Implications of Big Data for Knowledge Organization

Fidelia Ibekwe-SanJuan* and Geoffrey C. Bowker**

*Aix-Marseille University, School of Communication & Journalism (EJCAM),
21 rue Virgile Marron Marseille, France 13005,
<fidelia.ibekwe-sanjuan@univ-amu.fr>
** University of California Irvine, Donald Bren School of Information and Computer Sciences,
University of California, Irvine, 5019 Donald Bren Hall, Irvine, CA 92697-3440, USA,
<gbowker@uci.edu>

Fidelia Ibekwe-SanJuan is Professor at the School of Journalism and Communication, University of Aix-Marseille in France. Her research interests span both empirical and theoretical issues. She has developed methodologies and tools for text mining and information retrieval. She is interested in the epistemology of science, in inter-disciplinarity issues and in the history of information and library science. She was editor with Thomas Dousa of *Theories of Information, Communication and Knowledge. A Multidisciplinary approach*. She is currently investigating the impact of big data, open data and Web 2.0 on science and on the society.

Geoffrey C. Bowker is Professor at the School of Information and Computer Science, University of California at Irvine, where he directs the Evoke Laboratory, which explores new forms of knowledge expression. Recent positions include Professor of and Senior Scholar in Cyberscholarship at the University of Pittsburgh iSchool and Executive Director, Center for Science, Technology and Society, Santa Clara. Together with Leigh Star he wrote *Sorting Things Out: Classification and its Consequences*; his most recent books are *Memory Practices in the Sciences* and (with Stefan Timmermans, Adele Clarke and Ellen Balka) the edited collection, *Boundary Objects and Beyond: Working with Leigh Star*. He is currently working on big data policy and on scientific cyberinfrastructure; as well as completing a book on social readings of data and databases. He is a founding member of the Council for Big Data, Ethics and Society.

Ibekwe-SanJuan, Fidelia and Geoffrey C. Bowker. 2017. "Implications of Big Data for Knowledge Organization." *Knowledge Organization* 44(3): 187-198. 33 references.

Abstract: In this paper, we propose a high-level analysis of the implications of big data for knowledge organisation (KO) and knowledge organisation systems (KOSs). We confront the current debates within the KO community about the relevance of universal bibliographic classifications and the thesaurus in the web with the ongoing discussions about the epistemological and methodological assumptions underlying data-driven inquiry. In essence, big data will not remove the need for humanly-constructed KOSs. However, ongoing transformations in knowledge production processes entailed by big data and Web 2.0 put pressure on the KO community to rethink the standpoint from which KOSs are designed. Essentially, the field of KO needs to move from laying down the apodictic (that which we know for all time) to adapting to the new world of social and natural scientific knowledge by creating maximally flexible schemas—faceted rather than Aristotelean classifications. KO also needs to adapt to the changing nature of output in the social and natural sciences, to the extent that these in turn are being affected by the advent of big data. Theoretically, this entails a shift from purely universalist and normative top-down approaches to more descriptive bottom-up approaches that can be inclusive of diverse viewpoints. Methodologically, this means striking the right balance between two seemingly opposing modalities in designing KOSs: the necessity on the one hand to incorporate automated techniques and on the other, to solicit contributions from amateurs (crowdsourcing) via Web 2.0 platforms.

Received: 23 October 2016; Revised: 27 February 2016; Accepted: 9 March 2017

Keywords: big data, knowledge organization, KOSs, classification, information

1.0 Introduction

The field of knowledge organisation (KO) is mainly concerned with constructing and maintaining knowledge artefacts, otherwise known as knowledge organisation systems (KOSs) which are structured and controlled vocabularies such as bibliographic classification schemes, taxonomies and thesauri for libraries and organisations (Hodge 2000).

https://doi.org/10.5771/0943-7444-2017-3-187 https://www.inlibra.com/de/agb - Open Access - (cc) BY

Since the late twentieth century, mainstream KO research has focused largely on evolving bibliographic classification languages from the print and local computer media to the Internet and the World Wide Web and on bibliographic models and metadata for describing records. However, these systems have remained largely in the purview of libraries and librarians. The general public rarely has recourse to such systems to find things on the Internet, including books.

The late 1980s witnessed the advent of the Internet, the web and several advances in artificial intelligence (machine learning, natural language, text mining) which have since enabled the emergence of computer-supported techniques and tools for knowledge intensive tasks such as indexing, information retrieval, classification, ontologies and domain knowledge mapping. The KO community has been slow to integrate these automatic approaches into its methods. To the best of our knowledge, few KO publications (Smiraglia 2009; Ibekwe-SanJuan and SanJuan 2010, 2004 and 2002; Chen et al. 2008) have leveraged these automatic techniques, such as bibliometrics and data clustering to represent, organise or retrieve domain knowledge.

Over the last decade, the world has become aware of the clarion call of big data. It has become one of the buzz phrases of our times, after Web 2.0 and social media, which pepper the utterances and writings of commentators, journalists, businesses, policymakers and scientists. While not a technology in itself, big data has huge technological, methodological, social and epistemological implications. It has put every sector of activity under enormous pressure as organisations grapple with scalability issues in their information ecosystem. Recent advances in data processing techniques have given rise to very robust machine learning algorithms that are capable of leveraging the huge amounts of data left by our daily use of digital devices to build predictive models. The list of applications where big data algorithms are now being used is constantly growing, from the more conventional search and retrieval to recommender systems such as the ones on Amazon and Netflix that suggest what individuals may want to read next or watch next based on their previous clicks and purchases to targeted advertising to stock exchange brokering to trend detection, opinion mining and information visualisation. While not infallible, these algorithms have attained a level of performance that is acceptable to humans. Moreover, they are programmed to work in the background in a non-intrusive manner, quietly gathering data and crunching them to provide users with suggestions and recommendations that can rival those of a human librarian or KO specialist. Thus, big data algorithms raise the question of the relevance of humanly constructed KOSs and their capacity to keep up with the ever-increasing size of available data on specific topics and domains.

