea seems even more relevant in relation to big data and is probably one of the main inspirations behind many people's engagement with data curation as well as with open data. This is caused by the idea that lots of data are relatively underutilized. About biology, for example, Marx (2013, 260) reported:

> A number of McClure's graduate students are microbial ecologists, and she teaches them how to rethink their findings in the face of so many new data. "Before taking my class, none of these students would have imagined that they could produce new, meaningful knowledge, and new hypotheses, from existing data, not their own," she says. Big data in biology add to the possibilities for scientists, she says, because data sit [sic!] "under-analysed in databases all over the world."

The idea should be considered one of pillars on which LIS and knowledge organization is based (although, as we shall see below, it may be associated with problematic epistemological assumptions).[36]

5.0 Towards a social epistemological understanding of data

As a theoretical frame for understanding data this article argues that the social epistemology originally introduced by Jesse H. Shera (1951) and since developed in the domain-analytic approach is the most fruitful approach. The theory will hopefully later be presented in an independent article in this encyclopedia. Here we shall just say that a core assumption in this theory is that different kinds of knowledge representations (information or data representations) serve different interests, and that such interests are connected to different social groups. A given representation often (if not always) is relevant only from a point-of-view but may falsely be understood as the only viable solution from all points of view, i.e., it represents an ideological position. This social epistemological view belongs to a family of philosophies (including, for example, pragmatism, critical theory and hermeneutics, cf. endnote 8) but it conflicts with other philosophical families (including empiricism). We shall consider empiricism in the next subsection, but it should be said that already the way the concept 'data' was defined above and discussed in relation to the DIKW hierarchy was informed in opposition to an empiricist understanding. Furthermore, social epistemology has specific implications for standards and database semantics as demonstrated in subsections below. These subsections provide information that are intended as indication of the implications of a social epistemological understanding of data for the field of knowledge organization, although each topic is here introduced very briefly.

5.1 The myth of empiricism

It is not a coincidence that big data is often associated with empiricism. Empiricism is the claim that all knowledge is based on our experiences, the physical signals that hit our senses, disregarding the theory-laden nature of observations. To get knowledge or to be informed is understood as collecting data, the more data, the better knowledge. Much hype about big data seems therefore to be just an ideology based on a naïve trust in empiricism and inductivism. Empiricism is a point of view which may be contrasted with rationalism, historicism and pragmatism (see, e.g., Hjørland 2017, section 4.2c: http://www.isko.org/cyclo/classification#4.2c).

Kitchin (2014, 133-6) describe four fallacies of empiricism in the context of big data. They are:

(1) big data can capture a whole of a domain and provide full resolution;
(2) there is no need for *a priori* theory, models or hypotheses;
(3) data can speak for themselves free of human bias or framing;
(4) meaning transcends context or do main-specific knowledge.

Frické (2018, Section 5.2 http://www.isko.org/cyclo/dikw#5.2) also considered such fallacies and wrote (referring to himself):

> Frické (2009) argues that the DIKW theory seems to encourage uninspired methodology. The DIKW view is that data, existing data that has been col-

> lected, is promoted to information and that information answers questions. This encourages the mindless and meaningless collection of data in the hope that one day it will ascend to information—i.e. pre-emptive acquisition. It also leads to the desire for "data warehouses," with contents that are to be analyzed by "data mining." Collecting data also is very much in harmony with the modern "big data" approach to solving problems. Big data, and data mining are somewhat controversial (Austin and Goldwasser 2008, Austin et al. 2006, Dye 2007, Frické 2015). The worry is that collecting data blind is suspect methodologically.

An alternative is the hermeneutical and pragmatic view of social epistemology that data are never "raw" (Gitelman 2013), that they must be carefully collected and processed and that it is important to understand the interests that have guided their capture, selection and processing, in particular when they are used in new contexts. Much hype connected with big data is based on a problematic empiricism, but that does not mean that big data becomes irrelevant if empiricism is criticized. It just means that the issues must be explored from other background assumptions.

Empiricism also lies behind some ideas connected with "e-science" or cyberinfrastructure two terms for the increasingly digitalization of all scientific processes and resources. While there is no doubt digitalization and new digital infrastructures in society is changing the way science, social science and the humanities is carried out, it is outside the focus of this article to discuss e-science except for the idea of a new data-driven science in which theory is obsolete, sometimes named "the fourth paradigm" (Gray 2009). Leonelli (2012) and Frické (2015) argued that this idea should not be blindly accepted.[37] As Frické (2015, 660) concludes:

> The ability to cheaply and easily gather large amounts of data does have advantages: Sample sizes can be larger, testing of theories can be better, there can be continuous assessment, and so on. But data-driven science, the "fourth paradigm," is a chimera. Science needs problems, thoughts, theories, and designed experiments. If anything, science needs more theories and less data.

This idea of a new scientific paradigm caused by some specific ways of handling data, seems clearly to represent a problematic empiricism. Gray's paper can be read as a fairly accurate, although superficial description of the consequences of digitalization for scientific communication. Its description of developments of scientific paradigms, is, however, problematic. Although scientific paradigms have a lot to do with empirical data and their interpretation, it is not just the technology for providing massive data sets that in itself causes shifts in scientific paradigms, and it is directly wrong when it is claimed that empiricism is a paradigm that was abandoned hundreds of years ago (Gray 2009, Figure 1, xviii). Paradoxically Gray's paper represents itself the empiricist standpoint: that data and data processing are sufficient for science, that theories and hypothesis are no longer needed. This way of thinking causes a block for better understanding theoretical problems related to data and knowledge organization. Therefore, although big data is the background for "e-science," e-science does not define big data and cannot do without theory.

5.2 Standards

To manage big data, they must be classified, indexed or assigned metadata. Such classifications can be local or general, where local can mean both domain specific and geographical local as in the example about ecological research below. The more general a classification is, the better are the possibilities for sharing data and equipment, for achieving interoperability (http://www.isko.org/cyclo/interoperability) among systems and processes (see Zeng, in press). Standards are often conceived as technical solutions being developed to ensure optimal interconnection, reuse and cooperation between people and systems. They are not usually considered political. It may be the case that some standards, such as standards for paper sizes are relatively neutral, but in the case of knowledge organization systems (see Mazzocchi 2018), standards cannot be neutral, but always tends to support certain goals at the cost of other goals,[38] corresponding to the view of social epistemology expressed above. A fine case-study is provided by Millerand and Bowker (2009): The U.S. network for long-term ecological research (LTER). It was presented on p. 152-3:

> The LTER program constitutes a distributed, heterogeneous network of more than 1200 research scientists and students. Formed in 1980, the network currently consists of 26 sites or research stations (ironically, some 'long term' sites have already closed; and more have been added). Each is arranged around a particular biome – for example a hot desert region, a coastal estuary, a temperate pine forest or an Arctic tundra – in the continental United States and Antarctica. A 27th site is charged with the administration and coordination of the group. The program's mission is to further understanding of environmental change through interdisciplinary collaboration and long term research projects.

> One of the chief challenges of LTER is to move beyond the 'plot' of traditional ecoscience to analyze change at the scale of a continent and beyond the 5-year funding cycle or 30 year career cycle of the scientist to create baselines of data spanning multiple decades. While the preservation of data over time, and their storing in conditions appropriate to their present and future use, has always been a priority within the different sites of the LTER network, there has been a new urgency with the development of a cyberinfrastructure project aiming to encourage data sharing across the community.

We shall not here consider the more technological challenges described by Millerand and Bowker but focus on the perspective of relevance to social epistemology. Although standards are technical and practical devices, they are still also theoretically influenced (in a broader understanding of "theory," see Hjørland 2015) and should be theoretically analyzed. Technical solutions are developed on the basis of theoretical assumptions and their implications. The case study in Millerand and Bowker describes LTER as a research field consisting of many different sites, each using its own standards and metadata. The overall network had a clear, recognized interest in being able to combine data from many different sites to make new kinds of analyses, which are long-term, geographical broad and involving different kinds of observations. Two major objectives were generally agreed: (a) the promotion of interdisciplinary collaboration through data sharing and (b) the improvement of long-term data preservation. To provide a standard for metadata allowing such combinations of data is, as already said, often conceived as being a simple technical solution (Millerand and Bowker 2009, 150) and (153): "In an ideal world, the metadata contain all the details necessary for all possible secondary users of a dataset."

