The Interplay of Big Data, WorldCat, and Dewey

Rebecca Green
Assistant Editor, Dewey Decimal Classification,
OCLC Online Library Computer Center, Inc.
Dewey Editorial Office, LM-548, Library of
Congress, 101 Independence Ave., S.E.,
Washington, DC, 20540-4330 USA
greenre@oclc.org

Michael Panzer
Editor in Chief, Dewey Decimal Classification,
OCLC Online Library Computer Center, Inc.
6565 Kilgour Place,
Dublin, Ohio 43017-3395 USA
panzer@oclc.org

ABSTRACT

As the premier example of big data in the bibliographic world, WorldCat has the potential to support knowledge discovery in many arenas. After giving evidence for a big data characterization of WorldCat, the paper explores this knowledge discovery potential from two perspectives related to the Dewey Decimal Classification (DDC) system: (1) how WorldCat data can inform development of the DDC (classification analytics) and (2) how DDC-classified content in WorldCat can shed light on the bibliographic world itself (collection analytics). In the realm of classification analytics, WorldCat data support decisions to modify the DDC by expanding or reducing the number of classes, adding topical coverage, or adding subject access points; data analysis can support recognition of (1) trending topics and (2) the faceted structure of subject domains. In the realm of collection analytics, the paper considers as possible applications the use of the DDC in the topical "fingerprinting" of categorized content in WorldCat or in performing a bibliographic gap analysis,

Keywords

Big data, WorldCat, Dewey Decimal Classification, DDC, trending topics, facet analysis, classification analytics, collection analytics.

INTRODUCTION

This paper looks at OCLC's WorldCat as a big data source of potential benefit to the development of the Dewey Decimal Classification (DDC) system. Logically, the paper is divided into three parts. The first part will set the stage by giving relevant background on big data, WorldCat data, and the DDC's use of WorldCat data. The second, and main, part presents specific scenarios in which WorldCat data can be used to illuminate specific aspects of development of the DDC. In some cases, analysis of the data needs only descriptive statistics; in other cases, the data are the basis for decision-making based on inferential analysis. The third part suggests how DDC classification of WorldCat data might provide insights from collection-related analysis.

SETTING THE STAGE

This first part summarizes what is meant by "big data," explores WorldCat's claim to big data status, and considers how the development of the DDC is anchored in literary warrant, with WorldCat as a primary reflection/source of that warrant.

Big data

"Big data" is found on Wikipedia's List of buzzwords (n.d.), where its meaning is given, tongue-in-cheek, as "larger data sets than last month." Many enterprises feel awash in data characterized by such volume, variety, and velocity that the capacity to manage their data is threatened, if not indeed overwhelmed.

Volume

What is the threshold for a data collection to qualify as big data? Whatever it is, it's a moving target. In recent years, the answer might have been given in terms of terabytes (1 TB = 1000^4 bytes), but characterizations are more commonly now given in terms of petabytes (1 PB = 1000^5 bytes) (Villanova University, n.d.) or even exabytes (1 EB = 1000^6 bytes) (Francis, 2012). To give a sense of what those orders of magnitude mean, we consider that the estimated volume for all the printed material in the Library of Congress is ("only") 15 TB (All too much, 2010).

Or maybe the number of transactions represented by the data is more critical than the number of bytes. However, since transaction is not a unit that is comparable across data scenarios, the implicit suggestion is not reflected in specific pronouncements of big data thresholds.

In the end, since it is not solely the volume of data involved that proves overwhelming (Gartner, 2011), but the combination and interaction of volume, variety, and velocity on the one hand, versus computing capacity—data communication, memory, processing, etc.—on the other hand, the truism "my big data is not your big data" applies.

Advances in Classification Research, 2013, November 2, 2013, Montreal, Quebec, Canada.

Variety

Problems caused by variety in the data is manifest in several different ways—there's variety in the variety! When data derive from different sources, we face standard issues of interoperability. If, for instance, the same "fact" comes from different sources, reflecting different perspectives, using different vocabulary, reflecting different standards, will it be recognized as the same fact?

In addition to the use of different standards, which might lead to different data structures, we may have unstructured data (e.g., full text) or partially structured data in which it is difficult to isolate data elements.

Big data situations often involve semantically related datasets. But if the datasets encounter the problems hinted at above, they could be conceptually like a relational database, but without the foreign keys that allow one to join data from different relations.