A lot has been written on the relative merits and pitfalls of big data for science and society but mainly from the point of view of other fields and disciplines. Big data is starting to become a topic of concern to KO scholars, as evidenced by recent discussions at the ASIST workshop of the "Special Interest Group on Classification Research" in 2014.[1] In that workshop, Shiri (2014, 18) argued that the formal and "organized nature of linked data and its specific application" for building SKOSs, linked controlled vocabularies and knowledge organization systems, should "have the potential to provide a solid semantic foundation for the classification, representation, visualization and the organized presentation of big data." However, Shiri's article is focused on practical applications of linked data to KOSs and it is also more programmatic rather than based on actual empirical evidence of the use of SKOSs in applications requiring the processing of large amounts of data. Indeed, most of the fields he listed as areas where KO research should have relevance such as natural language processing (NLP), machine learning, text mining, data mining, information visualisation, semantic search engines, recommender systems and query expansion have made significant advances without input from KO research.

To the best of our knowledge, few attempts have been made to offer high-level analyses of the implications of big data for KO research and for KOSs. Our paper does not deal with specific empirical studies nor with specific applications of KOSs to big data. Instead, we focus on the conceptual, epistemological and methodological implications that big data could have for KO and KOSs. More specifically, the questions we try to bring answers to are as follows:

1) How will the increasing dematerialisation of activities in every sector which leads to the huge increase in the amount of digital data available and in turn to the necessity to turn to algorithms to process and extract information from the data affect the design of KOSs?
2) In other words, will the era of "algorithmic governmentality"[2] (Rouvroy and Berns 2013) ushered in by the big data phenomenon signal the demise of KOSs in their current form?
3) How should KO scholars and practitioners position themselves with regard to the participatory paradigm of Web 2.0 which functions conjointly with big data on the field?

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.

In section 2.0, we will recall the ongoing debate within the KO community about the relevance of KOSs in the digital age. In section 3.0, we turn to the phenomenon of big data in order to highlight its characteristics and in section 4.0, underscore the conceptual, epistemological and methodological challenges it poses for KO research and practice. Finally, we will offer some concluding remarks in section 5.0 about the challenges facing KO in the near future.

2.0 On the relevance of bibliographic classification schemes and of the thesaurus in the digital and web era

Fears about the possible demise of an existing method or technique in the face of new ones are not new nor are they peculiar to the KO community. For instance, Almeida, Souza and Baracho (2015) issued a dire warning about the threats facing information science (IS) in the face of information explosion and other recent phenomena like big data, cloud computing and social networks. Instead of making IS stronger as a field, the pervasiveness of information, of digital data and of information technologies have in fact weakened IS considerably as most of its original subject matter has now become the object of research of fields like engineering, computer science, linguistics, sociology, anthropology or economics. The consequence, according to the authors is a much deflated IS which risks becoming "a mere niche among other fields," with information professionals becoming like "the 'remora,' which feeds on the thematic leftovers of topics that other fields develop." The authors concluded that the initial research program for IS laid out in Borko's 1968 seminal paper "Information science: what is this?," "has become too broad for the IS field" and indeed, is "too broad for any research field." They suggested turning to interdisciplinarity as a way of negotiating IS's relations with other fields on the overlapping subject matters.

Hjørland, a leading researcher of KO, has devoted several articles to analysing various issues in KO research and artefacts that may compromise the relevance of the field in the digital age. His criticisms revolve around three points which are of particular relevance to our discussion: 1) the possible obsolescence of universal bibliographic classification schemes; 2) the neglect of subject knowledge by library classifiers; and, 3) the reluctance of the KO community to leverage data analysis techniques as an alternative to manually constructed KOSs.

We will examine these criticisms in the light of the issues raised by big data for all sectors of activities dealing with digital artefacts of which KO is one.

2.1 Over-standardisation and over-normalisation of bibliographic classification schemes

In his 2012 article, Hjørland asked "Is Classification Necessary after Google?" While the title is provocative, it is also relevant. He made the observation that (299): "At the practical level, libraries are increasingly dispensing with classifying books" and "At the theoretical level, many researchers, managers, and users believe that the activity of 'classification' is not worth the effort, as search engines can be improved without the heavy cost of providing metadata." Search engines now offer access to full text of digital contents to end users, thus alleviating the need for lengthy library borrowing procedures. Also, the Google Books indexing project aims to digitise most of the human production of books. When this project is completed, it will challenge even more strongly the traditional role of libraries as primary custodians of knowledge artefacts, especially in print and book formats, as more publications migrate to the digital media. Concerning universal bibliographic classification schemes, (Hjørland 2012, 299) observed that the *Dewey Decimal Classification* (*DDC*) and Universal Decimal Classification (UDC) were built from the point of view of "standardisation" rather than "tailored to different domains and purposes" and that sections of the UDC are obsolete; thus, obsolete knowledge was being served in a flag-bearing product of KO. Aside from not being useful for online search and retrieval, their obsolescence has more profound implications; the most commonly used bibliographical classification schemes may not reflect the most current theories orienting research activities in some fields.

Furthermore, Hjørland (2015b) argued that KO should be concerned with theories of knowledge since theories are expressed on the linguistic level as concepts and concepts are the building blocks of KOSs. As observed by Hjørland, a classification is composed of statements of the sort that concept "A" is a kind of concept "B" or that concept "A" is related to concept "B." A classification can therefore be likened to a scientific theory, although we observe that it is of a much looser type with more limited implications and explanatory power than scientific theories. KOSs such as library classifications, thesauri and ontologies are therefore important auxiliaries of scientific theories because they reflect how concepts and objects in a domain are related to each other from the point of view of a given scientific theory which guided the classification task.

