Making Sense of Big Data: A Facet Analysis Approach

Ali Shiri

School of Library and Information Studies, University of Alberta,
Edmonton, Alberta, T6G 2J4, Canada, <ashiri@ualberta.ca>

Ali Shiri is a professor at the School of Library and Information Studies in the University of Alberta, Canada. He teaches in the areas of digital libraries and digital information organization and retrieval. His research areas include user interaction with digital information, digital libraries, search user interfaces, knowledge organization systems, social media and big data. His recent book titled "Powering Search: The Role of Thesauri in New Information Environments" addresses the role and importance of thesauri in developing semantically rich search user interfaces for digital information repositories.

Shiri, Ali. **Making Sense of Big Data: A Facet Analysis Approach.** *Knowledge Organization. 41(5)*, 357-368. 29 references.

Abstract: Understanding, exploring and investigating big data to inform the development of policies and best practices requires a solid analysis, identification and mapping of the key facets and aspects of big data. The objective of this paper is two-fold: a) to provide a facet analysis of big data key topics and issues; and, b) to present a select number of information science research methodologies and study frameworks that may have the potential to be applied to research on big data. Six facets, namely data type, environment, people, operations and activities, analytics, and metadata are introduced to capture the key aspects of big data. Furthermore, sub-facets are created for each facet to demonstrate specific aspects that constitute the key topics. This type of conceptualization of big data will contribute to our learning and understanding of big data and its key components and characteristics. A number of suitable methodological frameworks from information science are introduced along with their potential applications for big data.

Received: 13 November 2013; Revised: 16 July 2014; Accepted 16 July 2014

Keywords: Big Data, facet analysis, research, metadata

1.0 Introduction

The vast volume, variety and complexity of digital data available on the web has resulted in the emergence of what is called "big data." De Witt et al. (2012) note that:

> Facebook uploads three billion photos monthly for a total of 3,600 terabytes annually. The data are generated by a lot of humans, but each is limited in their rate of data production. In the 10 years to 2008, the largest current astronomical catalogue, the Sloane Digital Sky Survey, produced 25 terabytes of data from telescopes. By 2014, it is anticipated that the Large Synoptic Survey Telescope will produce 20 terabytes each night. By the year 2019, The Square Kilometre Array radio telescope is planned to produce 50 terabytes of processed data per day, from a raw data rate of 7000 terabytes per second.

Social media sites, search engines, cloud-based computing infrastructures as well as virtual collaboratories, e-science, e-humanities and e-social sciences projects produce massive volumes of data that call for proper management and preservation-planning approaches and strategies in order to provide users with effective and efficient access.

There are many different terms used in the literature that may refer to or be associated with the phenomenon of "big data," including such terms as research data, digital data, linked data, open data, web of data and data repositories. The availability and discourse of these data types presents new research, development and policy opportunities as well as challenges. Domains and disciplines within natural sciences, social sciences and humanities can leverage the power of big data to create new research initiatives and avenues and to inform the development of policies, practices, systems and services.

The objective of this paper is twofold. The first objective is to present a faceto-analytical perspective of big data. In particular, the paper presents a categorization of topics and issues important for the understanding, analysis, learning, teaching, research and policy development for big data. The second objective is to draw upon research methodologies and analytical frameworks developed in information science to briefly provide new ways of analyzing and making sense of big data. The main argument in this paper is that information science in general and knowledge organization methods in particular can provide a solid basis for the understanding and the study of big data. The first part of this paper provides a delineated view of big data using facet analysis, which is a well-established knowledge organization method. The second part of this paper argues that there is a broad array of information science research methodologies and approaches that have particular and advantageous applications for studying and making sense of big data.

Recent discussions and studies of big data have focused on individual big data initiatives and projects. The variety of terminology used to refer to the phenomenon of big data warrants the development of a typology of various facets and types of big data. This kind of typology may serve as a basis for the conceptualization of big data in the context of research, development and teaching activities. Furthermore, it has the potential to provide a theoretical and terminological framework that could be used to investigate the various facets and aspects of big data in different contexts, environments and disciplines.

