Future Internet **2013**, *5*, 429-438; doi:10.3390/fi5030429

OPEN ACCESS

future internet

ISSN 1999-5903

www.mdpi.com/journal/futureinternet

Article

Libraries' Role in Curating and Exposing Big Data

Michael Teets [1] and Matthew Goldner [2,*]

[1] OCLC Innovation Lab; OCLC, 6565 Kilgour Place, Dublin, OH 43017, USA; E-Mail: teetsm@oclc.org

[2] OCLC Library Services Division; 6565 Kilgour Place, Dublin, OH 43017, USA

* Author to whom correspondence should be addressed; E-Mail: goldnerm@oclc.org; Tel. +1-614-764-6405.

Received: 21 March 2013; in revised form: 16 May 2013 / Accepted: 12 July 2013 / Published: 20 August 2013

Abstract: This article examines how one data hub is working to become a relevant and useful source in the Web of big data and cloud computing. The focus is on OCLC's WorldCat database of global library holdings and includes work by other library organizations to expose their data using big data concepts and standards. Explanation is given of how OCLC has begun work on the knowledge graph for this data and its active involvement with Schema.org in working to make this data useful throughout the Web.

Keywords: WorldCat; knowledge graph; library data; bibliographic data; authority data; Schema.org; OCLC

1. Introduction

Libraries have amassed an enormous amount of machine-readable data about library collections, both physical and electronic, over the last 50 years. However, this data is currently in proprietary formats understood only by the library community and is not easily reusable with other data stores or across the Web. This has necessitated that organizations like OCLC and major libraries around the world begin the work of exposing this data in ways that make it machine-accessible beyond the library world using commonly accepted standards.

This article examines how the OCLC data hub is working to become a relevant and useful source in the web of big data and cloud computing. For centuries libraries have carefully cataloged and described the resources they hold and curate. OCLC was formed in 1967 to bring this data together in electronic form in a single database. Today WorldCat, formed by libraries around the world, has over

300 million records of physical and electronic books and journals, recordings, movies, maps and scores with more than 2 billion holdings that describe which libraries, archives and museums hold and license these resources. However, these 300 million individual records remain isolated in the Web, requiring a restructuring of how this data is exposed for use throughout the Web.

The organizing of information has been carried out for centuries. "The idea that a library should provide the opportunity for study of the texts and a means to discover the original words of the authors, even those who had lived long before, first became the manifest of the library of Alexandria.… In the middle of the third century BC, Callimachus the poet, was employed here. Organizing the material and carrying on scholarly work at the same time.… Callimachus' so-called *Pinakes*, in which the works of the authors were organized in alphabetical order, was the first ever library catalogue." [1]

This work advanced over the next thousand years with small improvements on how these lists were maintained, but they were intended only as inventory devices rather than finding aids. It wasn't until the seventeenth century that printed book catalogs were introduced at Oxford and began to have alphabetical arrangements of authors to serve as finding aids [2]. The end of the nineteenth century saw the introduction of printed cards and the card catalog [3]. This was the dominant library finding aid for the next century until the first commercially available Online Public Access Catalog (OPAC) was introduced in 1980 by CLSI in the United States [4].

However, as David Weinberger points out in his work, *Everything Is Miscellaneous*, these catalogs were organized for an analog world [5]. Somebody had to decide how to organize the information, whether it be alphabetical by author or title, or using a subject thesaurus or classification scheme, which predetermined how materials would be organized and then located.

For the most part these were closed schemas disallowing reuse of parts of the metadata in other systems. There were exceptions, such as authority files of authors' names, which could then be linked to multiple cataloging records, but even then there was no common schema for linking to these files to open them up for broader use. So the question raised by OCLC Research and OCLC's Innovation Lab is whether this metadata about library materials has a role in the current world of cloud computing and big data, and if so, how is that role defined and who fills it?

Gartner defines big data as "high-volume, velocity and/or variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision-making and process automation." [6] This ties to Weinberger's concept of the Web as leaves that can be made smart and piled together to meet any specific users need. He states "Smart leaves are not like catalog cards with more room and an extra forth IQ points. … An identifier such as an ISBN that enables distributed information to come together when needed turns a C-student leaf into a genius." [7]

We are often asked, "Is library data really big data?" If you consider just the metadata representing the collection of printed and electronic works held by libraries, it really cannot be considered big data in its current meaning. [8] Even when you consider the full-text of those works, the data management does not require "big data" techniques.