2.2 Neglect of subject knowledge

As argued by Hjørland (2013, 179), a corollary of the slowness to update universal bibliographic classification

https://doi.org/10.5771/0943-7444-2017-3-187 https://www.inlibra.com/de/agb - Open Access -

schemes is the neglect of subject knowledge from KO scholars and practitioners:

> My claim is that the neglect of the importance of subject knowledge has brought forward a crisis in KO, and that no real progress can be observed in the field. Of course, there is plenty of progress in the development of digital technologies which enable better kinds of knowledge representation and information retrieval. But such progress is brought to us from the outside; it is not something the field of KO has provided. It is important to realize that there is a need to make sure that the KOSs developed or studied within our field are sufficiently based on and related to current scientific theory (that is also the case with approaches based on numeric taxonomic methods). There is no short cut via user studies, common sense, or anything else.

Hjørland cites the example of the field of astronomy where the evolution of theory led to Pluto being demoted in 2006 from the status of a planet to a dwarf planet. He argued that one would have expected library classification schemes, taxonomies, and thesauri to reflect this "discovery" without much delay to ensure that people seeking information about planets are not served outdated or incorrect knowledge. Unlike library classification schemes which may take years to update, Wikipedia pages dedicated to planets and to Pluto updated the state of knowledge in this field as soon as the discovery was validated by the community of scholars in astronomy.

Because scientific theories which are the result of scientific discoveries are not immutable facts that are true at all times but can be overturned by other competing theories, it is important for KO to be concerned with scientific theories and domain knowledge. This is a fundamental issue, which can be illustrated by the attempts to build a semantic web tied to a fixed ontology. Ontologies are frequently changed by scientific disciplines as they grow, and categories have different ontological properties in related disciplines. The flexibility that is needed here cannot be generated within a universal subject classification, but this does not in the slightest obviate the need to classify—it just says something about the need for classifications to be flexible and adaptable. For example, in biology, a new fossil can uproot a classification system, which is not a problem if the change can propagate swiftly across multiple interrelated classifications (Bowker 2000).

While we agree with the soundness of Hjørland's fundamental criticisms, it is important to underline that the role of universal bibliographic classifications is not only to represent the state of domain knowledge at every given moment in time but also to organise knowledge artefacts in physical spaces like libraries such that their relationship with one another can be perceived. Furthermore, given the dynamic and evolving nature of digital data and the uncertainties underlying the knowledge contained therein (see section 3.0 hereafter for a discussion), universal bibliographic classifications cannot be expected to constantly change their classifications to follow every discovery made at each instant. This will not only prove an impossible task to accomplish in real time for libraries, but it can also be very disruptive for end users. There is, of necessity, a waiting period between a scientific discovery and its inclusion into universal bibliographic classifications that are known for portraying knowledge validated by the scientific community and which have acquired a certain degree of permanence. Also, the practical value of universal bibliographic classifications—that of enabling patrons to collocate material artefacts in a physical space, is not entirely dependent on the theoretical "up-to-dateness" of their class structure. Finally, universal classification schemes like the *DCC* and UDC which are the focus of Hjørland's criticisms form only a subset of KOSs. The other types—thesauri, ontologies and specialised classification schemes are all domain-dependent knowledge artefacts that make no claim to universalism and should therefore be amenable to more frequent updates.

2.3 The reluctance to leverage automatic data analysis and knowledge representation techniques to build more up-to-date KOSs

The same concerns about the relevance of KOSs in the digital age were perceptible at the fourth conference of the UK chapter of the International Society for Knowledge Organization (ISKO-UK) held on 14 July 2015 in London. The conference theme "Knowledge Organization—Making a Difference: The Impact of Knowledge Organization on Society, Scholarship and Progress" asked participants to "address the role that KO should have in the future, the opportunities that lie ahead for KO, and what difference it could really make for economic, scientific and/or cultural development."

In answer to that call, Soergel (2015) offered a somewhat mixed diagnosis. On the one hand, he asserted the pervasiveness of knowledge in all human endeavour which should logically ensure the necessity of KOSs in every domain and knowledge intensive applications. This is the optimistic viewpoint. On the other hand, he also acknowledged that many of the advances in automated techniques for knowledge extraction, representation and dissemination were brought about by other scientific communities. Large-scale ontologies, knowledge and expert systems, information search and retrieval platforms, taxonomies and semantic web technologies have been developed outside of the

KO community. The lack of interoperability of many KOSs only aggravates the situation. Even within the KO community, silos are erected around KOSs which slow down their integration with other knowledge repositories and their reuse outside of the specific KO targeted applications. He called for "more communication between the largely separated KO, ontology, data modelling, and semantic web communities to address the many problems that need better solutions" (401) and exhorted the KO community to recapture the terrain that it had abandoned to computer science by focusing not only on the classification of bibliographic metadata as it has done for centuries but to become involved in actually "structuring and representing the actual data or knowledge itself," issues that KO has "left to the ontology, artificial intelligence, and data modelling communities" (403).

This, he says, requires that the KO community should embrace computer applications such as information extraction, phrase sense disambiguation and information retrieval which can benefit from insights from KO. For KO to continue to be useful to today and tomorrow's world, it must be prepared to work with data analysts, and computer scientists amongst others.

Hjørland (2013) equally lamented the reluctance of the KO community to leverage automatic techniques to building KOSs. While both automatic data analysis techniques and manual approaches to designing of KOSs entail methodological and epistemological biases, automatic techniques represent a "bottom-up" approach to knowledge organisation. They can yield more descriptive domain knowledge organisation that reflects the current state of knowledge, rather than the prescriptive top-down approach to building KOSs. As such, they share some features with KOSs that rely on user-generated contents (UGC) such as folksonomies.