2.0 Context and definitions

Facet analysis as a knowledge organization and analysis technique was first introduced by Ranganathan (1967). Hjørland (2013) has recently provided a historical and logical examination of the facet analysis theory and notes that the "facet-analytic paradigm is probably the most distinct approach to knowledge organization within library and information science, and in many ways it has dominated what has been termed modern classification theory." Foskett (2009, 1819) notes that a facet may consist of entity terms, such as elements in chemistry, or crops in agriculture; forms of entities, such as solid, liquid, gas; operations made on entities, such as combustion, forging, harvesting; tools for operations, such as presses, X-rays for therapy, microscopes; states of being, such as health and disease. He also argues that the use of the term "analysis" versus the term "division" "has a wider connotation and may be applied to the study of complexes as well as to the entities." This technique has been widely used in the development of various knowledge organization systems, including classification systems, thesauri, taxonomies, as well as in the development of website architectures and visual and navigational information structures. The notion of web facet has been proposed to provide a meaningful approach to the presentation and categorization of search engine results (Milonas 2011). Facets and faceted classification seem to be among the critical thematic areas that North American Knowledge Organization (NASKO) researchers have studied (Smiraglia 2009). La Barre (2010) provides a comprehensive review of the facet analysis theory and its historical and developmental stages, providing various recent applications such as databases, retrieval systems, interfaces, faceted metadata, faceted data modeling, and faceted search and browsing systems. A number of studies have made use of facet analysis as a way of delineating the various characteristics, attributes and aspects of complex, compound and multi-faceted topics. For instance, interactive information retrieval research has made use of the facet analysis technique for the analysis and proper understanding of such complex concepts as "task" in the information seeking and retrieval process (Li and Belkin 2008) and the concept of query in interactive information retrieval (Shiri 2013). In this paper, the goal was to benefit from facet analysis as an approach to the analysis of the concept of big data and how it is emerging and evolving as a subject area.

A number of definitions have been proposed for big data in the literature. Because of the multifaceted nature of this phenomenon, scientists, information technology managers, information scientists, policy makers and funding agencies have approached it from various perspectives. This is, in part, due to the vague nature of the term big data and what it means to people from various educational and occupational backgrounds. For instance, the National Science Foundation report on *Long-lived Digital Data Collections* (2005) avoids using the term "big." Rather it focuses on the longevity and proper management of digital data. The report defines digital data as follows:

> The term "data" is used in this report to refer to any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc. Such data may be generated by various means including observation, computation, or experiment.

This definition has a clear focus on demonstrating the vast variety of data, its origins and the associated techniques for its analysis and maintenance. A more technologically and industrially focused definition is offered by Kusnetzky (2011) who defines big data as follows: "In simplest terms, the phrase [big data] refers to the tools,

processes and procedures allowing an organization to create, manipulate, and manage very large data sets and storage facilities." This definition takes a more pragmatic approach to big data and places emphasis on the volume of data and the challenge of its technical management.

Jacobs (2009, 40) approaches big data from a database technology perspective and notes that the fact that most large datasets have inherent temporal or spatial dimensions, or both, is crucial to understanding one important way that big data can cause performance problems, especially when databases are involved. His meta-definition of big data stresses the significance of temporal data as a key factor and believes that big data should be defined at any point in time as "data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time." In today's world, it may mean that data is too large to be placed in a relational database and analyzed with the help of a desktop statistics/visualization package—data, perhaps, whose analysis requires massively parallel software running on tens, hundreds, or even thousands of servers. Dumbill (2013, 1) provides a more recent definition for big data: "Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it."

In line with technological approaches to big data, Warden (2011) provides a particularly useful glossary of big data that provides a listing and description of 60 most recent technological innovations in the area of big data that can help those working with large data sets navigate the large number of new data tools available. These technologies vary from noSQL databases, MapReduce, storage and servers to natural language processing, machine learning, acquisition and visualization.

In the context of the sciences, Borgman (2007) makes use of the term "data deluge" and refers to the variety of data created, ranging from laboratory and field notebooks, slides from talks and composite objects to graphic visualizations of data. Examples of data in the science may include X-rays, protein structures, spectral surveys, specimens and events and objects. She argues that it is difficult to separate data from software, equipment, documentation and knowledge required to use them. This observation points to the challenges of defining data and data carriers. Borgman (2007) also provides a categorization for the types of data created by social scientists. The first category is data collected by researchers through experiments, interviews, surveys, observations. The second type is data that is collected by other people or institutions usually for purposes other than research. These include government and institutional data such as census figures, economic indicators, demographics and other public records. Other data sources such as mass media content and records of corporations, she notes, can be useful sources of social science data. She suggests that in the area of humanities the distinction between documents and data is the least clear due to the fact that almost any document, physical artifact and any record of human activity can be used to study culture. Further, Borgman (2012) discusses the approaches to handling data and notes that data collection can be viewed from various perspectives, including observatory vs. exploratory, empirical vs. theoretical, describing phenomena vs. modeling systems, data collection by hand vs. by machine, collaborative vs. individual data collection.

Bizer et al. (2011) argue for the meaningful and semantic use and applications of big data by providing four challenges, namely: a) the fact that big data integration is multidisciplinary; b) web of data and structured data as part of big data faces processing and integration challenges; c) lack of good use cases to provide the opportunity for experimenting with open linked data on the Web; and, d) demonstrating the value of semantics in data linking and integration.

The idea behind the Digging into Data Challenge, an international grant competition:

> Was to address how 'big data' changes the research landscape for the humanities and social sciences. Now that we have massive databases of materials used by scholars in the humanities and social sciences—ranging from digitized books, newspapers, and music to transactional data like web searches, sensor data or cell phone records—what new, computationally-based research methods might we apply? As the world becomes increasingly digital, new techniques will be needed to search, analyze, and understand these everyday materials. 'Digging into Data' initiative challenges the research community to help create the new research infrastructure for 21st century scholarship.