The Ecological Metadata Language (EML) is a standardized metadata description language for the generation of metadata in the domain of environmental sciences and was the standard which the LTER research community decided to adopt when it engaged in the process of standardizing its scientific data management practices. Millerand and Bowker described two narratives on how the implementation of the standard was received (or perceived):

> Narrative 1: "EML is a success: the entire LTER community has adopted it" and
> Narrative 2: "EML is not (yet) a success: it needs to be redeveloped before it can be used."

However, everybody seems enthusiastic about its implementation. Millerand and Bowker (2009, 157):

> It is striking the degree to which all of the actors involved in the standardization process (EML developers, LTER network coordinators, information managers, domain researchers, and so on) have supported – and continue to support – the EML project ...., they all believe in the idea of a metadata standard permitting the exchange and sharing of data throughout the LTER network and beyond. In this sense, it's not the case of the imposition of a standard by one group of actors (developers and coordinators) on a hostile, resistant group (information managers). The latter have always been highly supportive of the project.

The quote continues: "It is at the moment of the actual implementation of the standard at a given site when critical problems emerge, and discordant voices can be heard." So, on one level everybody was enthusiastic because they could see great possibilities for new kinds of research questions and answers. But when it came to the implementation at their own site, they discovered losses if their own local standards had to be changed to the new standard.

The conclusion that Millerand and Bowker draws from their case study is that the standard had to be redefined and be the object of multiple versions over the course of its development. The implementation of EML is not simply a case of upgrading an existing system. It consists above all in redefining the sociotechnical infrastructure which supports this articulation of technical, social and scientific practices and (165): "Both standards and ontologies (the one apparently technical and the realm of machines, the other apparently philosophical and the realm of ideas) need to be socially, organizationally bundled – not as a perpetual afterthought but as an integral necessity."

A simpler conclusion can be made: Metadata organized for one purpose may not be optimal for other purposes. Any use of data should ideally have metadata (and other procedures and tools) optimized for the specific purpose. When new standards serving other and broader goals are implemented, discordant voices arise when locally developed metadata have to fit local goals less in order to serve other interests. A core sentence in the article (Millerand and Bowker 2009, 153) is this: "In an ideal world, the metadata contain all the details necessary for all possible secondary users of a dataset."[39]

From a philosophical perspective, the dream of the perfect standard thus seems to be related to the rationalist dream of a perfect language (cf., Laporte 2018). However, from the point of view of social epistemology, this will forever remain just a dream. What *is* possible, is to construe standards which are carefully developed to support specific goals – or compromises based on careful negotiations between different goals, which have first to be iden-

tified. This imply a movement of focus from a narrow perspective to a broader social and philosophical perspective.

5.3 Database semantics

Database semantics is about the meaning of data in databases. The field of data semantics seems to be well established in computer science. In 2012 the *Journal of Data Semantics* was established, and in its first editorial Stuckenschmidt (2012, 1) wrote:

> Data Semantics (Sheth 1997) is a topic that has been investigated in computer science for more than 30 years. It is typically associated with a formal definition of the intended interpretation of the data often in terms of logic or algebraic formalisms (Veltman 1984). Over the time [sic!], the goals of defining data semantics as well as the ideal of having a clear formal representation of semantics has not change [sic!], what has changed and is constantly changing, however, are ways of capturing and using the semantic of data as well as the formalisms used to represent it.

What is important, however, (cf., Section 3.3) is the semantic ambiguity (related to variety, diversity, heterogeneity and messiness) in both the production, merging and use of data. (Stuckenschmidt 2012, 3) wrote about this that in closed systems the meaning of data is defined by its intended use determined by the systems' developers and users and often reflected in the specific schema or the data structures used for representing it. To illustrate if you have "1984" in a bibliographic database, this can mean different things such as printing year or the title of a book (by George Orwell). In standard databases this is solved by having titles and printing in separate fields, to which searches can be specified. If all such data were merged with specified fields, users could not use "1984" to search for the book, because the overwhelming part of the retrieved references would be to documents having 1984 as the printing year. This is a simple case of semantic holism, the principle that the meaning of a word or sign depends on the context in which it is located. In Kuhn's (1962) theory semantic holism is related to scientific theories: Words and signs gets their meaning from the theories or "paradigms" of which they form part, and when "paradigm shifts" occur terms get new meanings. Example:

– Paradigm one: Ptolemaic astronomers might learn the concepts [star] and [planet] by having the Sun, the Moon, and Mars pointed out as instances of the concept [planet] and some fixed stars as instances of the concept [star].
– Paradigm two: Copernicans might learn the words "star", "planet", and "satellites" by having Mars and Jupiter pointed out as instances of the concept [planet], the Moon as an instance of the concept [satellite], and the Sun and some fixed stars as instances of the concept [star]. Thus, the terms "star," "planet," and "satellite" got a new meaning and astronomy got a new classification of celestial bodies.

This mean that if different paradigms are represented in the same database, that database may use the same terms in different ways. Another example in knowledge organization literature is provided by Gnoli's (2014, 130) analysis of Marc Bloch's methodology of historical knowledge, where the French historian observes that the meaning of such terms as bourgeoisie or feud is dependent on their specific chronological context and cannot be applied to different contexts unproblematically.

Of course, as Stuckenschmidt wrote (2012, 3), the systems designer in closed systems may standardize the meanings (by using controlled vocabularies, CV), which however, introduces a new layer of interpretation that users have to relate to. This is an issue that has been neglected in much literature on CV. We shall here limit this discussion by referring Maniez (1997), who pointed out that paradoxically the information languages increase the difficulties of cooperation between the different information databases, confirming what Lancaster (1986b, 181) observed earlier: "Perhaps somewhat surprisingly, vocabularies tend to promote internal consistency within information systems but reduce intersystem compatibility" (see also Zeng In Press). In order to understand database semantics, we need therefore to consider how the meanings of terms are applied in the primary literature and from there consider how these meanings are represented in the secondary literature (such as bibliographical databases). The web is probably a much better approximation to how meanings are developed in different kinds of discourses.

How can we work with database semantics in open systems, such as the WWW? Stuckenschmidt presents three approaches providing data semantics:

> 1. Semantics from models
> This is about the use of standards such as RDF and OWL, which Stuckenschmidt (3) quotes as being "a cornerstone of data semantics on the web", but (3) "publishing the ontology along with a data set does not really solve the problem as long as every data set comes with its own ontology. In this case the problem of possible misinterpretations is just lifted from the data to the conceptual level." Stuckenschmidt mentions the use of a jointly shared ontology, a top-level ontology or the use of semantic mappings.[40]

> 2. Semantics from data
> This is, for example, about the use of statistical approaches, linguistic approaches and learning approaches to extract semantic information from data sets.
>
> 3. Semantics from users
> This is described by Stuckenschmidt (4) as a new trend in data semantics. He further describes this approach: "In contrast to classical knowledge acquisition scenarios, it is not assumed that the users are experts in the particular domain. The idea is more that asking many users the same question will ultimately generate the right result because a majority of users will give the correct interpretation of the piece of data. The most prominent approach of this type is tagging, where users attach simple descriptions (tags) to information objects"

These approaches outlined by Stuckenschmidt (2012) seem based on the rationalist idea of one neutral and best interpretation of concepts: a view from nowhere rather than a view from a particular perspective. There seem to be a lack of understanding that semantic relations are not neutral but related to specific goals, interests, and paradigms etc. That means, that there is a danger that using a given approach – e.g. user tagging (see Rafferty (2017, Section 4: http://www.isko.org/cyclo/tagging#4) – may just represent a major ideology, not the interest of the specific purpose for which a system should be designed.

From the point of view of social epistemology, the core issue with databases is that data are taken out of some contexts and put into other contexts, where the different contexts may represent implicit knowledge or conflicting perspectives, paradigms, interests and goals; as put by Leonelli (2016): Data make journeys, they travel. Such journals have to be considered from the perspective of semantic holism.