The debate initiated during the fourth ISKO-UK conference generated a lively discourse on the relevance of the thesaurus for online information retrieval and other knowledge intensive applications. Through both a face-to-face meeting[3] and scholarly publications gathered in a special issue of *Knowledge Organization* (volume 43, number 3, published in 2016), many authors defended the continued relevance of bespoke thesauri for several online applications such as domain knowledge representation, multilingual search and image retrieval. However, there was a consensus that the thesaurus has been displaced by general-purpose search engines, such as Google, as the standard tool for knowledge representation and search. The reasons given were similar to those that had dethroned the relevance of universal bibliographic classification schemes for document retrieval and knowledge representation: difficult-to-implement construction models, a tendency to over-standardisation and over-normalisation of semantic relations also known as "bundling" which reduces the diversity of possible semantic relations between domain concepts and objects to only "is-a" and "related-to," leading to tools that are inadequate for real-life situations. When confronted with building a knowledge representation scheme in enterprises and organisations, many information professionals admitted to not building "ISO standard compliant thesauri." According to their testimonies, "flexibility and pragmatism," rather than strict adherence to ISO guidelines, govern the endeavour. Several information professionals also stressed the necessity to seek ways to integrate user-generated contents (UGC) such as tags and looser synonyms into future KOSs. We will return to this point in section 4.4. In a blog post following the debate on the future of the thesaurus, (Dextre Clarke 2015) surmises that:

> Given a discerning team of developers, curators, IT support staff and indexers, this sophisticated tool can and should function interoperably alongside statistical algorithms, NLP techniques, data mining, clustering, latent semantic indexing, linked data, etc. Networking and collaboration, not rivalry, are the future.[4]

While such optimism is commendable, the operative words here are "can" and "should." Indeed, the expected interoperability and integration of KOSs into data mining and clustering techniques has not happened, despite all the common sense arguments advanced by KO practitioners and scholars. Perhaps one of the reasons lies in the fact that the two scientific communities have very little interaction with one another. More fundamentally, KOSs and indexing and clustering algorithms are designed from different epistemological and methodological assumptions. This makes their integration if not contradictory, at least difficult to achieve in practical terms. By their very nature, statistical and probabilistic models underlying indexing and clustering algorithms are designed to select data units based on their distribution, to model the behaviour of data units in a corpus and produce statistical tests and measures. Of course, there are cases of combined approaches integrating some humanly constructed knowledge bases into automatic systems but such architectures rarely scale up to industrial applications and would be intractable in the case of today's big data (Ibekwe-SanJuan and SanJuan 2002; 2010). Machine learning models are precisely designed to "learn" from existing data in order to be able to classify unseen units in the future. Furthermore, decades of experimental studies in information retrieval, NLP or semantic knowledge extraction and modelling tended to show that systems relying on consultation of external knowledge-rich databases during an online search slow down the retrieval speed without necessarily guaranteeing a significant increase in precision.

As a participant in the ISKO-UK debate concluded, "in my experience findability is the big driver rather than interoperability."[5] This is in line with Hjørland's assessment (2012, 301) of the challenge facing KO on the practical level, which he framed as follows: "how can LIS professionals contribute to the findability of documents, given the availability of many competing services in the 'information ecology?'"

It seems that researchers and practitioners agree on this point: if KO artefacts do not help people achieve the goal for which they were built, namely finding documents and information in our current web-centred information ecology, then they risk being relegated to the ash heap of history, replaced by technological solutions powered by NLP, statistical and machine learning algorithms.

Having recalled the ongoing debates about the relevance of KO research and of KOSs in the digital age, we now turn to the concept of big data in order to determine how they may in turn affect KO research and artefacts.

3.0 What is big data?

The first task that awaits anyone who embarks on a discourse on big data is to define what it is. Apart from the fuzziness surrounding the nature and size of big data, there has been some debate about the origins of the term itself. 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."[6] Another occurrence of the term was found in a 1998 computer science paper by Weiss and Indurkhya. 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.[7] Laney's 3 Vs later expanded into 4 Vs (3 Vs + Validity) and now has a fifth V as well (4 Vs + Veracity).

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. This led Gray (2009) to consider that data-driven science will be the "fourth science paradigm." However, people rarely venture to indicate a minimum size after which data can undisputedly be said to become big. At what point is one truly justified of speaking of "big data"? Boyd and Crawford (2012, 664) offered a characterisation of the different dimensions of big data:

> Big Data is less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets. We define Big Data 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.

As wearable electronic devices become more pervasive especially in the health and fitness, home and car insurance sectors, more and more data will be collected such that the term "big data" will eventually lose its distinctive meaning since most digital data will be "big." If, as Soergel (2015, 402) asserted (and we are in agreement), "Knowledge organization is needed everywhere, it is pervasive," then one will expect that big data, which has also become a pervasive phenomenon that embodies knowledge, will have an impact on KO research and artefacts.

The next section recasts the discussion of the epistemological assumptions underlying big data-driven inquiry in the light of the current concerns about the future of KO research and KOSs.

4.0 Possible implications of big data-driven inquiry for knowledge organisation research and artefacts

Broadly speaking, publications on big data seem to fall into three categories: 1) enthusiastic; 2) critical; and, 3) nuanced. For the first category of big data apostles, it represents the new *El Dorado* whose exploitation has the potential to accelerate the rhythm of scientific discoveries and innovations (Andersen 2008; Mayer-Schönberger and Cukier

2014; Gray 2009). Undeniably, the exploration of big data by sophisticated data exploration techniques has accelerated the rhythm of discoveries in some fields. The era of e-science it has brought in its wake implies a culture of international collaboration. Although Kitchin (2014, 10) is not a big data apologist,—he has indeed offered more nuanced and somewhat critical appraisals of the phenomenon—he nevertheless acknowledged its opportunities for scientific inquiry:

> There is little doubt that the development of Big Data and new data analytics offers the possibility of reframing the epistemology of science, social science and humanities, and such a reframing is already actively taking place across disciplines. Big Data and new data analytics enable new approaches to data generation and analyses to be implemented that make it possible to ask and answer questions in new ways.