However, when we look at these collections as a training set for all human knowledge, we can follow obvious data trails to generate massive collections of new relationship assertions. From a single work, we can extract relationships from co-authors, citations, geo-location, dates, named entities, subject classification, institution affiliations, publishers and historical circulation information. From

these relationships, we can connect to other works, people, patents, events, *etc*. Creating, processing and making available this graph of new assertions at scale is big data. It requires the source data to be structured in such a way that many can consume and operate on the corpus in common ways.

In 2008, *Nature* published an entire issue on big data. Authors Frankel and Reid [9] made the prescient statement:

> "To search successfully for new science in large datasets, we must find unexpected patterns and interpret evidence in ways that frame new questions and suggest further explorations. Old habits of representing data can fail to meet these challenges, preventing us from reaching beyond the familiar questions and answers."

This is precisely the issue being grappled with concerning library metadata. Though it is structured, it is not usable in a Web world of big data. This has led us to using linked data technologies as a primary enabler for interoperability at scale on the Web. In 2006, Tim Berners-Lee [10] recommended four rules—or "expectations of behavior"—that are useful in this discussion.

Early innovation in exposing library data as linked data has shown there is a useful place in the larger web of big data. Some examples are shown in Table 1.

Table 1. Examples of linked data.

Organization	Type of data
British Library	Bibliographic (item descriptions)
Deutsche National Bibliothek	Authority (names) and bibliographic
Library of Congress	Bibliographic
OCLC	Classification (Dewey Decimal), authority, bibliographic

In 2011, the British Library (BL) [11] announced the release of the British National Bibliography as linked data. Its purpose went beyond just encoding bibliographic data as RDF; rather, "...they set out to model 'things of interest,' such as people, places and events, related to a book you might hold in your hand."

The Deutsche National Bibliothek (DNB) [12] first released its authority data as linked data in 2010. Authority data are authorized name headings that are used by libraries to ensure the works of a single author are associated with each other. DNB followed this in 2012 when it released its bibliographic data, which describes specific titles and works, joining the BL in releasing the data that describes a major national library collection as open linked data.

In May 2012, the US Library of Congress (LC) [13] announced it also would pursue releasing its bibliographic data as linked data. This was part of a larger project "to help accelerate the launch of the Bibliographic Framework Initiative. A major focus of the project is to translate the MARC 21 format to a Linked Data (LD) model while retaining as much as possible the robust and beneficial aspects of the historical format." MARC (MAchine Readable Cataloging) [14] had been developed since the late 1960s to provide a method to encode bibliographic and authority data so it could be moved between systems; however, it was used only in the library community and is not appropriate for use in today's Semantic Web.

OCLC's own efforts to expose linked data have spurred several innovative uses of existing data and increased traffic from search engines. Over this same time period OCLC has released several datasets

as linked data. Starting in late 2009, OCLC released Dewey classification data and has extended this data over the years to include deeper levels of the classification scheme [15]. This was followed in 2009 with the release of the Virtual International Authority File (VIAF) as linked data [16], and then the 2012 OCLC release of WorldCat as linked data using Schema.org as the primary vocabulary. As shown at Datahub [17], this trend is rapidly accelerating as more and more library data is released as open linked data.

2. Problem Faced

The transition that OCLC and the library industry are going through is simultaneously a significant change and a return to the roots of the industry. Organizing the authoritative descriptions of library objects is the root of library cataloging and a well-understood concept by library staff. However, the industry participants have primarily interoperated by exchanging text strings encoded in industry-proprietary MARC record formats. Our systems have evolved collectively to consume, manage and reproduce these text strings. Authoritative text strings have been stored and managed separately and are somewhat loosely connected through applications to item descriptions. The transition to effectively operating with data at scale requires much more focus on accessible structures, persistent identifiers and a comprehensive modeling of the data under management. Linked data is the formalized method of publishing this structured data so that large data repositories can be interlinked using standard Web protocols.