The implication of this view is that an important issue in data semantics is to uncover the main different meaning structure*s* (paradigms, perspectives or voices) in databases, on the web and in different fields of knowledge. One should not start by studying the data isolated but start by studying their overall contexts in a top-down fashion (or interactively top-down and bottom-up). Hjørland (1998, 27) provided an example searching for "the effects of lead on behaviour" in the psychological database PsycINFO; here there is no need to indicate that you are searching lead's effects on behaviour because it is implicit in PsycINFO that all records are on animal or human psychology/behaviour. However, if the records in PsycINFO are merged with the records in Chemical Abstracts, you will have to change your search strategy and specify that you are searching studies on how lead influences behaviour and performance. This new strategy would probably be less than optimal regarding the part of the records originally indexed in PsycINFO (because implicit information is lost by the merging). At another level PsycINFO can be seen as a merging of records which were once presented in individual journals, some of which may be American, some European, some behaviouristic, other psychoanalytic, etc. Originally, to the readers of those journals their selection policy and their way of writing titles and composing articles reflected some implicit meanings in those journals. By making a controlled vocabulary, a classification scheme, a certain structure in the records and so on the people behind PsycINFO made certain decisions which were coloured by their view of knowledge.

In other words, a given set of texts represents different "paradigms"[41] (or voices as understood by Bakhtin, 1981 and 1986[42]), which should be identified and thereby knowledge organization and information retrieval should provide choices for users to consciously select given "voices" based on informed choices. In a way this corresponds to the former mentioned terms variety, diversity, heterogeneity and messiness, but from the perspective that this diversity contains different values and may have their own conceptual structures, which should be identified and made visible. The more heterogenous the big data sets are, the more implicit knowledge is lost, and the more important the reconceptualization becomes.

Leonelli (2014) is a paper from the philosophy of biology which is based on an idea related to the idea of merging and re-identify conceptual structures. It describes three stages of data travel: (1) De-contextualization (2) Re-contextualization and (3) Re-use:

> 1. Data de-contextualization involves according to Leonelli (2014, 4) the assignment of metadata "to decontextualize the data that are included in their resources, so that they can travel outside of their original production context and become available for integration with other datasets (thus forming a Big Data collection). The process of de-contextualisation involves making sure that data are formatted in ways that make them compatible with datasets coming from other sources, so that they are easy to analyse by researchers who see them for the first time."
>
> 2. Data re-contextualization involves according to Leonelli (2014, 4) use of data in new context: "Within biology, re-contextualisation can only happen if database users have access not only to the data themselves but also to the information about their provenance – typically including the specific strain of organisms on which they were collected, the instruments and procedures used for data collection, and the composition

of the research team who originated them in the first place. This sort of information, typically referred to as 'metadata' (Edwards et al., 2011; Leonelli, 2010), is indispensable to researchers wishing to evaluate the reliability and quality of data. Even more importantly, it makes the interpretation of the scientific significance of the data possible, thus enabling researchers to extract meaning from their scrutiny of databases."

3. Data re-use Leonelli (2014, 5-6):

One of the central themes in Big Data research is the opportunity to re-use the same datasets to uncover a large number of different correlations. After having been de-contextualised and recontextualised, data are therefore supposed to fulfil their epistemic role by leading to a variety of new discoveries. From my observations above, it will already be clear that very few of the data produced within experimental biology make it to this stage of their journeys, due to the lack of standardisation in their format and production techniques, as well as the absence of stable reference materials to which data can be meaningfully associated for re-contextualisation .... Whenever data and metadata are added to a database, curators need to tag them with keywords that will make them retrievable to biologists interested in related phenomena. This is an extremely hard task, given that curators want to leave the interpretation of the potential evidential value of data as open as possible to database users. Ideally, curators should label data according to the interests and terminology used by their prospective users, so that a biologist is able to search for any data connected to her phenomenon of interest (e.g. "metabolism") and find what the evidence that she is looking for is. What makes such a labelling process into a complex and contentious endeavour is the recognition that this classification partly determines the ways in which data may be used in the future – which, paradoxically, is exactly what databases are not supposed to do.

The important conclusions are of this section are:

1. Databases are supposed to answer a broad range of different needs, but nonetheless, the way metadata are assigned (or not assigned), and data thereby classified, determines partly the ways the database can fruitfully be used.
2. Data are always produced for some purposes and perspectives. Before data "travel" and are merged in databases, relevant distinctions may be implicitly present, and this implicit knowledge may be lost.
3. To make a database effective, data must be classified and coded. But this classification cannot be neutral in relation to the purposes for which the database is going to be used. The data/information scientist involved must understand that she is a part of a theoretical struggle in the domain. His/her main task is to study, identify and understand the mostly implicit theoretical positions in this struggle and make decisions based on informed arguments. This means that the coding of data reestablishes some of the implicit knowledge lost during the merging.

6.0 Conclusion

The examination of the issues related to data and big data confirms the need to look at different domains. Rather than having many kinds of generalists such as information scientists and data scientists there is a need for specialists in, for example, biological information/data scientists or specialists in digital humanities. The relevance of epistemological perspectives exists at two levels:

– Data are always produced from certain perspectives, which contain their own criteria of relevance. The information scientist (data scientist, knowledge organizer) thus faces a merging of different views, often implicit, competing for influence, and this calls for "data criticism" (Beaton 2016); and, information science (data science, knowledge organization) is dependent on tools, approaches, theories etc. which also are based on epistemological assumptions, often implicit assumptions.

We find that the best theoretical frame to study big data is an approach which recognizes the importance of epistemology, which is a much-neglected perspective today. "Social epistemology" originally founded by Shera (1951) but recently made more concrete and forceful in knowledge organization is here considered the most fruitful philosophy. We should always consider how we in LIS conceptualize the phenomena, we study, including data and big data (cf., Johansson 2012, 28ff). This is an argument for more conceptual, theoretical and philosophical studies (cf., Furner 2017).

Notes

1. Lindberg (2012, 199; emphasis original): "In Latin, *data* is the plural of *datum* and, historically and in specialized scientific fields, it is also treated as a plural in English, taking a plural verb, as in *the data were collected and classified*.In modern non-scientific use, however, it is generally not treated as a plural. Instead, it is treated as a mass noun, similar to a word like *information*, which

takes a singular verb. Sentences such as *data was collected over a number of years* are now widely accepted in standard English."

The IEEE Computer Society allows usage of data as either a mass noun or plural based on author preference. Other professional organizations and style guides require that authors treat data as a plural noun. For example, the Air Force Flight Test Center specifically states that the word data is always plural, never singular. Furner (2016, 300, note 25) wrote: 'The first [document] is a count noun; the second [information] is a mass noun; and the third [data] is sometimes used as a count noun, sometimes as a mass noun. It makes sense to talk of "ten documents," just as it does to talk of "ten dollars"; but we would not usually choose to talk of "ten informations," just as we would not normally say "ten monies." "Data," of course, is an odd case: its origin in Latin as a plural form leads many writers to insist on its taking a plural verb (e. g., "The data are …"), but presumably even those scholars would balk at using formulations like "ten data."

Wikipedia writes: "Data is most often used as a singular mass noun in educated everyday usage. Some major newspapers such as *The New York Times* use it either in the singular or plural. In *the New York Times* the phrases "the survey data are still being analyzed" and "the first year for which data is available" have appeared within one day. In scientific writing data is often treated as a plural, as in These data do not support the conclusions, but it is also used as a singular mass entity like information. British usage now widely accepts treating data as singular in standard English, including everyday newspaper usage at least in non-scientific use. UK scientific publishing still prefers treating it as a plural. Some UK university style guides recommend using data for both singular and plural use and some recommend treating it only as a singular in connection with computers."

2. "Graphic record" was, for example, used by Shera (1951, 80) and elsewhere.
3. "Information explosion" is, according to the *Oxford English Dictionary* a term first used in 1941.

   Rider (1944) estimated that American university libraries were doubling in size every sixteen years. Given this growth rate, Rider speculates that the Yale Library in 2040 will have "approximately 200,000,000 volumes, which will occupy over 6,000 miles of shelves… [requiring] a cataloging staff of over six thousand persons."