Hodson (2012), the Research Manager for JISC Digital infrastructure names a number of areas that deal with the big data issue, namely web archiving, learning analytics, usage statistics and research data. In line with big data in the context of social sciences, JISC has sponsored a project to be conducted by the Oxford Internet Institute titled Big Data: Demonstrating the Value of the UK Web Domain Dataset for Social Science Research, which aims to enhance JISC's UK Web Domain archive, a 30 terabyte archive of the .uk country-code top level domain collected from 1996 to 2010. It will extract link graphs from the data and disseminate social science research using the collection. In his final remarks for the Eduserv Symposium 2012: Big Data, Big

Deal? held in London, UK in May 2012, Powell (2012) suggests that there seems to be confusion about open data and big data and that there is a potential confusion between big data and data that happens to be big. He notes that open data is considered to be big data. He also suggests that we need to think carefully about the kinds of questions we need to ask when deal with big data.

The National Science Foundation and the National Institutes of Health's *Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA)* solicitation states the aim of big data as:

> To advance the core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed and heterogeneous data sets so as to: accelerate the progress of scientific discovery and innovation; lead to new fields of inquiry that would not otherwise be possible; encourage the development of new data analytic tools and algorithms; facilitate scalable, accessible, and sustainable data infrastructure; increase understanding of human and social processes and interactions; and promote economic growth and improved health and quality of life.

These two organizations emphasize that big data does not exclusively refer to the volume of the data, but also to its variety and velocity. They note that: "Big data includes 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." Davenport et al. (2013) approaches the notion of big data from the perspective of business processes and lists three ways in which the organizations that capitalize on big data differ from traditional data analysis environments, namely:

- They pay attention to data flows as opposed to stocks.
- They rely on data scientists and product and process developers rather than data analysts.
- They are moving analytics away from the IT function and into core business, operational and production functions.

Wu et al. (2014) propose a theorem to model big data characteristics called HACE. The HACE big data model starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data.

A review of definitions and characteristics of big data demonstrates the complexity and variety of concepts and terms used to identify what constitutes big data. One could argue that research data, open data, linked data and semantic web data can be construed as part of big data. These terms refer to the growing volume of different types of structured and unstructured data, their complex and heterogeneous nature and machine-processability and the challenges they pose for creators and users of big data. The organization, curation, exploration, management, preservation, visualization and access to and use of these types of data pose similar technological and computational challenges.

A succinct analysis of the definitions provided above illustrates the different characteristics and properties of big data as presented below:

- Very large integrated and linked data sets
- Variety of data and its typology
- Storage facilities
- Processing capacity
- Temporal and spatial dimensionality
- Heterogeneous, diverse, distributed, complex, evolving nature
- Analytical and visualization tools, technologies and models
- Semantic vagueness and confusion around big data

As can be inferred from these characteristics, one can note the reason behind the fact that many different disciplines and subject domains are interested in and have started conducting research in the area of big data.

The review above of big data literature shows that there does seem to be a confusion surrounding big data terms and concepts and their definitions. The present paper, therefore, aims to provide a basic categorization of big data terms and concepts to facilitate the understanding and the development of the discourse surrounding big data. This categorization makes use of the facet analysis technique to capture and present concepts in a meaningful and logical order.

3.0 Big data topics and issues: a facet analysis approach

A number of publications have proposed categorizations of big data. For instance, The NSF (2005) report on Long-Lived Digital Data Collections suggests that "Data can also be distinguished by their origins – whether they are observational, computational, or experimental. This distinction is crucial to choices made for archiving and preservation." The report also proposes three types of digital data collections, namely research data collections, resource and community data collections and reference data collections. The European Bioinformatics Institute (*EBI*) and the Natural Environment Research Council (NERC) refer to canonical