For its critics, the power of big data is grossly overrated. The much mediatised errors of the Google Flu Trend algorithm in predicting the outbreaks and peaks of seasonal flu worldwide are often cited as a blatant case of big data algorithm failure. The refusal by Google analysts to release information on the exact datasets used to calculate such trends and the algorithmic processes involved only increased scholarly distrust (Auerbach 2014a, 2014b and 2014c; Marcus and Davis 2014; Thatcher 2014) as their study cannot be replicated by other scientists. Indeed, this is a general problem since big data algorithms used by private companies such as Facebook and Twitter tend to be proprietary and, thus, are not directly amenable to academic analysis.

The third category of more nuanced publications, recognises that big data-driven inquiry has the potential to accelerate the rhythm of discoveries in some fields[8] but at the same time that it has pitfalls of which we ought to be aware. Boyd and Crawford (2012, 664) summarised the duality of the big data phenomenon thus:

> Like other socio-technical phenomena, Big Data triggers both utopian and dystopian rhetoric. On one hand, Big Data is seen as a powerful tool to address various societal ills, offering the potential of new insights into areas as diverse as cancer research, terrorism, and climate change. On the other, Big Data is seen as a troubling manifestation of Big Brother, enabling invasions of privacy, decreased civil freedoms, and increased state and corporate control. As with all socio-technical phenomena, the currents of hope and fear often obscure the more nuanced and subtle shifts that are underway.

Boyd and Crawford were also amongst the first authors to frame high-level critical questions that we should be asking about the implications of big data driven inquiry for science. In the sections below, we will recall some of the ontological, epistemological and methodological implications of big data-driven inquiry which have been much debated in the big data literature and which may have implications for KO research and artefacts, given their current shortcomings discussed in section 2.0 above.

4.1 Data are social artefacts

Although data are often presented as a natural phenomenon just waiting to be collected, nothing could be farther from reality. Pushmann and Burgess (2014) discussed how the metaphors of "gold," "ocean," "torrent," "mineral" and "oil" are attached to the term big data, thus giving the impression that it is a natural phenomenon. As all data analysts know, data gathering is not a neutral nor an objective endeavour. It is governed by pragmatism (the goals of the study) and bound by technical constraints imposed by the data providers. This limits possibilities in terms data sources and content.

Ekbia et al. (2015, 1531) also offered a timely reminder of the whole gamut of tamperings involved in the data processing stage: from the intent to collect governed by pragmatic goals and involving "multiple social agents with potentially diverse interests" to its generation which is often "opaque and under-documented," to "incomplete or skewed" data without even talking of "instrument calibration and standards that guide the installation, development, and alignment of infrastructures" nor of human practices involving filtering (deciding which variables to keep and which to discard, in the case of personal data, anonymisation which often leads to loss of context and distortion), cleaning and even intentional distortions which all annihilate the pretensions to "rawness" of data."

This ensemble of tweaking makes data highly subjective and dependent on the aims of the project for which it is being collected. Data is therefore something that is constructed to suit a particular project and is by necessity always "incomplete." This point was aptly captured by Bowker (2013) when he wrote that "raw data is an oxymoron" and that "data should be cooked with care," an opinion echoed by the French sociologist, Bruno Latour (2014) who suggested that the French term for data "donnée" should be replaced by "obtenu" (obtained). Etymologically, "data" is the plural form of the latin word "datum" which means "that which is given?" (i.e., the perfect passive participle of the verb "do, dare," "to give"). Hence the suggestions by Bowker and Latour reflect the etymology of the word.

If all the data tweaking that takes place during data processing was not sufficient reason to adopt a critical attitude towards data-driven discoveries, consider the fact that big data studies from social media such as Facebook and Twitter are notoriously hard to replicate due to the restrictions in data gathering imposed by the private companies that have appropriated the data. Hence, the timely reminder by Boyd and Crawford (2012, 669) not to confuse data from social media with real people or that "people" and "Twitter users" are not synonymous, all the more so when a proportion of the data is generated by computer bots. A change in the dataset from which a study is conducted will also alter the "discoveries" made therein. Bowker (2014, 1797) similarly underscored the theoretical incompleteness of data:

> As Derrida (1998) argues in *Archive Fever* and Cory Knobel (2010) so beautifully develops with his concept of ontic occlusion, every act of admitting data into the archive is simultaneously an act of occluding other ways of being, other realities. The archive cannot in principle contain the world in small; its very finitude means that most slices of reality are not represented. The question for theory is what the forms of exclusion are and how we can generalize about them. Take the other Amazon as an illustration. If I am defined by my clicks and purchases and so forth, I get represented largely as a person with no qualities other than "consumer with tastes." However, creating a system that locks me into my tastes reduces me significantly. Individuals are not stable categories—things and people are not identical with themselves over time.

The foregoing observations underscore the transient and dynamic nature of big data which in turn render difficult if not impossible, the replicability of big data-driven studies. Yet, replicability is one of the canons of science. Thus, many studies reference the state of a database at the time the study was done but do not contain a copy of the database at that time. This holds *a fortiori* for studies of Twitter feeds where again one cannot access past states of the database.