   Saracevic (1999, 1052) wrote: "Information science is a field that emerged in the aftermath of the Second World War, along with a number of new fields, with computer science being but one example. The rapid pace of scientific and technical advances that were accumulating since the start of the twentieth century, produced by midcentury a scientific and technical revolution. A most visible manifestation of this revolution was the phenomenon of "information explosion," referring to the exponential and unabated growth of scientific and technical publications and information records of all kinds ("literature")."
4. Ekstrøm and Wildgaard (2019) is a book about data science librarianship.
5. The term *data science* sounds, from a Danish ear in particular, somewhat strange, because computer science in Danish is "datalogi" (meaning the study of data). It was named so by computer scientist Peter Naur (Naur, 1966); see also Wikipedia, 2018-05-07: https://en.wikipedia.org/wiki/Peter_Naur.
6. Rowley (2007, 170-1; references omitted) wrote:

   > 5.2. Defining data:
   >
   > Where definitions of data are offered these are typically clearly and succinctly stated, sometimes with examples. In summary the definitions variously suggest that:
   >
   > – Data has no meaning or value because it is without context and interpretation [27, 40–42].
   > – Data are discrete, objective facts or observations, which are unorganized and unprocessed, and do not convey any specific meaning [20, 37, 38, 41].
   > – Data items are an elementary and recorded description of things, events, activities and transactions [43–45].
   >
   > Choo [25] suggests that data are often elements of larger physical systems (such as books, or instrument panels) which give clues about what data to notice and how they should be read. Jashapara [39] and Choo [25] also introduce the concept of signals. Jashapara [39] suggests that we acquire data from the external world through our senses and try to make sense of these signals through our experience. Choo [25] develops this further and specifically identifies signals as the origin of data, and proposes the processes of sensing and selecting, together described as physical structuring, as transforming signals into data.
   >
   > Interestingly, these definitions are largely in terms of what data lacks; data lacks meaning or value, is unorganized and unprocessed. They lay the foundations for defining information in terms of data.
7. Kitchin (2014, 2; emphasis original) citing Jensen's quote commented: "Strictly speaking, then, this book

should be entitled *The Capta Revolution* [rather than The Data Revolution]"

8. See, for example, Hjørland (2005) and Hjørland (2017, section 4.2c: http://www.isko.org/cyclo/classification#4.2c), for a brief presentation of empiricism, rationalism, historicism and pragmatism.
9. Concerning the development of the distinction between data and information Gray (2003) is relevant although that article only covers findings from the information systems literature.
10. The reliability of DNA test depends on different issues, including how the test was done and how closely related the persons are. The quote about Mr. Jim Doe being the biological father of John Test is from this report: https://cdnmedia.eurofins.com/vaterschaftsanalyse/media/105508/positive-dna-report-eurofins-forensics.pdf Saved in WebCite: http://www.webcitation.org/72wPWlP8K
11. Karen Spärck Jones (1987, 9), for example, wrote "we are concerned with access and, more materially, indirect access to the information user wants: he wants the information in the documents, but the system only gives him the documents." Again, this is a view connected with an empiricist/positivist philosophy, which, as Capurro wrote in the quote above has been proven wrong. As van Rijsbergen and Lalmas (1996, 386) wrote: " In the early days of Information Retrieval (van Rijsbergen, 1979), people used to qualify their statements about information retrieval (IR) by saying that really they were working on document retrieval. It was denied strenuously that information was being retrieved. As Lancaster (1968a) wrote, "An information retrieval system does not inform (i.e., change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request."
    The situation has changed. We believe that the purpose of an information retrieval system is to provide information about a request and that a request is a representation of an information need that an IR system attempts to satisfy. Hence, a fundamental problem is how to compute the information contained in one object (e.g., a document) about another (e.g., a query). Thus, if a user states a query then it behooves the IR system to find the objects that contain information about that query."
    van Rijsbergen and Lalmas (1996) seems to be the only serious argument that information retrieval can be justified as a correct term, and it seems itself related to the empiricist and positivist philosophy which was rejected by Capurro (in Zins, 2007, 481) and by many others.
12. Borgman (2007, 41) wrote: "Meadows's (2001) treatise on information examines related concepts such as data, classification, storage, retrieval, communication, knowledge, intelligence, and wisdom. He distinguishes between data, facts ('inferences from data' [8]), and information, and yet does not offer firm definitions of any of them. Rather, he acknowledges the 'hazy border' (14) between data and information. What are raw and unprocessed data to one person may be meaningful information to another. Similarly, data that are background context to one researcher, such as micrometeorological measurements, may be the focus of research to another."
13. National Research Council (1999, 15;emphasis original): "*Data* are facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors. A data element is the smallest unit of information to which reference is made .... For purposes of this report the terms *data* and *facts* are treated interchangeably, as is the case in legal contexts.
    Data in a database may be characterized as predominantly *word oriented* (e.g., as in a text, bibliography, directory, dictionary), *numeric* (e.g., properties, statistics, experimental values), *image* (e.g., fixed or moving video, such as a film of microbes under magnification or time-lapse photography of a flower opening), or *sound* (e.g., a sound recording of a tornado or a fire). Word oriented, numeric, image, and sound databases are processed by different types of software (text or word processing, data processing, image processing, and sound processing).
    Data can also be referred to as *raw, processed*, or *verified*. Raw data consist of original observations, such as those collected by satellite and beamed back to Earth, or initial experimental results, such as laboratory test data. After they are collected, raw data can be processed or refined in many different ways. Processing usually makes data more usable, ordered, or simplified, thus increasing their intelligibility. Verified data are data whose quality and accuracy have been assured. For experimental results, verification signifies that the data have been shown to be reproducible in a test or experiment that repeats the original. For observational data, verification means that the data have been compared with other data whose quality is known or that the instrument with which they were obtained has been properly calibrated and tested.
14. This definition was adapted by some ISO standards. 5127:2017 defined: "3.1.1.15 Data, pl.: reinterpretable representation of *information* (3.1.1.16) in a formalized manner suitable for *communication* (3.1.8.04), interpretation, or processing (Source: ISO/IEC 2382:2015 (en), 2121272). Note 1 to entry: *Data* are often understood as taking the form of a set (3.1.1.09) of values or qualitative or quantitative variables."

15. Leonelli (2016, 77; emphasis original): "I propose to define data as any product of research activities, ranging from artifacts such as photographs to symbols such as letters or numbers, that is collected, stored, and disseminated *in order to be used as evidence for knowledge claims*." Rapley and Rees (2018) call their paper "Collecting Documents as Data," but is rather about documents as sources of data.
16. Borgman (2015, 29; emphasis original): "However, use of the term data varies by context and by user, often in subtle and significant ways. Particularly in the case studies, usage follows the conventions of the field being discussed. Unless used as a concept or when conventions are noted, *data refers to entities used as evidence of phenomena for the purposes of research or scholarship*."
17. UlrichsWeb does not list data journal among its serial types but see Candela et al. (2015) for a survey of data journals.
18. The literature about data handbooks is very small, almost non-existing. One of the best-known examples of data handbooks is *CRC Handbook of Chemistry and Physics: A Ready-reference Book of Chemical and Physical Data* (Haynes, Lide and Bruno 2017; known as the Rubber Bible). Among the papers about data-handbooks is Gurr (1974), which is a review of some data-handbooks in political science.
19. Kaden and Kleineberg (2018, 3): "The publication of your research data increases the transparency and reproducibility of your research results, makes hidden scientific work visible and opens up possibilities for reuse. Your research data will be given a citable persistent identifier (e.g., DOI - Digital Object Identifier) and can be considered as genuine scientific publications."
20. Among the published studies on data citation are, for example, Zhao, Erjia and Li (2018).
21. Voss (2013) considered data as signs with departure in the understanding of the concept "document" in LIS.
22. In a discussion 2018-09-25 Professor Niels Ole Finnemann argued that the concept of data must be relative to medium, that the concept from the print media cannot be used in today's complex computer processes. However, there seems not to be any reason the Kaase's definition cannot be used also in the context suggested by Finnemann: Different researchers may define units of analyses and criteria of measuring them that are relevant for their research.
23. Compare Leonelli (2016, 71-6): "3.1 Data in the philosophy of science", who distinguished two main views: (1) "the representational view of data" (2) "the practice turn."

23a. Richard Wilkin on the "Talk page" in Wikipedia on 12 August 2015 (https://en.wikipedia.org/wiki/Talk:Data#Data_Definition_vs_Information): "proposed the following definition: "Data is any set of quantitative and/or qualitative values. An information system either captures data by measuring the attributes of a phenomenon, or it is provided from another information system – be it mechanical (e.g. a computer), or biological (e.g. a human).." However, very similar definitions were formerly used in Wikipedia.