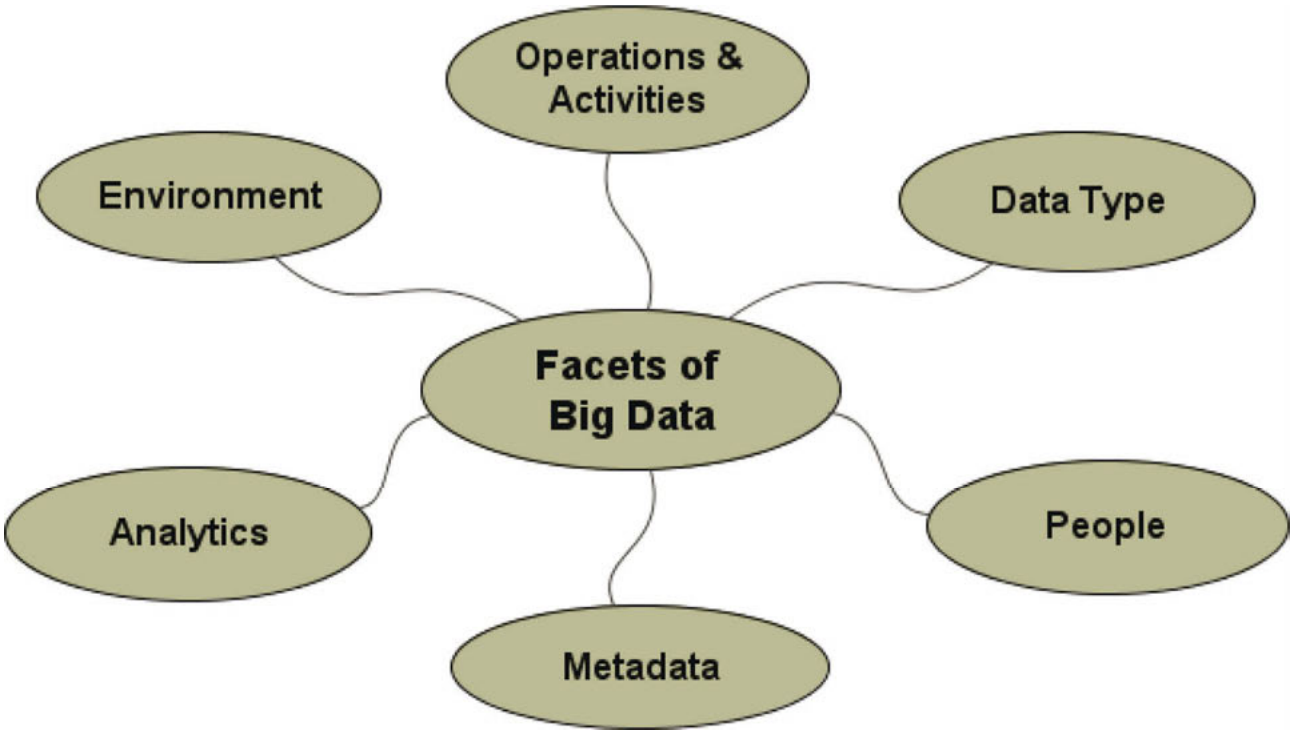

Figure 1. Facets of big data

data (data which has minimal variation) and episodic data (changing data e.g. in life of a cell), which may be unique in time and place e.g. climate information. A further categorization is into raw, processed, derived data and metadata (Lyon 2007). More recently, Wallis et al. (2012) in a study of Center for Embedded Network Sensing (CENS) data identified six dimensions of CENS data: a) background and foreground; b) observation, experimental, and simulation data; c) old and new; d) collection in lab or field; e) raw versus processed; and, f) collection by the team or obtained from external sources.

In order to provide a more comprehensive perspective of the topics and issues surrounding big data, the general principles of facet analysis is used to develop high level categories as well as sub-facets that represent specific types, instances or aspects of the high level facets. Facet analysis was introduced by Ranganathan (1962) as a model for the development of knowledge organization systems such as library classifications and thesauri. Based on this theory, Aitchison et al. (2002) provide a more specific and descriptive set of fundamental categories that are useful as a practical basis for facet analysis. These are as follows:

1. Entities, things, and objects subdivided by characteristics and function
2. Actions and activities
3. Space, place, location, and environment
4. Time
5. Kinds or types; systems and assemblies; applications and purposes

In this paper, a set of facets was developed to provide a framework for the conceptualization, discussion, exploration and research on topics and issues related to big data. This analytical framework does not claim to be comprehensive, rather it aims to provide a starting point for developing and documenting the discourse of big data in order to support research, teaching, learning and practice in the area of big data.

To develop a set of facets to categorize topics and issues related to big data, a wide range of sources were consulted. These include research reports produced by the funding agencies in the US, Canada and in the UK, journal articles, scholarly monographs, technology blogs, and conference proceedings. Particular attention was paid to the ways in which the literature conceptualized and categorized topics and themes related to big data. The review of literature demonstrated an evident gap for a concept map that could illustrate the key facets and aspects of big data in a coherent and meaningful fashion. Based on this analysis, six high level facets were developed, namely data type, environment, people, operations and activities, analytics, metadata. Figure 1 shows a visual representation of these facets.

The proposed facets here can be mapped onto the fundamental categories proposed by Aitchison et al. (2002) as follows:

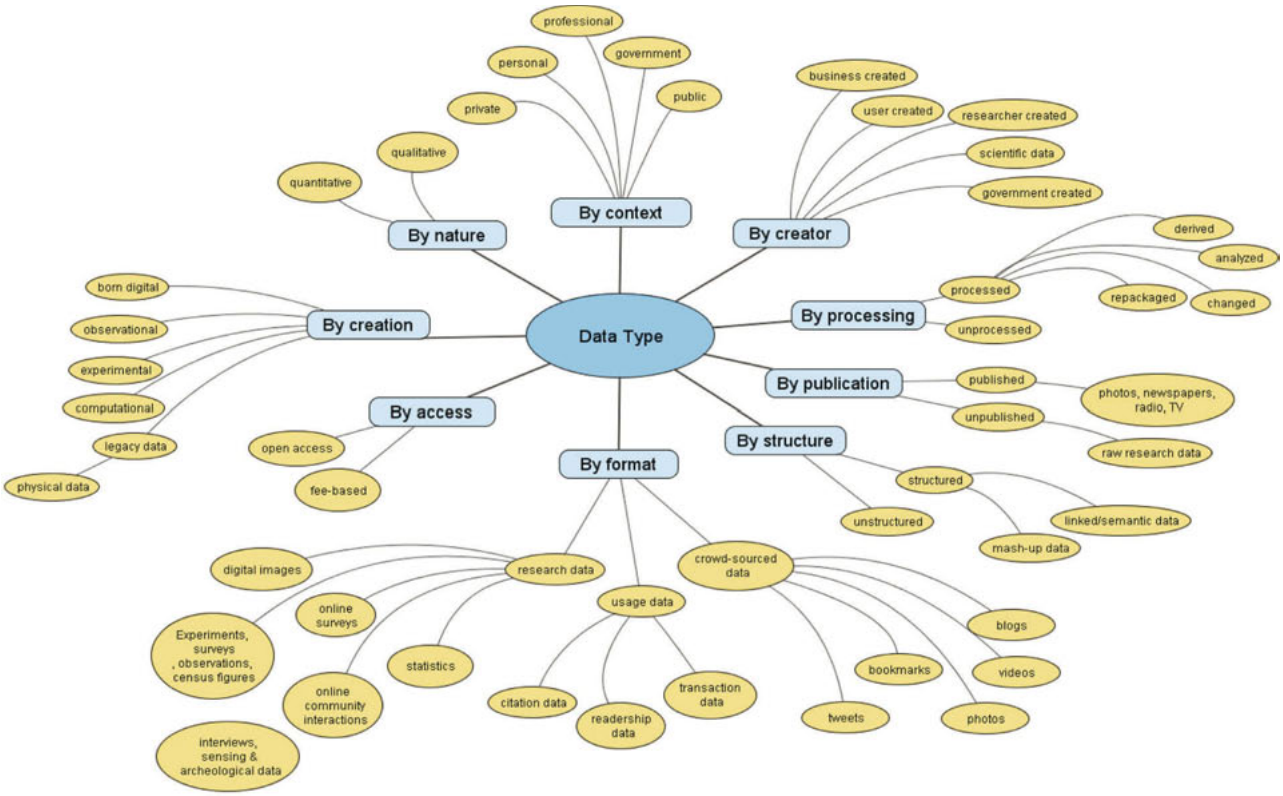

Figure 2. Visual presentation of the "Data Type" facet

Activities and operations = Energy
People = Agent
Environment = Space
Data Type = Property
Metadata = Entities
Analytics = Kinds or types of (systems)

While the first four categories namely "activities and operations," "people," "environment" and "data type" are easier facets to map onto the fundamental categories, the last two, namely "metadata" and "analytics" prove to be more subtle. Metadata is viewed as an entity here because of its unique function in identifying and locating data packages and should be distinguished from the "data type" facet. The distinction between data and metadata in this context is important as the review of emerging literature on big data points to a vague conceptualization of big data without any particular reference to how crucial a role metadata can play in this context. The "analytics" facet is proposed to cover the systems of analysis and visualization, since these two are among the most referenced topics in the big data literature. Further, they tend to be among the terms that co-occur particularly frequently with big data in the literature.

It should be noted that due to the highly conceptual and theoretical nature of facet analysis and the various approaches to its implementation, the mapping between the fundamental categories and the big data facets proposed in this paper could be subject to a variety of interpretations. This kind of mapping is conducted to demonstrate how facet analysis can be used to make sense of new and emerging areas of research and developments. As a result, the analysis and the facets may not be representative of a mutually exclusive set of categories.

Table 1 provides the high level facets as well as sub-facets, values for each sub-facet and instances of each value. Each facet has its own sub-facets, which aim to provide a more specific, detailed and categorized account of a facet. The values listed, provide a more specific set of aspects or areas related to each sub-facets. In some cases, Table 1 provides instances of a particular value. This is to provide examples and instances to clarify each sub-facet or value. It should, once again, be stressed that the "analytics" and "metadata" facets are considered and highlighted as separate facets due to their importance in the process of managing and making sense of big data. With the emergence of big data sets and repositories, it is crucially important to discuss the role and importance of metadata for the organization, access, retrieval and reuse of big data.

Figure 2 provides a delineated and visual representation of the Data Type facet and its many different dimensions and aspects.