Velocity

Problems associated with data velocity involve a mismatch between two speeds: the speed at which data are created and transmitted vs. the speed at which data can be analyzed. Is the system in a steady state where the rate of data output is at least as great as the rate of data input, or does the backlog of data to be processed continue to increase?

WorldCat as big data

Before exploring how OCLC's WorldCat stacks up against the conventional sense of big data, we must first ascertain what WorldCat comprises. Just as a day (i.e., a 24-hour period) consists of a day (i.e., a period of light) and a night (i.e., a period of darkness), so WorldCat has both larger and smaller senses, in which the smaller sense is part of the larger sense. On the one hand, WorldCat refers to a bibliographic database. While holdings and authority data inform what the user sees when searching WorldCat, the predominant sense of WorldCat in this smaller sense is a set of bibliographic records. On the other hand, WorldCat can also refer to all the data made available to member institutions through a WorldCatLocal implementation.

The full array of ("bigger") WorldCat data includes records in the MARC Bibliographic Format, in the MARC Holdings Format, in the MARC Authority Format (e.g., LCSH, FAST, BISAC, MeSH, VIAF records), vendor records, WorldCat knowledge base data, institutional registry data, and institution-specific acquisitions, circulation, ILL data.

WorldCat has recently surpassed the 300 million record mark for bibliographic data and the 2 billion record mark for holdings data. WorldCat's authority data includes ca. 26.4 million Library of Congress Subject Headings and 24.2 million VIAF clusters, with 21 million links between records. Article data in WorldCat Local, with its more than one billion records, dwarfs the number of records in the traditional bibliographic data set.

While the volume and variety of data in WorldCat is respectably large, these measures do not compete with volume and velocity measures for many social media sites, e.g., Twitter, YouTube. However, in its sphere, that is, in the bibliographic world, WorldCat has a better claim to the big data moniker than any other collection of data.

Moreover, absolute volume is not always as important as completeness of coverage. A smaller data collection that is 90% "complete" may trump a huge collection that is not as exhaustive in its domain. WorldCat's unparalleled inclusion of bibliographic records from national libraries throughout the world result in broader and deeper coverage of the universe of published knowledge than is the case for any other source.

Early in OCLC's life—back when the "OC" stood for Ohio Colleges—circulation and ILL were envisioned as being handled through the cooperative. Now that the supporting infrastructure is available, as libraries adopt a WorldShare Management Services solution, WorldCat could become big data in a more conventional sense.

Literary warrant and the DDC

Development of the DDC is based on literary warrant, most commonly as reflected in WorldCat. The DDC editorial rules specifically call for literary warrant to be taken into account for a number of situations, the most important of which are expansions (i.e., the development of new classes) and reductions (i.e., the discontinuing of entire classes to superordinate notation). But we will find additional and innovative uses to which WorldCat data can be put.

CLASSIFICATION ANALYTICS

As we look at specific uses of WorldCat data in support of DDC development, we will see that in some instances, the data analysis is of a straightforward, descriptive nature, while in other instances, the data analysis is of a more inferential nature. Current uses of literary warrant data from WorldCat for DDC development tend to be of the former type, while the more inferential analysis forms part of our future research agenda.

Classified works

Periodic profiles of the distribution of classified works across the classification constitute one sort of literary warrant based on descriptive statistics. These profiles can be used to identify appropriate expansions or reductions.

For example, we see below the current development under 306.44, which is part of the Sociology schedule. The topical scope of 306.44 is quite broad, with only bilingualism and multilingualism, on the one hand, and language planning and policy, on the other, having their own numbers. Are there other aspects of 306.44 that should be developed further?

306.44 Language

Including pragmatics

Class here anthropological linguistics, ethnolinguistics, sociolinguistics

306.446 Bilingualism and multilingualism

306.449 Language planning and policy

306.449 4–.449 9 Specific continents, countries, localities in modern world

Add to base number 306.449 notation 4–9 from Table 2, e.g., language policy of India 306.44954

Table 1 displays the growth in the use of 306.44 itself over the past six five-year time periods. During this time, the number of records[1] classed at 306.44 have increased twenty-fold, as shown in the middle column. The right-most column shows how many of the works classed in 306.44 were specific to English, French, German, and Spanish, numbers that have increased thirty-fold over the same time period, with a particular boost in recent years. These statistics suggest that an expansion under 306.44 in which the specific language can be expressed is warranted.