24. Perhaps, in most cases the definition could be simplified: "Data are information on units of analysis." For example, the sentence "my mother has 39° in fever" is a datum informing about the unit of analysis, which is temperature measured in Celsius degrees. "X is the child of Y" is a statement informing about the datum "parent child relation" between two persons X and Y.
25. Jensen (1950, ix) only considered the unit-phenomena in science, but the data concept is also used outside science, and also here are data about units of various kinds. Units need not be about measurements. Male/female and colors, for example, can be data and assigned without measurement and birth certificates reports sex of a child as a datum.
26. About the understanding of "raw data" as an oxymoron see also Gitelman (2013).
27. Kaase's definition of data thus also covers humanities although qualitative researchers may find with Markham (2018, 521) "that putting data at the forefront of qualitative inquiry is the worst sort of trap because it deludes us into thinking we have already collected the knowledge when, in actuality, what qualitative inquiry produces is a bricolage of multiple voices, actors, and perspectives filtered through our own unique gaze and interpretive lenses."
    But we may ask if, for example, such multiple perspectives be considered "data" in the kind of research suggested by Markham?
28. "[note 6 in Ibekwe-SanJuan and Bowker 2017] The company's overview affirms its chief scientist's claim to paternity of the term "In the late 90s, SGI's Chief Scientist at the time, John R. Mashey coined the term 'Big Data.'" https://www.sgi.com/company_info/overview.html" (Ibekwe-SanJuan and Bowker 2017, 196).
29. The quote continues: "[note 7]. Laney's 3 Vs later expanded into 4 Vs (3 Vs + Validity) and now has a fifth V as well (4 Vs + Veracity)." Note 7 (196): "According to Diebold (2012), "META is now part of Gartner.."
30. A bit is an abbreviation for 'binary digit'; the unit in the binary code (0 or 1) used to store and process data. There are different ways to express larger quantities of data as indicated in the quote below, the traditional one being:
    - A byte is 8 bits,
    - A kilobyte (KB) is 2^{10} bytes or 1,024 bytes;
    - A megabyte (MB) is 2^{20} bytes or 1,024 kilobytes;

- A gigabyte (GB) is 3^{30} bytes or 1,024 megabytes;
- A terabyte (TB) is 2^{40} bytes or 1,024 gigabytes;
- A petabyte (PB) is 2^{50} bytes or 1,024 terabytes;
- An exabyte (EB) is 2^{60} bytes or 1,024 petabytes;
- A zettabyte (ZB) is 2^{70} bytes or 1,024 exabytes;
- A yottabyte is 2^{80} bytes or 1,024 zettabytes.

However, to avoid confusing the terminology of the *International System of Units* (SI), where one kilo means 1,000 (and not 1,024) other prefixes have been suggested: "These SI prefixes refer strictly to powers of 10. They should not be used to indicate powers of 2 (for example, one kilobit represents 1000 bits and not 1024 bits). The IEC [International Electrotechnical Commission] has adopted prefixes for binary powers in the international standard IEC 60027-2: 2005, third edition, Letter symbols to be used in electrical technology—Part 2: Telecommunications and electronics. The names and symbols for the prefixes corresponding to 2^{10}, 2^{20}, 2^{30}, 2^{40}, 2^{50}, and 2^{60} are, respectively: kibi, Ki; mebi, Mi; gibi, Gi; tebi, Ti; pebi, Pi; and exbi, Ei. Thus, for example, one kibibyte would be written: 1 KiB = 2^{10} B = 1024 B, where B denotes a byte. Although these prefixes are not part of the SI, they should be used in the field of information technology to avoid the incorrect usage of the SI prefixes." (Bureau International des Poids et Mesures 2006, 121). See also Wikipedia: https://en.wikipedia.org/wiki/Megabyte.

31. Zikopoulos et al. (2012, xxvi): "Chapter 1 talks about the three defining characteristics of Big Data: volume (the growth and run rates of data), variety (the kinds of data such as sensor logs, microblogs—think Twitter and Facebook—and more), and velocity (the source speed of data flowing into your enterprise)."
32. "We have chosen to capitalize the term 'Big Data' throughout this article to make it clear that it is the phenomenon we are discussing" (Boyd and Crawford 2015, 663).
33. It is not healthy for scholarly fields to develop terminology, that is not well motivated, but just reflects fashion or "hot" terms and attempts to look smart. A serious field works on construing a set of necessary and sufficient terms that are associated with its basic theories. Social pressure to deal with hot terms rather than with well-constructed terms and theories may indeed be a barrier for scientific progress.
34. One of Google's many projects is Google Books with its Ngram viewer. Pechenick, Danforth and Dodds (2015) and Zhang (2015) demonstrates how the interpretation of big data often errors because of missing knowledge about the composition of the data sets. Zhang, for example, describes an erorounous interpretation that there in English books was a drop of the use of the term "fuck" in the period 1810-1960. This is due to problems with OCR-recognition: The lowercase long s in certain old books looks a lot like a f. Pechenick, Danforth and Dodds (2015) demonstrates how important it is to know, for example to which degree Google Books cover fiction, nonfiction and scientific literature from different periods. This example demonstrates one of the core assumptions of the present article, that data need interpretation on the basis of knowledge in specific domains.
35. The reason "big data" has become a common term is that other characteristics than size are very important for managing data from an IT point of view, but that does not indicate they are important for a theory of data from the point of view of LIS and knowledge organization. However, the task here is to evaluate the term in the theoretical context of knowledge organization. Whether or not it is useful in other contexts is not a relevant discussion here.
36. Ibekwe-SanJuan and Bowker (2017, 188) wrote: "We hope that this discussion will help bridge the current gap between two research communities (and their literatures) which have existed separately until now: the KO community on the one hand, and the data analysis and machine learning community on the other." This can also be understood as a claim that undiscovered public knowledge exists in the two communities. As such, this example points to an issue, that was relatively neglected by Swanson: That discovering of relevant knowledge is closely related to issues of scientific paradigms. (Whether Ibekwe-SanJuan and Bowker's claim is true or false is irrelevant for the point here made).
37. Hine (2006, 292) wrote: "Consequently, while practices and outcomes of knowledge production may change with increasing use of information and communication technologies, such changes do not do away with existing frameworks or necessarily produce new epistemic cultures."
38. An anonymous reviewer asked me to back-up the claim that standards support certain interests at the cost of other interest with references to the literature. The best source is Lampland and Star (2009) (including the chapter by Millerand and Bowker discussed in this Section 5.2). Most literature on this topic addresses specific domains, such as the ideology of language standardization (Milroy 2001) or standardization of accounting rules (Ramanna 2015) etc. etc. The general claim is supported by the growing literature about the impossibility of neutral knowledge organization systems (e.g., Mai 2011 and Guimarães 2017). The most important arguments come from voices in epistemology and the philosophy of science.
39. The quote continues: "(an ideal solution that evokes Baruch Spinoza's problem — to know a single fact

about the world, we need to know every fact about the world)."

40. Finnemann (in press, Section 6): "A related project is the semantic Web, initiated by Tim Berners-Lee, the creator of the Web protocols, aiming to 'bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users' (Berners-Lee, Hendler and Lassila 2001, 3; Berners-Lee, Shadbolt and Hall. 2006). The project is built on the claim that it is possible to automatize semantic analyses of materials to create coherent semantic metadata, which can be used by the machine either by help of an AI inference system or as automatic creation of linked data. Whether this is possible beyond controlled vocabularies within in a formalized semantic universe remains to be seen. In a linguistic perspective it is difficult to perceive such systems remain stable in a long-term perspective."
41. The concept "paradigm" became influential after Kuhn (1962), but not in the sense indicated in that book. Whereas Kuhn only saw one paradigm at a given time, today the idea of multiple, competing paradigms is the norm. Kuhn did not recognize different schools in the social sciences as paradigms, but this is nonetheless the way paradigms is mostly understood today.
42. Bakhtin (1981, 1986) introduced the concept of "voice" according to which an utterance is always produced by a certain voice, a speaking personality with a specific viewpoint. Specific voices, being invoked and informed as responses in the conversational and collaborative situation at hand, are also informed by a broader socio-cultural context with a particular history. Such social languages correspond to what Bakhtin called speech genres. These refer to types of utterances produced by types of voices. Speech genres can be recognized by typical situations of speech communication, by typical themes and meanings of words that are addressed. Advancing a certain theoretical viewpoint involves talking in terms of that theory, although it simultaneously depends on how the specific person understands that theory, what he or she places within that category.