3. Solution in Knowledge Graph and Linked Data

In order to manage authority of text strings, we must first think about the organization of top-level entities relevant to our industry needs. This organization is often referred to as an "upper ontology" or, more recently, as entities in a knowledge graph. A knowledge graph is a model of relationships between entities or objects in a given space. Given libraries' broad responsibility in organizing knowledge, the entities in a library knowledge graph must be quite broad and include the following key entities:

- **People:** Traditionally people who have formally published works, but increasingly we must account for all people whether they are writing, reviewing, publishing or simply using library content;
- **Items:** Physical items held within libraries such as books and media, but also electronic-only information such as e-books, journal articles and digital scans of real objects;
- **Places:** Geographic locations past and present must be maintained to understand context of published works. Political names are important, but so are more abstract names such as Northern Gulf States;
- **Events:** Events can account for grand events such as a public performance by a rock music group, and also minor events such as a user viewing a book text;
- **Organizations:** Organization entities encompass corporate names, publishers, political parties and bodies, and associations;
- **Concepts:** Subject classification systems such as the Dewey Decimal Classification system and Library of Congress Class system are well-known concept organization schemes. More free-form tagging systems also round out this category.

The models produced are being defined in RDF schemas. These schemas or vocabularies represent the most tangible output of the work in the knowledge organization space. Two recent efforts are attracting the most attention. The first is the Schema.org effort led by Google [18] and the library extension to this vocabulary initiated by OCLC and evaluated within the W3C Community Groups [19]. Some parts of these extensions are being integrated back into the core schema. The second effort is the Bibliographic Framework Transition Initiative launched by the Library of Congress [20]. The former focuses on simplicity and interoperability, including systems outside of the direct sphere of the library. The latter focuses on the transition from MARC21 to modern standards necessary to preserve the precision and detail required for library workflows that include cataloging, resource sharing, preservation and inventory management.

The selection of a schema, framework or vocabulary can be daunting, technically complex and fraught with emotion and politics within a given community. We have adopted the strategy of selecting broadly used models and developing accessible systems to achieve the most interoperability. It is not our strategy to select and prove a single schema is the best for all. We have selected the Schema.org vocabulary as the root of our data model as it provides an accessible core.

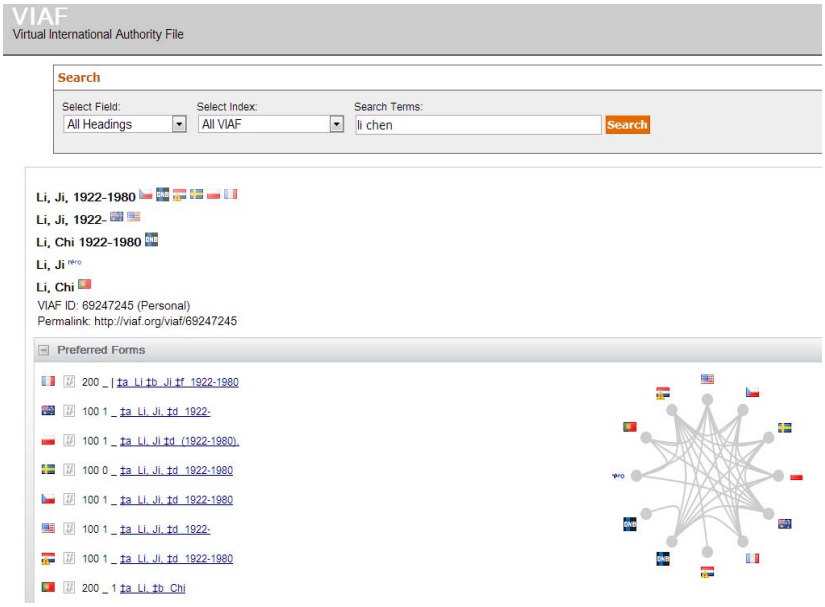
Schema.org was launched as a collaboration among Google, Bing and Yahoo! to create a common set of schemas to encourage structured data markup within Web pages. While initially viewed as overly simplistic, the organization has been very collaborative with established communities of practice to extend and generally improve the quality of their schemas. OCLC was an early collaborator focusing on items typically held by libraries. We worked with staff within Google to understand the possibilities and determine the best places to contribute. Some of our initial extensions were to recognize the differences between content and the carrier in published works. For example, when requesting or offering a movie within libraries, or anywhere on the Web for that matter, it is important to distinguish whether it is a digital download, a DVD, a CD or VHS. This is data well-structured and used in libraries but not immediately necessary for a movie review site. This ability to cross industries with similar data is met through common schemas that allow multiple industries to improve simultaneously [21].