Time period	Records retrieved	Language-specific
1981–1985	120	14
1986–1990	412	59
1991–1995	912	134
1996–2000	1230	163
2001–2005	1603	199
2006–2010	2369	446

Table 1.Use of 306.44, 1981–2010.

In contrast, let us consider part of the development under 006.33 Knowledge-based systems, provided in anticipation of future literature growth**:**

006.336 Programming for knowledge-based systems

006.336 3 Programming languages for knowledge-based systems

006.337 Programming for knowledge-based systems for specific types of computers, for specific operating systems, for specific user interfaces

006.338 Programs for knowledge-based systems

Table 2 displays the number of works classed in each of these numbers over the past six five-year periods (including the current incomplete five-year period). The numbers have been, in general, modest, unlikely to have prompted such a development had literary warrant not have been anticipated in advance.

Time period	006 .336	006 .3363	006 .337	006 .338
1986–1990	0	1	0	0
1991–1995	1	1	1	0
1996–2000	1	0	5	3
2001–2005	6	0	5	1
2006–2010	14	1	10	3
2011–2015	3	0	0	1

Table 2.Use of subdivisions of 006.33, 1986–2015.

Access Points

Subject heading data in DDC categorized content can also be used to identify areas where expansions of new classes should be considered. Such analysis can also support the adding of new Relative Index or mapped headings for DDC classes, as well as the adding of new topics to class description. For example, a general principle of DDC development is that if and when a topic in standing room achieves a sufficient level of literary warrant, an expansion should be provided for the topic. When a topic is matched by a subject heading, straightforward searches on that subject heading provide the data needed. Consider the DDC class below:

004.678 Internet

Including extranets, virtual private networks

Class here World Wide Web

While topics in class-here notes approximate the whole of the class and are thus not subject to expansion, the standing-room topics in including notes are potentially ripe for expansion. The LCSH Extranets (Computer networks), for which Virtual private networks (Computer networks) is a lead-in term, fully matches the including-note topics here. The WorldCat search *dd:004.678* and (hl:extranets w computer w networks)* retrieves 69 records, easily surpassing the literary threshold of 20 titles needed to support an expansion.

[1] Unless specified otherwise, numbers of records retrieved represent raw results, neither grouped by work, nor weighted by number of holdings, as would often be most appropriate.

Another use of WorldCat subject heading data is made in checking for new topics that need to be reflected in class descriptions and/or indexing. For example, the WorldCat search *yr:2010-2013 and (dd:004* or dd:005* or dd:006*)* located recent computer science literature. From these records were drawn LCSH main headings that occurred 5 or more times that had not been mapped to a DDC class and that did not match a Relative Index term. The topics represented by these subject headings were then studied according to standard procedures to identify which of them should be added to class descriptions and indexed. On the basis of such analysis, microcontrollers, sensor networks, and wireless sensor networks are being added to the descriptions of the classes indicated.

004.6 Interfacing and communications

. . .

Including sensor networks

. . .

006.22 Embedded computer systems [formerly 004.1]

Class here microcontrollers

For a specific aspect of embedded computer systems, see the aspect, e.g., systems analysis and design of embedded computer systems 004.21, wireless sensor networks 004.6, software for embedded systems 005.3

Trending Topics

But we don’t want to have to depend on sustained use of a subject heading before recognizing the need to account for the topic it reflects. Can we identify trending topics so that classifiers can find direction on how to classify them as literature emerges?

Just as my big data is not your big data, so my trending topics are not your trending topics. That is, trending topics in the bibliographic world are not the same as trending topics on Twitter. In social media contexts trending topics register “a sudden high-magnitude spike in activity over some baseline of activity” (Nikolov & Shah, 2012); that is, overall popularity doesn’t matter as much as novelty, as reflected in the sudden emergence of interest (@Twitter, 2012). But for bibliographic classification, “quickness” of achieving literary warrant comes in terms of months and years, and some degree of sustained literary warrant should be at least anticipated before provision is made for the new topic.

In the best of all possible worlds, the first literature on bibliographic trending topics would always be accompanied by the minting of new subject headings. But such is not always the case. Big data has been in the air for more than a decade, but the LCSH Big data was not created until August 2012. Tim Berners-Lee introduced a clearly delineated notion of linked data in 2006, but the LCSH Linked data was not created until July 2013. Table 3 shows results from searching on the phrases “big data” and “linked data” as subjects or titles in WorldCat for 2008–2013. Literature on these two subjects can be seen to be substantial. Why would subject headings not have been created earlier? The term “big data” has a breeziness to it that could easily have given way or may yet give way to a more grown-up name. (For this reason, the topic is being reflected in the DDC in the category description of 005.7 Data in computer systems as “high-volume data sets”; “big data” will be used only as an access point.) “Linked data,” as used by Tim Berners-Lee, has a very specific characterization, but at the same time people were talking about linked data in more general terms, and linked data structures have been around for many decades.