Facets	Sub-facets	Values	Instances
Data Type	**By creation**	-Born-digital Observational Experimental Computational	-Machine-generated -Human generated
		-Legacy data	-Physical data
	By nature	-Qualitative -Quantitative	
	By creator	-User created -Researcher created-Scientific data -Government created -Business created	
	By context	-Public -Private -Personal -Professional -Government	
	By format	-Research data (both qualitative and quantitative)	-Experiments, surveys, observations, census figures economic data and demographic, interviews, sensing and archeological data -Statistics -Digital images -Online surveys, online community interactions
		-Usage data	-Citation data -Readership data -Transaction data -Transaction logs
		-Open crowd-sourced data	-Streams of tweets, blogs, photos, and videos, bookmarks
	By publication	-Published	-Manuscripts, photographs, diaries, Television, radio and newspapers
		-Unpublished	-Raw research data (transaction logs)
	By processing	-Processed -Derived -Analyzed -Changed -Repackaged -Unprocessed (raw)	
	By structure	-Structured -Linked data -Semantic web data -Mash-up data -Unstructured	
	By access	-Open access -fee-based	
Environment	**Physical**	-Libraries -Archives -Museums -Publishers -Universities -Funding agencies -Statistical agencies -Media organizations -Laboratories -Field	
	Web-based	-Recommendation systems -Social networks -Social media -Search engines -e-business sites -Data archives	

(Table 1.)

Facets	Sub-facets	Values	Instances
		-Institutional repositories -Digital libraries -Virtual organizations -Cloud-based systems and services -Mobile computing providers -Information providers -Harvesters -Data commons -Data centres -Collaboratories -Observatories	
People	**Creators**	-Scholars -Scientists -Social scientists -Humanities scholars	
	Organizers and curators	-Archivists -Curators -Librarians -Records managers -Information managers	
	Users	-Researchers -Scholars -Students -Readers -Shoppers -Gamers	
	Information technology managers	-Database managers -Knowledge engineers -Information scientists -Data engineers -Data scientists	
Operations and activities	**Management and preservation**	-Data capture -Data curation -Data archiving -Data management -Data preservation -Data access -Data interoperability -Data discovery -Data privacy management	
Analytics	**Qualitative, quantitative, textual & learning analytics**	-Mathematical and computer modeling -Visualization -GIS-based analysis -Data mining and analytics -Web analytics -Informetric and webometric analytics -Simulation -Statistical analysis -Exploratory & confirmatory analysis -Transaction log analysis -Textual, discourse, content, conversation & interpretive analysis	-Community mining in social networks -Social recommenders - Data and information interaction behaviours including: Gaming Reading Reviewing Researching Shopping Studying Using Viewing
Metadata	**By creation**	-Manually assigned -Automatically assigned -Semi-automatically assigned	

(Table 1.)

Facets	Sub-facets	Values	Instances
	By creator	-Author generated -User generated -Librarian/indexer generated -Automatically extracted or harvested	
	By type	-Identification and descriptive	Title, author, creator
		-Administrative metadata	Condition, control, access
		-Content ratings metadata	Audience, use metadata
		-Linkage and relationship metadata	Relation, origin
		-Provenance metadata	Source, creator
		-Terms and conditions	Rights, reproduction restrictions
		-Structural and technical metadata (Greenberg, 2005)	Compression ratio, format, file type
	By content	-Collection level metadata -Item level metadata	

Table 1. Facet analysis of big data topics and issues

This kind of conceptualization of big data does not claim to be all-encompassing, but it aims to provide a framework for thinking and talking about big data in a more systematic manner. Research, teaching and development related to big data can benefit from the facets proposed in Table 1.

4.0 An information science perspective: Research areas and methodologies

Taking a broader perspective, this section aims to highlight some of the contributions that information science can make to the better understanding and studying of big data. As was noted in the introduction, the second objective of this paper was to draw on the methodological and theoretical frameworks in information science to propose new ways of looking at and researching big data. A number of research methodologies and approaches have been devised and developed in information the potential to benefit research into big data. Analysis and evaluation of information search behaviour, user transaction and interaction data analysis, usability evaluation, semantic and subject analysis of content as well as citation analysis and webometric methodologies are examples of research methods and approaches that could be utilized to study big data. For instance, textual, semantic, qualitative and subject analysis of large data sets can benefit from knowledge organization systems such as ontologies, thesauri, taxonomies and other types of controlled vocabularies that have been widely used by information scientists for decades. These tools, most of which available digitally, may be used for the analysis of and provision of access to big data repositories. Further, they could be used for automatic description and assignment of subject metadata to big data repositories and collections. Currently, there are a number of prototype systems that have incorporated knowledge organization systems to support the organization and management of and access to linked data repositories. These projects make use of Simple Knowledge Organization System (SKOS), a World Wide Web Consortium standard for organizing large open data collections. Standards such as SKOS could be introduced to support the description and discovery of big data.

Table 2 provides a select number of areas of research methodologies and approaches in information science that could contribute to the study, exploration and development of big data. The specific areas listed in the second column provide a more granular set of methodological frameworks that can be utilized in the context of big data. The third column provides specific examples of analysis and evaluation in relation to big data. For instance, the use of big data repositories by scientists, social scientists and humanities scholars could draw upon the frameworks developed for the evaluation of user information interaction behaviour. The ways in which researchers may make use of big data for research and teaching purposes can be traced using webometirc, informetric and bibliometric approaches. Best practices developed in the area of digital libraries in the past twenty years can contribute to the management, preservation, and sustainable development of big data repositories.