Once primary entities are identified, three primary activities are required to make them usable. First, the entities must be modeled, but using existing schemas will reduce the work substantially. Modeling experts must work with expected contributors and consumers of the entities to understand their requirements. Those requirements must be transformed into conceptual and formal data models that govern how the data is stored and accessed. Secondly, the relationships between the entities must be defined in a similar manner. Finally, those entities must be given unique, persistent and accessible identifiers. And this is where we transition from just having lots of data to what is known as "big data." With accessible data and identifiers, the relationships can be auto-discovered by our systems. These new relationships can be turned immediately into action by these same systems.

The use of linked data in a larger knowledge graph, and its application in big data, is not just theory. It provides immediate and tangible benefits to the consumers of the data being managed. In our case, one of the most prominent examples is an author's name. While a person's name may seem obvious in text form, it can be quite complex in a global data landscape. A common problem to solve is to distinguish between authors of the same name. There could be hundreds of people named "Ji Li" who publish journal articles even within a specific technology discipline. The name can have multiple Romanized transcriptions of the original Chinese name. Birth and death dates, co-authors, countries of

publication and publishers can be included with the name to provide clarity. In reality, however, it is not until you get to a globally unique identifier for each that the problem can be mostly solved. Storing the identifier instead of the name, as in the Virtual International Authority File [22], and resolving at use allows disambiguation that would be impossible on any single data store (see Figure 1). Beyond disambiguation, it allows immediate updating should the author's information change. When Ji Li becomes Dr. Ji Li, or even completely changes his name, all systems using the authoritative identifier could be immediately current. These same benefits apply to geographic place names, titles, events, subjects, *etc*.

Figure 1. Author representation in Virtual International Authority File (VIAF).



Following from the author example above, a metadata record describing a creative work in a collection can now be expressed as links rather than text strings to be maintained. The following human-readable display of a "record" (see Figure 2) is actually not a record at all but instead is a series of entities with defined relationships. The entities have identifiers resolvable on the Web using standard http requests. We can see that Ji Li published a book in 1961 titled, *Beijing di 1 ban*. From the Library of Congress link through the VIAF identifier we can discover that Ji Li passed away in 1980 even though this specific reference was unaware of the death date. You can see how this method of data organization is critical as we move toward massive, globally accessible data stores.

Our first task was to select and model our primary data. Rather than attempt to model all of the data under management at OCLC, we chose to be opportunistic by modeling and exposing the data we viewed as important in WorldCat and nearest to the model exposed by Schema.org. We had a good handle on concepts with Dewey [23], LC Subject Authorities [24] and FAST subject headings [25], and have services to crosswalk between these and other classification systems. The new work was to make them accessible in the Schema.org vocabulary and embed this as html markup. We followed the same path for the data elements the library community is comfortable modeling that falls under the schema:creative work hierarchy such as: author, name, publisher, inLanguage, *etc*. See an example of schema description below (Figure 3). We applied these models across our own WorldCat catalog and we also made the data accessible to include in others' catalogs with the intent of improving library and

nonlibrary descriptions. We exposed this data directly in our production systems in html markup in the human-accessible systems and also in downloadable datasets. Because this was new to the community we serve, we made the data visible in a human-readable "view source" display in the bottom of each page we produced (seen above in the book example, Figure 2).

Figure 2. Human-readable display of a "record".

<http://www.worldcat.org/oclc/27168676>

library:holdingsCount	"14"
library:oclcnum	"27168676"
library:placeOfPublication	rdf:type: schema:Place schema:name: "Beijing :"
rdf:type	schema:Book
schema:author	<http://viaf.org/viaf/69247245> madsrdf:isIdentifiedByAuthority: <http://id.loc.gov/authorities/names/n81041682> rdf:type: schema:Person schema:name: "Li, Ji, 1922-"
schema:bookEdition	"Beijing di 1 ban."
schema:datePublished	"1961."
schema:name	"Wang Gui yu Li Xiangxiang"
schema:numberOfPages	"70"
schema:publisher	rdf:type: schema:Organization schema:name: "Xin hua shu dian jing shou"
schema:publisher	rdf:type: schema:Organization schema:name: "Ren min wen xue chu ban she"

Figure 3. Schema description.