Year	Big data	Linked data
2008	2	14
2009	0	34
2010	7	72
2011	74	84
2012	227	152
2013	413	114

Table 3. Subject or title searches on “big data” and “linked data” in WorldCat, 2008–2013

The creation of new Library of Congress subject headings can then serve as only a possible source of data as to newly emerging and trending topics. This leaves us not only with the task of detecting trending topics, but also with the task of detecting new topics, some of which may trend.

Another source of information we may look to for early detection of bibliographically trending topics is conference themes, as commonly reflected in their titles. Consider, for example, the following workshop, congress, and conference titles:

- Big data: 29th British National Conference on Databases
- 1st Workshop on Architectures and Systems for Big Data
- Workshop on big data
- Big Data Analytics: First International Conference
- The Semantic Web: Semantics and Big Data: 10th International Conference
- 2012 workshop on Management of big data systems
- 2nd Workshop on Research in the Large : Using App Stores, Wide Distribution Channels and Big Data in UbiComp Research
- IEEE International Congress on Big Data
- Big Data 2 Knowledge (Workshop)

Chapter and paper titles may provide corroborating evidence of the emergence of new and trending topics.

We need to explore some number of topics like big data and linked data across a variety of disciplines to identify possible relationships between occurrence of topics in chapter, paper, and conference titles, etc., relative to the appearance of monographic literature on the topic.

Structure of Discipline

The Classification Research Group (1955) called for facet analysis—arguably the most important contribution to knowledge organization theory and practice of the twentieth-century—to become "the basis of all methods of information retrieval."

Kwasnik (1999), however, identifies the "difficulty of establishing appropriate facets" as a limitation of facet analysis. The postulating of five fundamental (and universal) categories (Personality, Matter, Energy, Space, and Time; Ranganathan, 1985/1962) doesn't suffice: Ranganathan himself characterized facets as "general manifestations" of these fundamental categories. For example, the Organ and Problem facets in Medicine are general manifestations of the Personality and Energy fundamental categories, respectively.

How then are the facets of a subject established? Facet analysis begins with the identification of a number of topics (possibly complex) in the subject area, followed by the determination of specific attributes by which to characterize them. While the choice of facets is circumscribed by the topics needing to be described, their identification is nonetheless subjective to a not-insignificant degree. To the extent that the choice of facets turns out to be an ongoing decision, their identification is also intense and laborious.

If possible, we wish to detect the facet structure of a subject literature automatically, based on its own metadata. In particular, the titles of the monographic literature of a subject often include noun phrases that are, in Ranganathan's words, "particular manifestations" of its fundamental categories. But how can we generalize appropriately from such particular manifestations to facets?

We are proposing to undertake such a task by following these steps:

- Retrieve bibliographic records from WorldCat (2013) for the monographic literature[2] of 2–3 subjects (currently under consideration: education, medicine, music); isolate title data.
- Identify noun phrases in the titles. Because occurrence of noun phrases in WordNet (2010), a lexical database for English, plays a critical role in the generalization process, simple heuristics will be used to identify strings of words in the title that might be or might contain noun phrases. Words will be deleted from the end and then from the beginning of each string and matched against the WordNet noun index until a longest match is identified or no words remain.
- Use the conceptual density measure of Agirre & Rigau (1996) to disambiguate noun phrases and to identify appropriate generalizations.

The overall approach parallels that described in Green & Dorr (2004).

As existing classification systems—such as the Dewey Decimal Classification, the basis of bibliographic retrieval in the first step above—adopt a trend toward greater faceting (Mitchell, 1996), the procedures explored here hold out promise in identifying facets appropriate to the literature to be classified, not only at the discipline level, but also at the level of subclasses within the discipline.

COLLECTION ANALYTICS

Instead of looking at specific uses of WorldCat data in support of DDC development, it should also be possible to shift focus to the topical texture of (subsets of) works in WorldCat themselves. Can Dewey help with topical "fingerprinting" (or a gap analysis) of categorized content in WorldCat?