Interoperability between and among big data repositories can be facilitated through the effective use of collection level metadata and subject description. Lynch (2008, 28)

Research approaches and methodologies	Specific frameworks & areas	Big data applications
Information retrieval interaction methodologies	-Information searching and retrieval models -Cognitive, affective and emotional aspects of information search and retrieval -Human information interaction -Relevance research	-Term level analysis -Search level analysis -Interaction level analysis -Behaviour level analysis -Context level analysis -Situation level analysis -User level analysis -System level analysis
Information behaviour	-Information needs and use behaviour assessment	-Potential, perceived and actual needs and uses of big data sources and repositories in the context of teaching, research and learning
Webometric, informetric and bibliometric methodologies	-Web impact factor -Link and path analysis -Citation, co-citation and domain analysis -Scholarly communication -Research evaluation	-Establish methodological frameworks to automatically explore and evaluate links and citations between and among different big data repositories, in particular the process of creating, publishing, re-using and repackaging
Transaction log analysis methodologies	-Search behaviour patterns -Query formulation and expansion behaviour -User-web interaction behaviour -Usage analysis -Viewing, reading and downloading behaviour	-Analysis of different types of users and their interaction with big data, including the evaluation of the use, re-use, integration, visualization, as well as a delineation of types and nature of interaction (viewing, searching, and making sense of data, , data manipulation, data integration, data presentation)
Knowledge organization and representation	-Simple Knowledge organization System (SKOS) -Controlled vocabularies -Semantic web, open and linked data -Resource description and discovery -Metadata (item and collection level)	-Identification, consistent description and registry of big data sources and repositories using ontologies, thesauri, taxonomies and classification schemes -Evaluation of subject access to data -Evaluation of metadata-enhanced access to big data based on new data-specific metadata elements and access points -Exploring the effectiveness of various metadata generation approaches for bi data
Digital libraries	-Digital objects -Digitization -Digital preservation -Interoperability -Rights management -Search and retrieval of heterogeneous digital information	-Effective identification, management and preservation of big data -Cross-searching and cross-browsing different big data sets using interoperable systems and services -Integration and management and use of hybrid data sources including born-digital and digitized data

Table 2. Select list of information science research areas and methodologies and their applications for big data

stresses the importance of metadata for big data. He notes that one of the key aspects of data stewardship is:

> To define and record appropriate metadata—such as experimental parameters and set-up—to allow for data interpretation. This is best done when the data are captured. Indeed, descriptive metadata are often integrated within the experimental design. Description includes tracing provenance—where the data came from, how they were derived, their dependence on other data and all changes made since their capture. Proper stewardship requires documenting the storage formats.

The crucial role of metadata in relation to big data becomes increasingly evident as many big data repositories are created and require efficient access mechanisms. Proper metadata assigned to big data could have many advantages, including facilitating collaboration among organizations and institutions responsible for the creation and maintenance of big data collections. The key concept of metadata interoperability suggests that big data sets could be described using standard metadata in order to support the re-use and re-purposing of big data sets held by various institutions and organization. In order to achieve this, there is an evident need to develop and use metadata interoperability models and practices to allow big data to be flexibly and effectively used across many different platforms, domains, disciplines, systems and services. Some of the key questions that metadata could answer in the context of big data initiatives and projects are: How do we collect, code, describe and cite data? How do we describe and provide access to legacy data? How do we ensure consistent description and constant access to various big data collections and their associated technologies? How do we integrate digitized collections into big data collections? How do we develop big data-specific registry and metadata application profiles?

The rationale behind Table 2 lies in the recognition of some of the long standing research traditions and methodologies in information science that can now serve us in thinking, conceptualizing, analyzing and making sense of big data. This not only provides a new frontier for information scientists and information professionals to be involved in current digital data developments, but it will also present new opportunities for cross and interdisciplinary information work that will benefit researchers in information science as well as in other domains and disciplines. A number of American LIS schools have already started developing big data and data science courses and programs. It is timely and important to conceptualize and discuss the role of information science with regards to big data developments.

5.0 Conclusion

The overarching aim of this paper was to create conceptual and concrete links between information science and knowledge organization methods and traditions and the emerging area of big data. This paper provided a facet analytical approach to big data to lay a basic framework for the study, exploration and discussion of various big data related topics and issues. Six high level facets, namely data type, environment, people, operations and activities, analytics, and metadata, were introduced to map the big data issues and areas along with sub-facets and instances of those sub-facets. In line with the second objective of this paper, a number of information science research areas and methodological frameworks were introduced to demonstrate their applicability and suitability for research on big data.

Following the emergence of search engines, digital libraries and various types of information repositories in the 1990s and 2000s, the notion of big data is gradually finding its way into our new digital information environment. The increasing pace of data-intensive teaching, learning, business, research, and development necessitates a well-rounded understanding of the key concepts and issues of big data. This understanding will support effective and efficient planning and management of the processes and procedures for the identification and streamlined use of big data. The successful operations of many organizations and institutions that produce, process, manage, use and maintain big data hinges on a clear understanding of the complexity and multifaceted nature of big data and its associated challenges.