What are the implications of the aforegoing considerations for KO? Traditionally, many KO practitioners and researchers building classification and indexing systems have justified the inclusion of terms and their relations based on literary warrant. This involves gathering data about the usage of these terms in books and other knowledge artefacts in a given field. How will literary warrant be construed given that the available size of data from which such warrants can be drawn has grown exponentially and will continue to do so, and also that the said data is constantly changing? If literary warrant is an important criterion for constructing KOSs, it means that KO practitioners and researchers will have to better account for how the corpora guaranteeing this literary warrant are built. The KO community will need to confront data representativity issues by documenting precisely how the data were collected, who or what is represented, who or what is left out, the types of processing the data underwent from its collection to the knowledge acquisition and representation. This will better inform end users about how the knowledge artefacts were built and what they can be used for. Doing so will also lend more credibility to those KO artefacts that claim universal subject coverage like the encyclopedic bibliographic classification schemes.

However, literary warrant is not the sole basis for warrants in building KOSs. Other types of warrants have been suggested such as "use warrant," "structural warrant," "educational warrant," "scientific/philosophical warrant," "semantic warrant" and "cultural warrant." For a typology of warrants, see Howarth and Jansen (2014).

If there exist other kinds of warrants that are not based on literary warrant, then the necessity for KOS construction to scale up to big data becomes less crucial. This is not to say that data representativity is unimportant nor should it be neglected. It means however that the credibility of KOSs is not uniquely linked to their adherence to literary warrant.

The essential argument here is that KO as a field needs to adapt to the changing nature of output in the social and natural sciences, to the extent that these in turn are being affected by the advent of big data. One model might be the high-energy physics community (itself closely linked with the rise of the web) where not only are data generated and shared in vast quantities in real time, but also literary warrant is created more rapidly than traditional publications through open archives like ArXiv.

4.2 Big data changes what it is to know

Epistemology is a philosophical account of what knowledge is and what knowing is. This is of particular import to the field of KO, a field which deals with the classification of existing knowledge accumulated over thousands of years of scientific inquiry. The sheer size of data and their dynamic and heterogeneous nature (e.g., image, text, sound) make it difficult to subject big-data driven inquiries to rigorous scientific verification. This could in turn result in being forced to abandon the principles of falsifiability and fallibilism of scientific theories laid down by Karl Popper and Charles S. Peirce which have guided scientific activity up till now—for some fields at least. If rapidly changing ontologies (characteristic of big data) are creating incommensurabilities on the fly, then we are moving into a more Kuhnian (1957) world in which old theories just cannot be compared with new—rendering falsifiability somewhat obsolete—especially until we create institutional mechanisms for preserving all states of a given database (a monumental and almost impossible task).

If as Hjørland (2015b) convincingly argued, KO should be concerned with theories, then knowledge derived from big data-driven discoveries should impact the theories and epistemological positions from which future KOSs are constructed. Essentially, the field of KO needs to move from laying down the apodictic (that which we know for all time) to adapting to the new world of social and natural scientific knowledge by creating maximally flexible schemas—that, faceted rather than Aristotelean classifications. The great knowledge organisation schemata of the nineteenth and twentieth centuries drew on a period when the sciences were being "invaded" by statistics (Hacking, Empire of Chance). This led social science (specifically Durkheim) to creating reified categories which in turn informed social policy. If the reified categories go away and "big data" replaces statistics (it is surely invading all branches of human learning just as effectively), then we need to rethink the nature of our enterprise.

4.3 Yes, classification is still necessary after Google!

In answer to his provocative title "Is Classification still Necessary after Google?," Hjørland (2012) concluded in the affirmative, despite advances in machine mediated information retrieval and Google, because any classification is a choice between different viewpoints. We are in agreement with this analysis. Big data and Google have not removed the need for classification; quite the contrary! The ever-increasing amount of data means we need classification more than ever, but the nature of that classification needs rethinking. There is a simple logic to this: big data only works if it comes with good metadata. Each form of metadata in turn relies on fixed categories of one kind or another. It is not that classifications will go away, rather, they just become less visible. Thus, even in the world of big data, it may be very difficult to parse backwards a retronym such as "biological mother" since the "mother" category may be part of the data flow—where "surrogate mother" or "adoptive mother" may not.

4.4 Big data paradox: between automation and human labour

Whereas the big data phenomenon entails increased automation of tasks in many fields, data-driven algorithms also require constant human input to learn and improve their models and hence performances. We therefore have a paradoxical situation where big data leads ultimately to the replacement of humans by algorithms whilst at the same time requiring human labour (crowdsourcing) to improve the predictive powers of the said algorithms. Ekbia et al. (2015) called this paradox "heteromation" whereby big data relies on the co-existence of two seemingly opposing modalities: human labour and automation.

Initially, there was a lot of scepticism about the quality and value of large-scale knowledge resources built through the crowdsourcing model but the success of the Wikipedia project in harnessing public participation to make it the most consulted and cited web site, as well as the proliferation of participatory science projects have silenced even the most vocal sceptics.

To cite only a few examples in the field of astronomy, the SDSS project gave rise to *Galaxy Zoo,*[9] a crowdsourcing project to identify and annotate 3D images of celestial objects taken by the SDSS telescope. Started in 2007, *Galaxy Zoo* has mobilised more than 150,000 amateur contributors who helped astronomers classify more than 230 million celestial objects. Likewise, the *eBird*[10] project led by the Ornithology Lab at Cornell University crowdsourced the immense task of inventorising of bird species. This enabled scientists to collect up to 160 millions observations from more than 1,000 bird watchers all over the world, thus accounting for more than 95% of bird species. Lagoze (2014) acknowledged that it was "a highly successful citizen science project that for over a decade has collected observations from volunteer participants worldwide. Those data have subsequently been used for a large body of highly-regarded and influential scientific research." This would have been impossible for scientists alone given the scale of the task. Web 2.0 technologies have made harnessing the contributions of the masses possible at an unprecedented scale. As scientists and other professionals in giant tech companies rely more and more on machines and on volunteers, the existence of online communities of citizen scientists can lead to a blurring of frontiers between amateurs and specialists. By inviting large members of the public to partake in the scientific adventure, scientists ultimately relinquish some of their prerogatives and areas of past expertise. Lagoze (2014) called this a "fracturing of the control zone." This makes many scholars and professionals understandably uneasy because it breaks down well-established barriers between experts and amateurs.