Properties from CreativeWork		
about	Thing	The subject matter of the content.
accountablePerson	Person	Specifies the Person that is legally accountable for the CreativeWork.
aggregateRating	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.
alternativeHeadline	Text	A secondary title of the CreativeWork.
associatedMedia	MediaObject	The media objects that encode this creative work. This property is a synonym for encodings.
audience	Audience	The intended audience of the work, i.e. the group for whom the work was created.
audio	AudioObject	An embedded audio object.
author	Organization or Person	The author of this content. Please note that author is special in that HTML 5 provides a special mechanism for indicating authorship via the rel tag. That is equivalent to this and may be used interchangeably.
award	Text	An award won by this person or for this creative work.
awards	Text	Awards won by this person or for this creative work. (legacy spelling; see singular form, award)
comment	UserComments	Comments, typically from users, on this CreativeWork.
contentLocation	Place	The location of the content.
contentRating	Text	Official rating of a piece of content—for example,'MPAA PG-13'.
contributor	Organization or Person	A secondary contributor to the CreativeWork.

From the first public step, we set out on two parallel but complementary paths. One path took an internal focus; the other took an external focus. Internally, we began this formal modeling for the most important data resources under our management. This required modeling and a program of education among the consumers of this data within our product and engineering teams. Externally, we sought

partners that would augment our data, provide connectivity through modeled relationships, or would simply exploit this data within their systems. Momentum on both internal and external fronts has been increasing, and we find our data being used to create innovative relationships inside and outside our services. Library developers have begun to integrate our data into their services. Zepheira [26] is a Semantic Web company that has assisted both OCLC and the Library of Congress in implementing these concepts.

In the midst of our efforts to move to more modern data structures for interoperating both inside and outside the library community through data, the Library of Congress kicked off an effort to replace the MARC cataloging record structure with a modern solution. In many ways, LC followed a similar path as described above but with a more specific focus on the detailed information required in specific library services. The rich detail required at an item level has led LC to a more industrial-strength model optimized for its needs. OCLC was invited to the table early in this discussion and has been both a contributor and consumer of LC's efforts. Because we are both operating with accessible and machine-understandable data models, our collective efforts do not produce significant conflicts. When either group looks at an object from the other, it will see not only the string of text data but also the identifier of that object, the model under which it was created, and even the authority under which it is managed. Essentially, the data becomes interoperable at scale.

Using big data to automatically discover relationships is opening the doors for rapid innovation. Our systems could recognize that a PhD student at a small European institution is focusing on the same subject matter as a professor at a major US academic research organization. The systems could recognize that research in a specific area is rapidly increasing while an adjacent category is rapidly decreasing. Collaboration forums, collections and repositories could be spawned or dismantled without distracting the researchers from their primary task. In theory, better solutions to research problems can be found more quickly.

4. Conclusions

We stated earlier that OCLC and major libraries around the world need to expose the vast wealth of library collections data produced in the last 50 years beyond the library community. As pointed out by analyst Anne Lapkin in a Gartner report [27]:

> "Big data is not just about MapReduce and Hadoop. Although many organizations consider these distributed processing technologies to be the only relevant "big data technology," there are alternatives. In addition, many organizations are using these technologies for more traditional use cases, such as preprocessing and the staging of information to be loaded into a data warehouse."

In other words, just making big data sets accessible is not a desired end point. It is about making the data reusable in combination with other data sets across the Web.

Just moving to RDF alone could not accomplish this, leading OCLC to make the decision to work with an already accepted cross-industry vocabulary to improve access. OCLC and others involved are still early in this process. However, there is already strong evidence that once exposed, library data is useful to other communities and is accessed and repurposed. Because librarians have invested so much time and energy into authoritatively describing resources, their creators and subjects covered, this data

serves a valuable role in connecting the many facets of people, items, places, events, organizations and concepts into a meaningful knowledge graph.

As Weinberger said in *Everything Is Miscellaneous*, this work allows any searcher on the Web to pull together information based on relationships that are meaningful to the searcher, not based on a predefined organization created for an analog world [28]. By exposing massively aggregated library data in ways that make this possible, information seekers around the world will find items of interest in ways not previously possible.

Acknowledgments

The authors wish to thank Jeff Young, Software, Architect, OCLC, for his contributions to the article content, and Brad Gauder, Editor, OCLC, for editorial assistance.

Conflicts of Interest

The author declares no conflict of interest.