Future research needs to expand and enhance this typology to cover the more subtle and nuanced aspects and areas of big data. Furthermore, due to the multidisciplinary nature of big data, various disciplines can build on this typology and can contextualize it as a framework for the discussion, conceptualization and exploration of big data.

References

Aitchison, Jean, Gilchrist, Alan and Bawden, David. 2002. *Thesaurus construction and use: a practical manual*. London: Fitzroy Dearborn.

Bizer, Christian, Boncz, Peter. A., Brodie, Michael and Erling, Orri. 2011. The meaningful use of big data: four perspectives - four challenges. *SIGMOD record* 40 no. 4: 56-60.

Borgman, Christine L. 2007. *Scholarship in the digital age: information, infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Borgman, Christine L. 2012. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology* 63: 1059-78.

Davenport, Thomas, H., Barth, Paul and Bean, Randy. 2013. How 'big data' is different. *MIT Sloan Management Review* 54 no. 1. Available http://sloanreview.mit.edu/article/how-big-data-is-different/.

De Witt, Shaun., Sinclair, Richard, Sansum, A. and Wilson, Michael. 2012. Managing large data volumes from scientific facilities. *ERCIM news* 89. Available http://ercim-news.ercim.eu/en89/special/managing-large-data-volumes-from-scientific-facilities

Digging into Data Challenge. http://www.diggingintodata.org/

Dumbill, Edd. 2013. Making sense of big data. *Big data* 1: 1-2.

Foskett, Douglas. J. 2003. Facet analysis. In Drake, Miriam, ed., *Encyclopedia of library and information science, second edition*. New York, NY: Dekker, pp. 1063-7.

Greenberg, Jane. 2005. Understanding metadata and metadata schemes. *Cataloging & classification quarterly* 40 no. 3-4: 17-36.

Hodson, Simon. 2012. JISC and big data. In *Eduserv symposium 2012: big data, big deal? May 10, 2012, London, UK.*

Hjørland, Birger. 2013. Facet analysis: the logical approach to knowledge organization. *Information processing & management* 49: 545-57.

Jacobs, Adam. 2009. The pathologies of big data. *Queue* 7 no. 6: 1-12.

Kuznetsky, Dan. 2010. *What is "big data?*" Available at http://www.zdnet.com/blog/virtualization/what-is-big-data/1708

La Barre, Katherine. 2010. Facet analysis. *Annual review of information science and technology*, 44: 243-84.

Li, Yuelin and Belkin, Nicholas J. 2008. A faceted approach to conceptualizing tasks in information seeking. *Information processing & management* 44: 1822-37.

Lynch, Clifford. 2008. Big data: How do your data grow? *Nature* 455: 28-9. Available at http://www.nature.com/nature/journal/v455/n7209/full/455028a.html.

Lyon, Liz. 2007. *Dealing with data: roles, rights, responsibilities and relationships. Consultancy report.* June 19, 2007. Available at http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf.

Milonas, Elizabeth. 2011. Wittgenstein and web facets. In Smiraglia, Richard P., ed. *Proceedings from North American Symposium on Knowledge Organization*, Vol. 3. Toronto, Canada, pp. 33-40

National Science Foundation. 2005. *Long-lived digital data collections enabling research and education in the 21st century.* Available at http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf.

National Science Foundation. 2012. *Core techniques and technologies for advancing big data science & engineering (BIGDATA) solicitation.* Available at http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.htm.

Oxford Internet Institute. 2014. *Big data: demonstrating the value of the UK web domain dataset for social science research.* Available at http://www.oii.ox.ac.uk/research/projects/?id=88.

Powell, Andy. 2012. *Final remarks*. In Eduserv Symposium 2012: Big Data, Big Deal? May 10, 2012, London, UK.

Ranganathan, Shiyali Ramamrita. 1967. *Prolegomena to library classification*. New York: Asia Publishing House.

Shiri, Ali. 2013. The many facets of 'query' in interactive information retrieval. In *Proceedings of ASIS&T Annual Meeting, Association for Information Science & Technology*, Montreal, November 1- 6, 2013.

Smiraglia, Richard. 2009. Modulation and specialization in North American Knowledge Organization: visualizing pioneers. In Jacob, Elin K. and Kwasnik, Barbara, eds., *Pioneering North American contributions to knowledge organization, Proceedings of the 2d North American Symposium on Knowledge Organization*, June 17-18, 2009, pp. 35-46.

Wallis, Jillian. C., Wynholds, Laura A., Borgman, Christine. L., Sands, Ashley and Traweek, S. 2012. Data, data use, and scientific inquiry: two case studies of data practices. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, JCDL '12*. Washington, D.C.: ACM, pp. 19-22.

Warden, Pete. 2011. *Big data glossary*. Sebastopol, CA: O'-Reilly.

Wu, Xindong., Zhu, Xingquan, Wu, Gong-Qing and Ding, Wei. 2014. Data mining with big data. *IEEE transactions on knowledge and data engineering* 26 no. 1: 97-107.