In the library, archives and museum realm, there are efforts to integrate the participatory model of knowledge production popularised by Web 2.0 and which the general public has come to expect. Concepts like "museum 2.0," "participatory museum" (Simon 2010) and "library 2.0" or "participatory libraries" have come to represent endeavours to leverage public participation and integrate UGC (e.g., folksonomies) into some of the KOSs artefacts[11]—al-

https://doi.org/10.5771/0943-7444-2017-3-187 https://www.inlibra.com/de/agb - Open Access -

though this practice has found little traction in academic libraries (see Ibekwe-SanJuan and Ménard 2015 for a review).

Some information professionals at the thesaurus debate organised by ISKO-UK in 2015 have emphasised the need to liberate thesaurus construction from the shackles of top-down ISO normalisation rules and to integrate more UGC and more uncontrolled vocabulary terms.

Making a similar argument, Hjørland (2012, 308) writes:

> While mainstream classification research is still based on the objectivist understanding (a document has a subject), the minority view (that document A is assigned subject X by somebody in order to support some specific activities) is gaining a footing. I believe this last view is decisive for making a future for classification in both theory and practice.

This will enable KOSs to reflect more up-to-date knowledge which will better serve the needs of specific applications and categories of users. It will also ensure that KOSs are integrating more diverse viewpoints through the implementation of recommender systems available on Web 2.0 platforms. Domain experts are more aware of term usages and of recent advances in their specific fields than cataloguers.

5.0 Conclusion

In this paper, we confronted the ongoing debates about the future of KO research with debates about the benefits and pitfalls of big data for scientific inquiry in order to determine how the latter might affect the former. In an era of big data, it appears even more unrealistic to hope that universal bibliographic classification schemes can be updated by a handful of "expert cataloguers or bibliographers" nor to ignore the participatory and collaborative paradigm which has made Web 2.0 platforms like Wikipedia, Facebook and Twitter successful. However, universal bibliographical classifications in libraries are typically not used alone but are integrated into broader systems (i.e., catalogs), which bring them together with subject heading systems, thesauri and sometimes folksonomies. Public participation can be better leveraged to update subject headings and thus enhance the effectiveness of library classification schemes. Thus, it is not a matter of "either...or," i.e., either expert-built classification systems or participatory/collaborative systems, but "and ... and," i.e., determining how both approaches can be combined in designing KOSs for specific applications and categories of users.

The challenge for KO is therefore to reinvent itself in an information ecosystem filled with algorithms that are continuously crunching data and delivering digital content tailored to users' profiles rather than focusing on one-size-fits-all knowledge bases constructed *a priori*. This calls for a "rapprochement" between the KO and the computer and artificial intelligence communities as well as a significant opening up of library and information science curricula to integrate subjects like epistemology, philosophy, statistics and data analysis techniques. Knowledge organisation will not go away as a field; it is central to the scientific endeavour. However, it needs to adapt to the new temporalities of theoretical development occasioned by the spread of big data across the social and natural sciences.

Notes

1. See https://journals.lib.washington.edu/index.php/acro/issue/view/1014
2. Term coined by Rouvrot Antoinette to refer to the reign of big data algorithms which now make most decisions for humans, from what we ought to read and buy, to which stocks we ought to invest in and to smaking cientific discoveries which we cannot account for because data-driven discoveries lack causality dimension. Algorithms use our digital traces (our personal data) to calculate our "digital selves" and serve us desires before we are even aware of having them.
3. See ISKO-UK event, accessed on 11th August 2016. Accessible at http://www.iskouk.org/content/great-debate
4. Dextre Clarke, Blogpost on "The Thesaurus Debate needs to move on." 27 February 2015. http://iskouk.blogspot.co.uk/2015/02/thesaurus-debate-needs-to-move-on.html
5. Notes taken on the thesaurus debate by a participant. http://www.iskouk.org/content/great-debate
6. 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
7. According to Diebold (2012), "META is now part of Gartner."
8. In particle physics, the discoveries of the Higgs Boson in 2012 and of the pentaquarks in 2015 are among some of the most significant recent scientific discoveries which would not have been possible without the Large Hadron Collider (http://home.cern/topics/large-hadron-collider) which generates massive data for physicists to analyse.
9. http://www.galaxyzoo.org/
10. http://ebird.org/content/ebird/
11. See for instance some of the papers in the bibliography of Jennifer Trant: http://www.archimuse.com/consulting/trant_pub.html but also museomix initiatives in different countries

References

Almeida, M.B., Renato Souza and R. M. A. Baracho. 2015. "Looking for the Identity of Information Science in the Age of Big Data, Computing Clouds and Social Networks." In *Re:inventing Information Science in the Networked Society: Proceedings of the 14th International Symposium on Information Science (ISI 2015), Zadar, Croatia, 19- 21 May 2015,* ed. Pehar Franjo. Glückstadt: Hülsbusch.

Auerbach, David. 2014a. "The Mystery of the Exploding Tongue: How reliable is Google Flu Trends?" *Slate.com* 2014 no. 3. http://www.slate.com/articles/technology/bitwise/2014/03/google_flu_trends_reliability_a_new_study_questions_its_methods.html

Auerbach, David. 2014b. "Big Data is Overrated and OkCupid's User Experiments Prove It." *Slate.com* 2014 no. 7. http://www.slate.com/articles/technology/bitwise/2014/07/facebook_okcupid_user_experiments_ethics_aside_they_show_us_the_limitations.html

Auerbach, David. 2014c. "The Big Data Paradox: It's Never Complete, and It's Always Messy—And If It's Not, You Can't Trust It." *Slate.com* 2014 no. 8. http://www.slate.com/articles/technology/bitwise/2014/08/what_is_big_data_good_for_incremental_change_not_big_paradigm_shifts.html

Bowker, Geoffrey. 2014. "The Theory/Data Thing: Commentary." *International Journal of Communication* 8: 1795-9.