    Multivoicedness does not, however, just mean the juxtaposition of voices; not just that persons said what they meant. New meaning, new insight and understanding is, according to Bakhtin, dependent on the tension between different voices, viewpoints and perspectives.

    Any sign has an orientation toward plurivocality, polylogism, and multivoicedness, and therefore an ability to adapt to new and different situational contexts.

References

Austin, Peter C., and Meredith A. Goldwasser. 2008. "Pisces did not have Increased Heart Failure: Data-driven Comparisons of Binary Proportions between Levels of a Categorical Variable can Result in Incorrect Statistical Significance Levels." *Journal of Clinical Epidemiology* 61: 295-300.

Austin, Peter C., Muhammad M. Mamdani, David N. Juurlink, and Janet E. Hux. 2006. "Testing Multiple Statistical Hypotheses Resulted in Spurious Associations: a Study of Astrological Signs and Health." *Journal of Clinical Epidemiology* 59: 871-2.

Bakhtin, Mikhail Mikhaĭlovich. 1981. *The Dialogic Imagination: Four Essays*, ed. and trans. Michael Holquist and Caryl Emerson. Austin: University of Texas Press.

Bakhtin, Mikhail Mikhaĭlovich. 1986. *Speech Genres and Other Late Essays*, ed. Caryl Emerson and Michael Holquist; trans. Vern W. McGee. Austin: University of Texas Press.

Beaton, Brian. 2016. "How to Respond to Data Science: Early Data Criticism by Lionel Trilling." *Information & Culture* 51: 352-72.

Berners-Lee, Tim, James Hendler and Ora Lassila. 2001. "The Semantic Web; A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities." *Scientific American* 284, no. 5: 34-43. http://www.jstor.org/stable/26059207?seq=1&cid=pdf-reference#references_tab_contents

Berners-Lee, Tim, Nigel Shadbolt and Wendy Hall. 2006. "The Semantic Web revisited." *IEEE Intelligent Systems* 1541-1672/06 2006. https://eprints.soton.ac.uk/262614/2/OLD_Semantic_Web_Revisted.pdf

Borgman, Christine L. 2007. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: The MIT Press.

Borgman, Christine L. 2010. "Research Data: Who Will Share What, With Whom, When, and Why?" Unpublished paper, China-North America Library Conference, Beijing (2010). Retrieved from: https://works.bepress.com/borgman/238/

Borgman, Christine L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: The MIT Press.

Bowker, Geoffrey C. 2005. *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.

Boyd, Danah and Kate Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." In *Information, Communication & Society* 15: 662–79.

Briet, Suzanne. 1951. *Qu'est-ce que la Documentation?* Paris: EDIT.

Briet, Suzanne. 2006. *What is Documentation?* English translation of the classic French text. Lanham, MD: Scarecrow Press.

Buckland, Michael. 2018. "Document Theory." *Knowledge Organization* 45: 425-36.

Bugaje, Maryam and Gobinda Chowdhury. 2018. "Data Retrieval = Text Retrieval?" In *Transforming Digital Worlds. iConference 2018*, edited Gobinda Chowdhury, Julie McLeod, Val Gillet and Peter Willett. Lecture Notes in Computer Science 10766. Cham: Springer, 253-62.

Bureau International des Poids et Mesures (BIPM). 2006. The International System of Units (SI). 8th edition. Organisation Intergouvernementale de la Convention du Mètre. Paris: STEDI Media. https://www.bipm.org/utils/common/pdf/si_brochure_8.pdf

Candela, Leonardo, Donatella Castelli, Paolo Manghi and Alice Tani. 2015. "Data Journals: A Survey." *Journal of the Association for Information Science and Technology* 66: 1747-62. doi:10.1002/asi.23358

Caputo, John D. 2018. *Hermeneutics: Fact and Interpretation in the Age of Information*. London: Penguin.

Clarivate Analytics. 2018. The Repository Selection Process: Repository Evaluation, Selection, and Coverage Policies for the Data Citation Index within Clarivate Analytics Web of Science. https://clarivate.com/products/web-of-science/repository-selection-process

Cohen, Aaron, Michael, P. Zoë Starvi and William R. Hersh. 2004. "A Categorization and Analysis of the Criticisms of Evidence-Based Medicine." *International Journal of Medical Informatics* 73: 35–43.

Consultative Committee for Space Data Systems. 2002. *Reference Model for an Open Archival Information System (OAIS)*. Washington, DC: CCSDS.

"Data." 2018. *Wikipedia.* http://en.wikipedia.org/wiki/Data

"Data" and "Datum." *s.v., WordNet* 3.1 (2012). http://wordnetweb.princeton.edu/perl/webwn?s=data&sub=

Diebold, Francis. 2012. "A Personal Perspective on the Origin(s) and Development of 'Big Data': The Phenomenon, the Term, and the Discipline." In *PIER Working Paper, n° 13-003, University of Pennsylvania - Department of Economics; National Bureau of Economic Research (NBER), November 2012*. http://ssrn.com/abstract=2202843

Drucker, Johanna. 2011. "Humanities approaches to graphical display." *Digital Humanities Quarterly* 5, no. 1. http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html

Dye, Lee. 2007. "Do Libras Live Longer?" http://abcnews.go.com/Technology/Story?id=2890150&page=1

Edwards, Paul N., Matthew S. Mayernik and Archer L. Batcheller, Geoffrey C. Bowker and Christine L. Borgman. 2011. "Science Friction: Data, Metadata, and Collaboration." *Social Studies of Science* 41: 667–90.

Ekbia, Hamid, Michael Mattioli, Inna Kouper, G. Arave, All Ghazinejad, Timothy Bowman, Venkata Ratandeep Suri, Andrew Tsou, Scott Weingart and Cassidy R. Sugimoto. 2015. "Big Data, Bigger Dilemmas: A Critical Review." *Journal of the Association for Information Science and Technology* 66: 1523-45.

Ekstrøm, Jeanette and Lorna Wildgaard. 2019. *Theories and Methods in Data Science Librarianship*. London: Facet Publishing. Forthcoming.

Finnemann, Niels Ole. In Press. "Web Archive." In *ISKO Encyclopedia of Knowledge Organization*. http://www.isko.org/cyclo/web_archive

Fjordback Søndergaard, Trine, Jack Andersen and Birger Hjørland. 2003. "Documents and the Communication of Scientific and Scholarly Information. Revising and Updating the UNISIST Model." *Journal of Documentation* 59: 278-320.

Floridi, Luciano. 2008. "Data." In *International Encyclopedia of the Social Sciences*, ed. William A. Darity. New York: Macmillan Reference 234–7.

Floridi, Luciano. 2012. "Big Data and Their Epistemological Challenge." *Philosophy and Technology* 25: 435–7.

Fox, Christopher, Anany V. Levitin and Thomas C Redman. 1994. "The Notion of Data and Its Quality Dimensions." *Information Processing & Management* 30: 9–19.

Frické, Martin. 2009. "The Knowledge Pyramid: A Critique of the DIKW Hierarchy." *Journal of Information Science* 35: 131-42. doi:10.1177/0165551508094050.

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

Frické, Martin. 2018. "The Knowledge Pyramid: the DIKW Hierarchy." In *ISKO Encyclopedia of Knowledge Organization*: http://www.isko.org/cyclo/dikw

Furner, Jonathan. 2016. "'Data': The Data." In *Information Cultures In The Digital Age: A Festschrift in Honor of Raphael Capurro,* ed. Matthew Kelly and Jared Bielby. Wiesbaden: Springer, 287-306. http://www.jonathanfurner.info/wp-content/uploads/2016/12/Furner-Final-Proof-18.4.16.pdf

Furner, Jonathan. 2017. "Philosophy of data: Why?" *Education for Information* 33: 55-70. doi:10.3233/EFI-170986

Gitelman, Lisa, ed. 2013. *Raw Data' is an Oxymoron*. Cambridge, MA: MIT Press.