Ad-hoc subcollection from WorldCat: For Dummies series

As an example of an ad-hoc subcollection we use a popular reference book series that seems to cover every conceivable topic on the planet: the "For Dummies" series. Can Dewey help to survey the subject landscape of this series with over 1800 published titles (according to the publisher)?

For the following, we adopt a work-based rather than a manifestion-based approach. Using a generous approach to searching for works by also including resources that are not books, the "For Dummies" series accounts for more than 5000 works in WorldCat. Table 4 shows how the series stacks up against the DDC main classes.

The series passes a first hurdle (covering all main classes) with flying colors. There are "For Dummies" works available for topics in all main classes, with a particularly strong presence in 000 Computer science, information, & general works (with an overwhelming presence in computer science [004 and structurally subordinate classes 005 and 006]); 300 Social sciences; 600 Technology; 700 Arts &

[2] The search *dd:[e.g.]78* and dt:bks and cs:dlc not su:(biography or juvenile)* is relatively successful in identifying works whose titles mention particular manifestations of facets; that is, works whose titles are more fanciful tend to be filtered out by such a such.

recreation; and a weaker showing in especially 200 Religion and 800 Literature. (Note, however, that all main classes are not equal; one cannot say, for instance, that the 400s, with their 99 "For Dummies" titles, are covered 50% more exhaustively than the 100s, with their 66 titles.)

DDC main class	Number of "For Dummies" titles
000	1726
100	66
200	37
300	435
400	99
500	124
600	1194
700	398
800	30
900	195

Table 4. Distribution of "For Dummies" works across DDC main classes

Figure 1 is a treemap visualization of all three levels of the summaries, which makes heavily populated subject areas easily visible. Let's have a closer look at the makeup of 600s. Almost half of the works are classed in 650 Management & public relations, again with half of them in 658 General management. Some examples: *Small business for dummies* (658.022), *Motivating employees for dummies* (658.314), or *Interpreting company reports for dummies* (658.1512). Also strong is 610, with a majority of sections populated. Notably absent are work that would class in 614 Forensic medicine; incidence of injuries, wounds, disease; public preventive medicine. If anyone is searching for a topic not yet covered by the "For Dummies" series, *Public health for dummies* (if it focuses on medicine, the interdisciplinary number would be 362.1) or *Forensic medicine for dummies* seem to be up for grabs.

The series is particularly soft in philosophy. There are gaps in the second summary, most notably 110 Metaphysics, 120 Epistemology, 180 Ancient, medieval & eastern philosophy, and 190 Modern western philosophy.

What are other notable holes in the topical coverage of the series? While there is a *Latin for dummies*, there does not seem to be a *Classical Greek for dummies* (belonging in the 480s). Also, it seems about every topic in the exact sciences is covered, but there is not a single work in 560 Paleontology. *Fossils for dummies*, anyone?

The same approach could be taken to find unexpected titles in sparsely populated areas.