Bowker, Geoffrey. 2013. "Data Flakes: An Afterword to Raw Data Is an Oxymoron." In *Raw Data Is an Oxymoron,* ed. Lisa Gitelman. Cambridge, MA: MIT Press, 167-71.

Boyd, Danah and Kate Crawford. 2012. "Critical Questions for Big Data." *Information, Communication & Society* 15: 662-79.

Chen, Chaomei and Fidelia Ibekwe-SanJuan, Roberto Pinho and James Zhang. 2008. "The Impact of the Sloan Digital Sky Survey on Astronomical Research the Role of Culture, Identity, and International Collaboration." In *Culture and Identity in Knowledge Organization: Proceedings of the Tenth International ISKO Conference, Montréal, Canada, August 5-8, 2008,* ed. Clément Arsenault and Joseph T. Tennis. Advances in Knowledge Organization 11. Würzburg: Ergon Verlag, 307-12.

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

Dextre Clarke, Stella. 2015. "Thesaurus Debate Needs to Move on." http://iskouk.blogspot.co.uk/2015/02/thesaurus-debate-needs-to-move-on.html

Ekbia, Hamid, Michael Mattioli, Inna Kouper, G. Arave, Ali 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.

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.

Hjørland, Birger. 2015a. "Are Relations in Thesauri Context-Free, Definitional, and True in All Possible Worlds?" *Journal of the Association for Information Science and Technology* 66: 1367-73.

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

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

Hjørland, Birger. 2012. "Is Classification Necessary after Google?" *Journal of Documentation* 68: 299-317.

Howarth, Lynne C. and Eva Hourihan Jansen. "Towards a Typology of Warrant for 21st Century Knowledge Organization Systems." In *Knowledge Organization in the 21st Century: Between Historical Patterns and Future Prospects: Proceedings of the Thirteenth International ISKO Conference, 19-22 May 2014, Krakow, Poland,* ed. Wiesław Babik. Advances in Knowledge Organization 14. Würzburg: Ergon Verlag, 216-21.

Ibekwe-SanJuan, Fidelia and Elaine Ménard, eds. 2015. *Archives, Libraries and Museums in the Era of the Participatory Social Web.* Special issue of *Canadian Journal of Information and Library Science* 39, nos. 3-4.

Ibekwe-sanjuan, Fidelia and Eric Sanjuan. 2010. "Knowledge Organization research in the Last Two Decades: 1988-2008." In *Paradigms and Conceptual Systems in Knowledge Organization: Proceedings of the Eighth International ISKO Conference Rome, 23-26 February 2010,* ed. Claudio Gnoli and Fulvio Mazzocchi. Advances in Knowledge Organization 12. Würzburg: Ergon Verlag, 115-21.

Ibekwe-SanJuan, Fidelia and Eric SanJuan. 2004. "Mining for Knowledge Chunks in a Terminology Network." In *Knowledge Organization and the Global Information Society: Proceedings of the Eighth International ISKO Conference, London, England, July 13-16, 2004,* ed. Ia C. McIlwaine. Advances in Knowledge Organization 9. Würzburg: Ergon Verlag, 41-7.

Ibekwe-SanJuan, Fidelia and Eric SanJuan. 2002. "From Term Variants to Research Topics." *Knowledge Organization* 29: 181-97.

Kitchin, Rob. 2014. "Big Data, New Epistemologies and Paradigm Shifts." *Big Data & Society* 1, no. 1: 1-12.

Lagoze Carl. 2014. "Big Data, Data Integrity, and the Fracturing of the Control Zone." *Big Data & Society* 1, no.2: 1-11.

Latour, Bruno, 2014. "Le mode d'existence du Politique." Lecture *Colloquium IXXI "La révolution numérique et la gouvernance, ENS Lyon, France, 4 April 2014*.

Marcus, Gary and Ernest Davis. 2014. "Eight (No, Nine!) Problems with Big Data." *New York Times*, April 6, 2014. https://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html

Mayer-Schönberger, Viktor and Kenneth Cukier. 2014. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Mariner Books.

Pushmann, Cornelius and Jean Burgess. 2014. "Metaphors of Big Data." *International Journal of Communication* 8: 1690-709.

Rouvroy, Antoinette and Thomas Berns. 2013 "Gouvernementalité algorithmique et perspectives d'émancipation." *Réseaux* no. 177: 163-96. doi:10.3917/res.177.0163

Shiri, Ali. 2014. "Linked Data Meets Big Data: A Knowledge Organization Systems Perspective." *Advances in Classification Research Online* 24: 16-20. doi:10.7152/acro.v24i1.14672

Simon, Nina. 2010. "*The Participatory Museum.*" Santa Cruz, Calif.: Museum 2.0.

Smiraglia, Richard P. 2009. "Modulation and Specialization in North American Knowledge Organization: Visualizing Pioneers." In *Pioneering North American Contributions to Knowledge Organization: Proceedings of the 2nd North American Symposium on Knowledge Organization, Syracuse, NY, USA, June 18-19, 2009,* eds. Elin K. Jacob and Barbara Kwasnik. Arizona: University of Arizona Library, 35-46. http://hdl.handle.net/10150/105092

Soergel, Dagobert. 2015. "Unleashing the Power of Data through Organization: Structure and Connections for Meaning, Learning and Discovery." *Knowledge Organization* 42: 401-27.

Thatcher, Jim. 2014. "Living on Fumes: Digital Footprints, Data Fumes, and the Limitations of Spatial Big Data." *International Journal of Communication* 8: 1765-83.