Gnoli, Claudio. 2014. "Boundaries and Overlaps of Disciplines in Bloch's Methodology of Historical Knowledge." In *Knowledge Organization in the 21st Century: Between Historical Patterns and Future Prospects: Proceedings of the Thirteenth International ISKO Conference 19-22 May 2014, Kraków, Poland*, ed. Wiesław Babik. Advances in Knowledge Organization 14. Würzburg: Ergon Verlag, 129-35.

Golub, Koraljka and Joacim Hansson. 2017. "(Big) Data in Library and Information Science: A Brief Overview of Some Important Problem Areas." *Journal of Universal Computer Science* 23 (Online), no. 11: 1098-108. http://lnu.diva-portal.org/smash/get/diva2:1194296/FULLTEXT01.pdf

Gray, Jim. 2009. "eScience: A Transformed Scientific Method." In *The Fourth Paradigm, Data-intensive Scientific Discovery*, ed. Tony Hey, Stewart Tansley and Kristin Tolle. Redmond, Wash.: Microsoft Research, 19-33. Second Printing, version 1.1 available at: http://itre.cis.upenn.edu/myl/JimGrayOnE-Science.pdf

Gray, Robert L. 2003. "Brief Historical Review of the Development of the Distinction Between Data and Information in the Information Systems Literature." In *9th Americas Conference on Information Systems*, hg. von. J. Ross und D. Galletta. Tampa: Association for Information Systems, 2843–9.

Guimarães, José Augusto Chaves. 2017. "Slanted Knowledge Organization as a New Ethical Perspective." In *The Organization of Knowledge: Caught Between Global Structures and Local Meaning*, ed. Jack Andersen, and Laura Skouvig. Studies in Information. Bingley, UK. Emerald Publishing Limited. doi:10.1108/S2055-537720170000011012

Gurr, Ted Robert. 1974. "The Neo-Alexandrians: A Review Essay on Data Handbooks in Political Science." *The American Political Science Review* 68: 243-52.

Haustein, Stefanie. 2016. "Grand Challenges in Altmetrics: Heterogeneity, Data Quality and Dependencies." *Scientometrics* 108: 413–23. doi:10.1007/s11192-016-1910-9.

Haynes, W. M., David R. Lide and Thomas J. Bruno, eds. 2017. *CRC Handbook of Chemistry and Physics: A Ready-reference Book of Chemical and Physical Data*. 97th ed. Boca Ratn, Florida: CRC Press.

Hey, Tony and Anne Trefethen. 2003. "The Data Deluge: An E-Science Perspective" *In Grid Computing: Making the Global Infrastructure a Reality*, ed. Fran Berman, Geoffrey Fox and Tony Hey. New York: Wiley, 809-24. https://eprints.soton.ac.uk/257648/1/The_Data_Deluge.pdf

Hine, Christine. 2006. "Databases as Scientific Instruments and Their Role in the Ordering of Scientific Work." *Social Studies of Science* 36: 269–98.

Hjørland, Birger. 1998. "Information Retrieval, Text Composition, and Semantics." *Knowledge Organization* 25: 16-31.

Hjørland, Birger. 2005. "Empiricism, Rationalism and Positivism in Library and Information Science." *Journal of Documentation* 61: 130-55. http://dx.doi.org/10.1108/00220410510578050

Hjørland, Birger. 2011. "Evidence Based Practice. An Analysis Based on the Philosophy of Science." *Journal of the American Society for Information Science and Technology* 62: 1301-10.

Hjørland, Birger. 2015. "Theories are Knowledge Organizing Systems (KOS)." *Knowledge Organization* 42: 113-28.

Hjørland, Birger. 2017. "Classification." *Knowledge Organization* 44: 97-128.

Hodge, Gail and Evelyn Frangakis. 2004. *Digital Preservation and Permanent Access to Scientific Information: The State of the Practice*. A Report Sponsored by The International Council for Scientific and Technical Information (ICSTI) and CENDI US Federal Information Managers Group. https://cendi.gov/publications/04-3dig_preserv.pdf

Ibekwe-SanJuan, Fidelia and Geoffrey C. Bowker. 2017. "Implications of Big Data for Knowledge Organization." *Knowledge Organization* 44: 187-98.

ISO 5127:2017. Information and documentation: Foundation and Vocabulary. 2nd ed.. Geneva: International Standards Organization.

Jensen, Howard E. 1950 "Editorial note." In *Through Values to Social Interpretation: Essays on Social Contexts, Actions, Types, and Prospects* by Howard Paul Becker. Durham, NC: Duke University Press, vii–xi.

Johansson, Veronica. 2012. "A Time and Place for Everything? Social Visualization Tools and Critical Literacies." PhD diss. University of Borås. https://www.diva-portal.org/smash/get/diva2:877028/FULLTEXT01.pdf

Kaase, Max. 2001. "Databases, Core: Political Science and Political Behavior." In *International Encyclopedia of the Social and Behavioral Sciences*, ed. Neil J. Smelser and Paul B. Baltes. Amsterdam: Elsevier, 5: 3251–5.

Kaden, Ben and Michael Kleineberg. 2018. *Guidelines for Publishing Dissertation-related Research Data*. Version 1.0. Berlin: Humboldt-Universität zu Berlin. doi:10.18452/19301

Keralis, Spencer D. C. 2012. CFP for Roundtable on Data Management for Humanities Research at MLA 2013. Blog post. *Spencer D. C. Keralis's Blog*. http://www.openaccessweek.org/profiles/blogs/cfp-for-roundtable-on-data-management-for-humanities-research-at

Kitchin, Rob. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. London: SAGE.

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

Lampland, Martha and Susan Leigh Star (Eds.). 2009. *Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*. Ithaca, NY: Cornell University Press.

Lancaster, Frederick Wilfrid. 1968a. *Information Retrieval Systems: Characteristics, Testing and Evolution*. New York: Wiley.

Lancaster, Frederick Wilfrid. 1986b. *Vocabulary Control for Information Retrieval*. 2nd ed. Arlington, Va.: Information Resources Press.

Laney, Douglas. 2001. "3-D Data Management: Controlling Data Volume, Velocity and Variety," *META Group Research Note*, February 6. http://goo.gl/Bo3GS

Laporte, Steven. 2018. "Ideal Language." *Knowledge Organization* 45: 586-608.

Leonelli, Sabina. 2010. "Packaging Small Facts for Re-Use: Databases in Model Organism Biology." In: *How Well Do Facts Travel? The Dissemination of Reliable Knowledge*, edited by Mary S. Morgan and Peter Howlett. Cambridge, UK: Cambridge University Press, 325–48.

Leonelli, Sabina. 2012. "Classificatory Theory in Data-intensive Science: The Case of Open Biomedical Ontologies." *International Studies in the Philosophy of Science* 26: 47–65.

Leonelli, Sabina. 2014. "What Difference does Quantity Make? On the Epistemology of Big Data in Biology." *Big Data & Society 1*: 1–11. doi:10.1177/2053951714534395

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago, IL: University of Chicago Press.

Lindberg, Christina A., comp. 2012. *Oxford American Writer's Thesaurus*. 3rd ed. New York: Oxford University Press.

Lyon, Aidan. 2016. "Data." In *The Oxford Handbook of Philosophy of Science*, ed. Paul Humphreys. Oxford, UK: Oxford University Press, 738-58.

Machlup, Fritz. 1984. "Semantic Quirks in Studies of Information." In *The Study of Information: Interdisciplinary Messages*, ed. Fritz Machlup and Una Mansfield. New York: Wiley, 641–71.

Mai, Jens-Erik. 2011. "The Modernity of Classification." *Journal of Documentation* 67: 710-30.

Maniez, Jacques. 1997. "Database Merging and the Compatibility of Indexing Languages." *Knowledge Organization* 24: 213-24.

Markham, Annette N. 2018. "Troubling the Concept of Data in Qualitative Digital Research." In *The SAGE Handbook of Qualitative Data Collection*, edi. Uwe Flick. London: SAGE Publications, 511–23.

Marx, Vivien. 2013. "Biology: The Big Challenges of Big Data." *Nature* 498: 255–60. http://www.nature.com/articles/498255a.pdf

Mazzocchi, Fulvio. 2018. "Knowledge Organization System (KOS): An Introductory Critical Account." *Knowledge Organization* 45: 54-78.