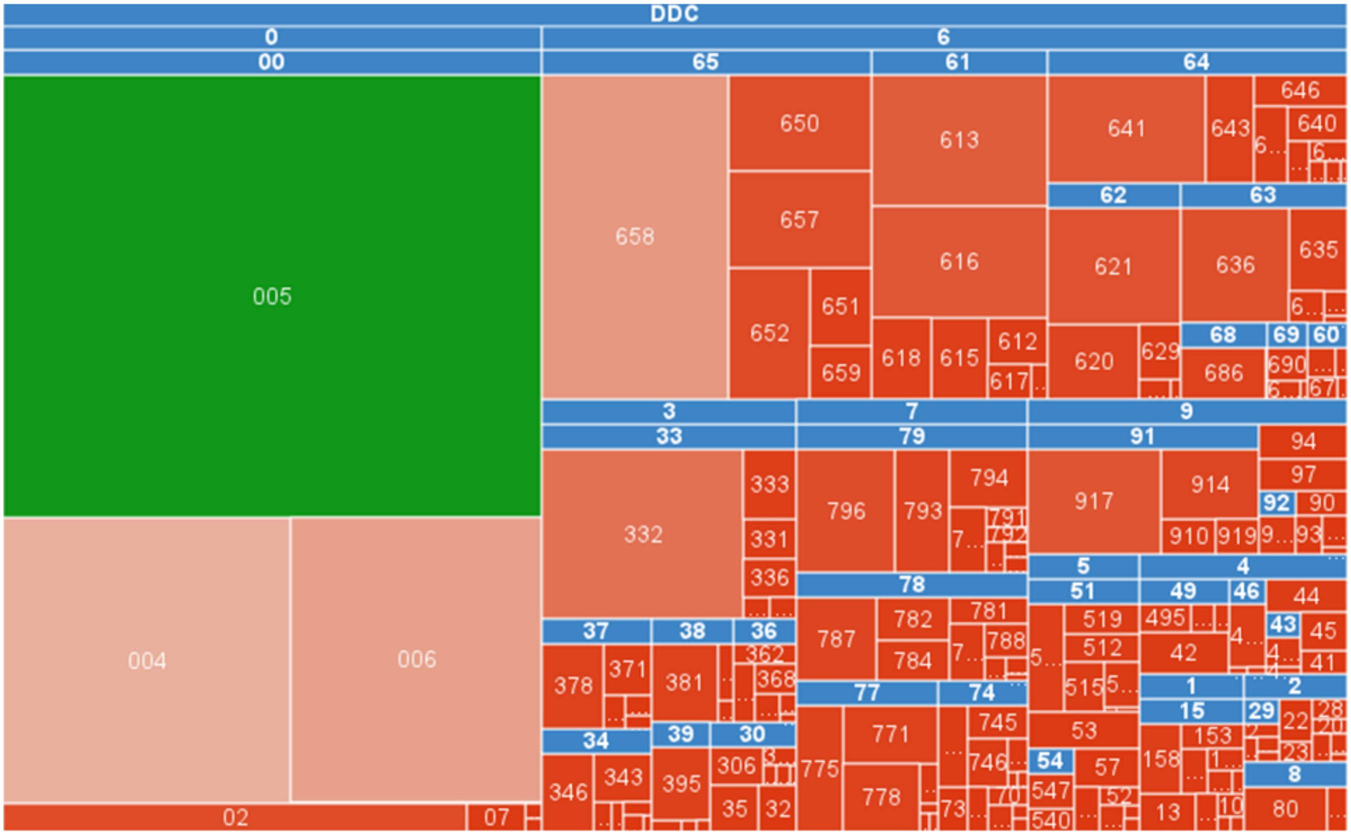

Figure 1. Treemap visualization of "For Dummies" titles across DDC

CONCLUSION

The DDC and WorldCat share the same bibliographic scope. Consequently, WorldCat data (especially in the context of DDC-categorized content) play a prominent role in supporting intelligent development of the DDC. We have seen how WorldCat data can be used to highlight needed expansions and discontinuations, to identify topics with sufficient literary warrant to support explicit mention in the scheme, and to recognize important new (i.e., bibliographically trending) topics. It has also been suggested that DDC-categorized data can be analyzed to identify the faceted structure of a subject domain. At the same time, DDC-categorized data can also be analyzed to support analysis of the bibliographic collection itself, of which topical "fingerprinting" and gap analysis are examples.

As the amount of DDC-categorized content available through WorldCat continues to grow, the interplay between big data, WorldCat, and the Dewey Decimal Classification promises to promote intelligent development of the classification system itself and understanding of the bibliographic world that it organizes.

REFERENCES

@Twitter. (2012). To trend or not to trend [Web log post]. Retrieved from https://blog.twitter.com/2010/trend-or-not-trend

Agirre, E. & Rigau, G. (1996). Word sense disambiguation using conceptual density. In *Proceedings of the 16th International Conference on Computational linguistics (COLING '96)* (vol. 1, pp.16-22). Copenhagen: Center for Sprogteknologi.

All too much. (2010). *Economist* (2010-2-25). Retrieved from http://www.economist.com/node/15557421

Big data definition. (n.d.). In *MIKE2.0* [Method for an Integrated Knowledge Environment wiki]. Retrieved from http://mike2.openmethodology.org/wiki/Big_Data_Definition

Broughton, V. (2006). The need for a faceted classification as the basis of all methods of information retrieval. *Aslib Proceedings 58 (1),* 49–72.

Chalmers, J. (2012). Big data part 3: Worldcat's Identities Network and semi-useless tools [Web log post]. Retrieved from http://joshchalmers.wordpress.com/2012/02/02/big-data-part-3-worldcats-identities-network/

Classification Research Group. (1955). The need for a faceted classification as the basis of all methods of information retrieval. *Library Association Record, 57(7)*, 262-268.