References

1. Eliot, S.; Rose, J. *A Companion to the History of the Book*; John Wiley and Sons: Hoboken, NJ, USA, 2009; p. 90.
2. Kent, A.; Lancour, H.; Nasri, W.Z.; Daily, J.E. *Encyclopedia of Library and Information Science*; Marcel Dekker: New York, NY, USA, 1968; Volume 4, p. 255.
3. Kent, A.; Lancour, H.; Nasri, W.Z.; Daily, J.E. *Encyclopedia of Library and Information Science*; Marcel Dekker: New York, NY, USA, 1968; Volume 4, p. 277.
4. Kent, A.; Lancour, H.; Nasri, W.Z.; Daily, J.E. *Encyclopedia of Library and Information Science*; Marcel Dekker: New York, NY, USA, 1968; Volume 58, p. 154.
5. Weinberger, D. *Everything is Miscellaneous: The Power of the New Digital Disorder*; Times Books: New York, NY, USA, 2007.
6. LeHong, H.; Laney, D. Toolkit: Board-Ready Slides on Big Data Trends and Opportunities. *Gartner*, 1 March 2013, G00238695.
7. Weinberger, D. *Everything is Miscellaneous: The Power of the New Digital Disorder*; Times Books: New York, NY, USA, 2007; p. 120.
8. Hessman, T. Putting big data to work. *Ind. Week* **2013**, *262*, 14–18.
9. Frankel, F.; Reid, R. Big data: Distilling meaning from data. *Nature* **2008**, *455*, doi:10.1038/455030a.
10. W3C. Design Issues Website. Available online: http://www.w3.org/DesignIssues/LinkedData.html (accessed on 8 May 2013).
11. Talis Systems Website. Available online: http://talis-systems.com/2011/07/significant-bibliographic-linked-data-release-from-the-british-library/ (accessed on 3 March 2013).
12. Svensson, L.G.; Jahns, Y. PDF, CSS, RSS and other Acronyms: Redefining the Bibliographic Services of the German National Library. Available online: http://conference.ifla.org/past/ifla76/91-svensson-en.pdf (accessed on 3 July 2013).

13. Library Journal. Info Docket. Available online: http://www.infodocket.com/2012/05/23/linked-data-the-library-of-congress-announces-modeling-initiative-contracts-with-zepheira/ (accessed on 3 March 2013).
14. Library of Congress. *Bibliographic Framework As a Web of Data Linked Data Model and Supporting Services*; Library of Congress: Washington, DC, USA, 2012. Available online: http://www.loc.gov/marc/transition/pdf/marcld-report-11-21-2012.pdf (accessed on 5 August 2013).
15. The OCLC Cooperative Blog. Available online: http://community.oclc.org/cooperative/2012/07/dewey-linked-data-interview-with-michael-panzer.html (accessed on 3 March 2013).
16. OCLC Website. Available online: https://www.oclc.org/en-US/news/releases/2012/201224.html (accessed on 3 March 2013).
17. Datahub. Available online: http://datahub.io/ (accessed on 12 August 2013).
18. Schema.org. Available online: http://schema.org/ (accessed on 12 August 2013).
19. Schema Bib Extend Community Group. Available online: http://www.loc.gov/marc/transition (accessed on 12 August 2013).
20. Bibliographic Framework Initiative. Available online: http://www.loc.gov/bibframe/ (accessed on 12 August 2013).
21. Godby, C.J. *The Relationship between BIBFRAME and the Schema.Org "Bib Extensions" Model A Working Paper*; OCLC Research: Dublin, OH, USA, 2013. Available online: http://www.oclc.org/content/dam/research/publications/library/2013/2013-05.pdf (accessed on 5 August 2013).
22. Virtual International Authority File. Available online: http://viaf.org/viaf/69247245/ (accessed on 12 August 2013).
23. OCLC Website. Available online: http://www.oclc.org/dewey (accessed on 3 March 2013).
24. Library of Congress Subject Headings. Available online: http://id.loc.gov/authorities/subjects.html (accessed on 3 March 2013).
25. OCLC Research Website. Available online: http://www.oclc.org/research/activities/fast.html (accessed on 3 March 2013).
26. Zepheira Website. Available online: http://zepheira.com/ (accessed on 3 March 2013).
27. Lapkin, A. Hype Cycle for Big Data, 2012. *Gartner*, 31 July 2012, G00235042.
28. Weinberger, D. *Everything is Miscellaneous: The Power of the New Digital Disorder*; Times Books: New York, NY, USA, 2007; pp. 94–96.

© 2013 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).