Meadows, Arthur Jack. 2001. *Understanding Information*. Munich: K. G. Saur.

Millerand, Florence and Geoffrey C. Bowker. 2009. "Metadata Standards: Trajectories and Enactment in the Life of an Ontology." In *Standards and Their Stories. How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*, ed. Martha Lampland and Susan Leigh Star. Ithaca, NY: Cornell University Press, 149–165.

Milroy, James. 2001. "Language Ideologies and the Consequence of Standardization." *Journal of Sociolinguistics* 5: 530-55.

Murray, Steve. 2017. "The LSST and Big Data Science: A New Kind of Telescope Will Need a New Kind of Astronomer." *Astronomy* [Magazine], December 15, 2017. http://www.astronomy.com/news/2017/12/the-lsst-and-big-data-science

National Research Council. 1999. Committee for a Study on Promoting Access to Scientific and Technical Data for the Public Interest, National Research Council. *A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases*. Washington, D.C: National Academy Press. https://www.nap.edu/read/9692/

National Science Board. 2005. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. Alexandria, VA: National Science Foundation. https://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

National Science Foundation. 2012. *Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA)*. NSF-12-499. Arlington, VA: National Science Foundation. https://www.nsf.gov/pubs/2012/nsf12499/nsf12499.pdf

National Science Foundation. 2014. *Critical Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA)*. NSF-14-543. Arlington, VA: National Science Foundation. https://www.nsf.gov/pubs/2014/nsf14543/nsf14543.pdf

Naur, Peter. 1966. "The Science of Datalogy." *Communications of the ACM* 9: 485.

Nielsen, Hans Jørn and Birger Hjørland. 2014. "Curating Research Data: The Potential Roles of Libraries and Information Professionals." *Journal of Documentation* 70: 221-40.

Owens, Trevor. 2011. "Defining Data for Humanists: Text, Artifact, Information or Evidence?." *Journal of Digital Humanities* 1: 6-8. http://journalofdigitalhumanities.org/1-1/defining-data-for-humanists-by-trevor-owens/.

Parsons, M. A. and P. A. Fox. 2013. "Is Data Publication the Right Metaphor?" *Data Science Journal* 12: WDS32-WDS46. doi:http://doi.org/10.2481/dsj.WDS-042

Pechenick, Eitan Adam, Christopher M. Danforth and Peter Sheridan Dodds. 2015. "Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution." *PLoS ONE* 10, no. 10: e0137041, 1-24. doi:10.1371/journal.pone.0137041

Peirce, Charles Sanders. 1955. "The Scientific Attitude and Fallibilism." In *Philosophical Writings of Peirce*, ed. Justus Buchler. New York: Dover.

Rafferty, Pauline M. 2018. "Tagging." *Knowledge Organization* 45: 500-16.

Ramanna, Karthik. 2015. *Political Standards: Corporate Interest, Ideology, and Leadership in the Shaping of Accounting Rules for the Market Economy*. Chicago: University of Chicago Press.

Rapley, Tim and Gethin Rees. 2018. "Collecting Documents as Data." *The SAGE Handbook of Qualitative Data Collection*, ed. Uwe Flick. London: SAGE Publications, 378-91.

Redman, Thomas C., Christopher Fox and Anany V. Levitin. 2017. "Data and Data Quality." *Encyclopedia of Library and Information Sciences*, 4th ed., ed. John D. McDonald and Michael Levine-Clark. Boca Raton, FL: CRC Press, 2: 1171-82.

Rider, Fremont. 1944. *The Scholar and the Future of the Research Library: A Problem and Its Solution*. New York: Hadham Press.

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

Roy, Suman Deb and Wenjun Zeng. 2015. *Social Multimedia Signals: A Signal Processing Approach to Social Network Phenomena*. Cham, Switzerland: Springer.

Saracevic, Tefko. 1999. "Information Science." *Journal of the American Society for Information Science* 50:1051-63.

Serafin, Robert J. et al. 1999. "A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases." Washington, DC: National Academy Press. https://www.nap.edu/read/9692/chapter/1

Shera, Jesse H. 1951. "Classification as the basis of bibliographic organization." In *Bibliographic Organization: Papers presented before the Fifteenth Annual Conference of the Graduate Library School July 24-29, 1950*, ed. Jesse H. Shera and Margaret E. Egan. Chicago: University of Chicago Press, 72-93.

Sheth, Amit. 1997. "Data Semantics: What, Where and How? In *Database Applications Semantics, Proceedings of the Sixth IFIP TC-2 Working Conference on Data Semantics (DS-6) Stone Mountain, Atlanta, Georgia U.S.A., May 30–June 2, 1995*, ed. Robert Meersman and Leo Mark. Cham: Springer. doi:10.1007/978-0-387-34913-8

Spang-Hanssen, Henning. 2001. "How to Teach about Information as Related to Documentation." Human IT 5, no. 1: 125-43. https://humanit.hb.se/article/view/168

Spärck Jones, Karen. 1987. "Architecture Problems in the Construction of Expert Systems for Document Retrieval." In *Knowledge Engineering: Expert Systems and Information Retrieval*, ed. Irene Wormell. London: Taylor Graham, 7–33.

Stuckenschmidt, Heiner. 2012. "Data Semantics on the Web." *Journal of Data Semantics* 1: 1–9. doi:10.1007/s13740-012-0003-z

Swanson, Don R. 1986a. "Fish Oil, Raynauds Syndrome, and Undiscovered Public Knowledge." *Perspectives in Biology and Medicine* 30: 7-18.

Swanson, Don R. 1986b. "Undiscovered Public Knowledge." *The Library Quarterly* 56: 103-18.

van Rijsbergen, Cornelis Joost. 1979. *Information Retrieval.*, 2nd ed. London: Butterworths.

van Rijsbergen, Cornelis Joost and Mounia Lalmas. 1996. "Information Calculus for Information Retrieval." *Journal of the American Society for Information Science* 47: 385-98.

Veltman, Frank. 1984. "Data Semantics." In: *Truth, interpretation, and information: selected papers from the third Amsterdam colloquium*, ed. Jeroen Groenendijk, Theo M. V. Janssen and Martin Stokhof. Berlin: Walter de Gruyter, 43–65.

Voss, Jakob. 2013. Describing Data Patterns. A General Deconstruction of Metadata Standards. PhD diss. Humboldt University Berlin. http://edoc.hu-berlin.de/dissertationen/voss-jakob-2013-05-31/PDF/voss.pdf

Wallis, Jullian C, Christine L. Borgman, Matthew S. Mayernik and Alberto Pepe. 2008. "Moving Archival Practices Upstream: An Exploration of the Life Cycle of Ecological Sensing Data in Collaborative Field Research." *International Journal of Digital Curation* 3, no. 1. Retrieved from http://www.ijdc.net/article/view/67 on June 14, 2018

Ward, Jonathan Stuart and Adam Barker. 2013. "Undefined by Data: A Survey of Big Data Definitions." *ArXiv e-print*, http://arxiv.org/abs/1309.5821

Weiss, Sholom and Nitin Indurkhya. 1997. *Predictive Data Mining: A Practical Guide*. San Francisco, CA: Morgan Kaufmann Publishers.

Zeng, Marcia Lei. In Press. "Interoperability." In *ISKO Encyclopedia of Knowledge Organization*. http://www.isko.org/cyclo/interoperability

Zhang, Sarah. 2015. "The Pitfalls of Using Google Ngram to Study Language." *Wired*. https://www.wired.com/2015/10/pitfalls-of-studying-language-with-google-ngram/

Zhao, Mengnan, Erjia Yan and Kai Li. 2018. "Data Set Mentions and Citations: A Content Analysis of Full-text Publications." *Journal of the Association for Information Science and Technology* 69: 32-46.

Zikopoulos, Paul, Chris Eaton, Dirk deRoss, Tom Deutsch and George Lapis. 2012. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. New York: McGraw Hill. https://www.immagic.com/eLibrary/ARCHIVES/EBOOKS/I111025E.pdf

Zins, Chaim. 2007. "Conceptual Approaches for Defining Data, Information, and Knowledge." *Journal of the American Society for Information Science and Technology* 58: 479-93.