Croll, A. (2013). Implications and opportunities of big data. Presentation at OCLC Americas Regional Council Member Meeting and Symposium, 2013 ALA Midwinter Meeting, Seattle WA.. Retrieved from http://www.youtube.com/watch?v=Ic_BlPesEls

Francis, M. (2012). Future telescope array drives development of exabyte processing. Retrieved from http://arstechnica.com/science/2012/04/future-telescope-array-drives-development-of-exabyte-processing/

Gartner. (2011). Gartner says solving 'big data' challenge involves more than just managing volumes of data [Press release]. Retrieved from http://www.gartner.com/newsroom/id/1731916

Green, R. & Dorr, B. J. (2004). Inducing a semantic frame lexicon from WordNet data. In *Proceedings of the Second Workshop on Text Meaning and Interpretation: held in cooperation with ACL-2004* (pp. 65–72). East Stroudsburg, PA: ACL. Retrieved from http://aclweb.org/anthology/W/W04/W04-0909.pdf

Kwasnik, B. H. (1999). The role of classification in knowledge representation and discovery. *Library Trends 48 (1)*, 22–47.

Lee, P. (2013). What is big data? The definition [Web log post]. Retrieved from http://techrux.net/big-data-definition/

Library of Congress. [199-?]-. *MARC standards*. [Washington, D.C.]: Library of Congress, Network Development and MARC Standards Office. http://purl.fdlp.gov/GPO/gpo13828.

List of buzzwords. (n.d.). In *Wikipedia*. Retrieved October 31, 2013, from http://en.wikipedia.org/wiki/ List_of_buzzwords

Lugg, R. & Harnish, K. (2013). Collection analytics: Using data to drive decisions. Presentation at the 2013 ALA Midwinter Meeting, Seattle, WA. Retrieved from http://www.youtube.com/watch?v=THjUUC8gMpc&list=PL4ekxRiBK0Fi-yIJkJk80yNiE2P5x8Xl8&index=1

Mitchell, J. S. (1996). New features in Edition 21. In *Dewey Decimal Classification and Relative Index*, devised by Melvil Dewey; Edition 21, edited by Joan S. Mitchell, Julianne Beall, Winton E. Matthews, Jr., & Gregory R. New. Albany, New York: Forest Press.

Mitchell, J. S. & Vizine-Goetz, D. (2010). Dewey Decimal Classification (DDC), *Encyclopedia of Library and Information Sciences*, *Third Edition, 1:1*, 1507–1517.

Nikolov, S. & Shah, D. (2012). A nonparametric method for early detection of trending topics. In V. Blondel, et al. (Eds.), *Interdisciplinary Workshop on Information and Decision in Social Networks, November 8 - 9, 2012, MIT* (pp. 43-44). Cambridge, MA: Connection Science and Engineering Center, MIT. Retrieved from http://wids.lids.mit.edu/wids-abstracts.pdf

OCLC Online Computer Library Center, Inc. (2013). OCLC partnership with Plum Analytics uses WorldCat to measure impact of research [Press release]. Retrieved from http://www.oclc.org/news/releases/2013/201343dubllin.en.html

Ranganathan, S. R. (1985/1962). Facet analysis: Fundamental categories. In *Theory of subject analysis: a sourcebook*, edited by Lois Mai Chan, Phyllis A. Richmond, and Elaine Svenonius.(Littleton, Colo:

Libraries Unlimited), 88–93. Reprinted from his *Elements of Library Classification*, 3rd ed. (Bombay, New York: Asia Publishing House).

ScaleDB. (n.d.). Big data and transactional databases: Exploding data volume is creating new stresses on traditional transactional databases [White paper]. Retrieved from http://www.scaledb.com/pdfs/ BigData.pdf

ScaleDB. (n.d.). Large database. Retrieved from http://www.scaledb.com/large-database.php

Teets, M. & Goldner, M. (2013). Libraries' role in curating and exposing big data. *Future Internet 2013, 5*, 429-438. Retrieved from www.mdpi.com/1999-5903/5/3/429/pdf

Villanova University. (n.d.). What is big data?. Retrieved from http://www.villanovau.com/university-online-programs/what-is-big-data/

WorldCat. OCLC Online Computer Library Center, Inc. (2013). A public-accessible interface to the database is accessible at http://worldcat.org. Accessed 29 October 2013.

WordNet. Princeton University. (2010). [http://wordnet.princeton.edu]. Accessed 29 October 2013.