Disruption or revolution? The reinvention of cataloguing (Data Deluge Column)

Donna Ellen Frederick

The Data Deluge Column previously looked at the changing roles and functions of the metadata specialist as well as the ongoing shift in the nature of library data. This installment of the column will discuss revolutionary developments in technology and look at five changes in the world of cataloguing and library data which the author argues are the catalysts for significant change for libraries in the near future.

For the past five years or so, the author has been studying disruptive change in library technical services. She uses Christensen's (1997) model of disruptive change to identify changes which are disruptive and Lucas' (2012) "survivor model" to analyze how well libraries are adapting to disruptions. Christensen's model suggests that disruptive technologies are typically smaller, easier to use and more affordable than the traditional product and that these characteristics make them more accessible and suitable to a wider population, which often makes the innovation threatening for the sustainability of the traditional industry leaders. While the model did not immediately appear to apply to libraries and their collections, after the author did a bit of reading and thinking about the topic, she began to see that eBooks have been disruptive to both library service and book publishing. She was able to, for example, identify the ways in which eBook package purchases create multiple disruptions in the traditional work of technical services. The disruptions she identified were so numerous that discussing all of them was well beyond the scope of a single paper. She took much of what she learned from her analysis and used it as the basis for writing her 2016 monograph *Managing eBook Metadata in Academic Libraries: Taming the Tiger*. In this book she points out to readers, examples of disruptions in selection, acquisitions, cataloguing, discovery and access of resources, preservation and other common technical services activities. There is no question that library workers have had to retool their skills and practices to deal with eBooks. If libraries, for example, were to attempt to continue to use their old practices and workflows, they would have been completely overwhelmed by the sheer volume of titles in eBook packages and the fluidity of content. Her use of Lucas' model helped to identify the major pain points which were making it difficult for many libraries to make appropriate adjustments in response to the introduction of the large eBook packages. Directing the attention of libraries toward these pain points and posing questions for the librarians to help them chart a new course toward adapting to the change are central to the overall approach taken in the book with regard to laying out a plan for effective eBook metadata management. The key in framing the questions and discussion was to view various aspects of eBooks as disruptive to library practices.

When the author attempted to apply the same process with regard to identifying disruptions in the larger and more general context of library data and its creation, she soon found that the various elements in Christensen's model did not apply. The model did not seem relevant. While it was easy to identify a new metadata environment based on a "linked data" model as the goal that cataloguers and metadata librarians are moving toward, analysis using Christensen's model beyond that point started to fall apart. Can, for example, a person talk about linked data as being "smaller and easier to use"? With eBooks, a person can imagine a number of large volumes downloaded onto an eReader. The eReader could be slipped into a relatively small bag, and then the user could scan through that content to quickly and easily find certain content. Imaging what it would be like to do similar things with print books helps to illustrate how eBooks can be described as "smaller and easier to use" despite that not all readers would agree that eBooks can be described as "easier to use" in all circumstances. But with linked data, it is not something that human beings interact with directly. Books and eBooks are technologies. Linked data is not a technology but a model or a way of organizing data which relies on a number of technologies to implement. Different technologies can be utilized in different ways in the process of implementing linked data, but the linked data itself is not something that the human mind or body can interact with directly. Linked data's quality of being physically imperceptible makes it very difficult for human beings, as creatures who rely on our senses for learning, to understand and discuss. Instead, what we often do is talk about the technologies which could be used when implementing the model. It is particularly confusing that many of the technologies that would be required to implement the model in libraries, such as BIBFRAME or a linked data-native library management system, do not yet exist in a complete and robust form. Therefore, average librarians cannot get the chance to

observe examples of the technologies operating in a real-life context. Without being able to see the technologies in action and in a familiar context, it is harder for many to understand the model and how it functions. Because it is difficult to imagine the complexity of how these new systems will operate in real-life contexts, even those who are experimenting are likely to not see the full ramifications of the upcoming technological change. Those librarians who are experimenting with and testing the new technologies are only getting a tentative glimpse of what is yet to come. As a result, some of the technologies that are being developed and tested today are likely to be replaced once, for example, librarians get some experience with implementing systems which are less reliant on locally stored data and much more reliant on linked open data on the Web.

Essentially, the author's attempt to analyze linked data as a disruptive technology revealed to her that it is important to differentiate between real things (technologies) and ideas (models) and that ideas in and of themselves cannot be disruptive. Instead, they inspire and give shape to the development of new technologies. It is only when an "idea" is applied through creative processes to innovate a new technology that the potential for disruption can arise. However, upon further reflection, she realized that a piece of the equation was missing in her analysis. One day she saw a photograph from the Second World War era of a family seated around a very large radio. The sheer size of the radio struck her as interesting. It likely only consisted of receiver, a speaker, some controls and, perhaps, a large battery. Today, these components could theoretically be so small that a human being could not use them. But why were they so large in the 1940s? Were radios status symbols? If they were, it makes sense that customers would have preferred a radio with a large cabinet. Perhaps that is part of it. However, the large size is likely partially because of the fact that the cabinet was likely full of the vacuum tubes which were in use previous to the transistor radio era. These tubes were significantly larger than today's electronic components and generated heat, so they required space for air flow. When transistor radios came into the general consumer market, the big radios and the tubes that used them were undoubtedly the victims of a disrupted household radio manufacturing industry. The large radio became "old technology" and largely fell out of style. To stop the analysis of the change at that point, the author realized, is premature. The reality is that the invention of the vacuum tube in the early twentieth century and their replacement with transistors around mid-century is part of the evolution of electronics. Electronics were born out of the idea that controlling electrical energy as it passes through certain types of materials creates an environment where signals can be used to achieve tasks. The way in which electronics developed over the twentieth and twenty-first centuries resulted in a succession of disruptive innovations and associated change as, for example, vacuum tube machines gave way to solid state and analog to digital. Ultimately, the driving force behind all of this change in technology was the idea of electronics and how they work. Disruption occurred and the materials and processes for implementing the idea were improved and refined. The idea behind "electronics" itself was not disruptive. Instead, electronics theory was a set of ideas that revolutionized machines, created new industries and changed lives.

In the twenty-first century, we now find ourselves dependent on electronics, yet few of us understand the theory behind electronics. We have an easier time understanding other technologies such as the internal combustion engine. We can understand that an explosion in a chamber can drive a piston which turns a wheel which can drive some other mechanical part such as another wheel or belt which, in turn, supplies power to machines. However, most of us struggle trying to imagine what is going on inside of our laptops. We know that if the battery runs out and we do not have our electrical cords, that the laptop would not work. We know that electricity is critical to the functioning of the machine, but few of us understand how that electricity is used by the laptop and its various parts. So, while we understand the technology of the laptop and can interact with it so that we can use it to accomplish work, most of us do not understand the theory of electronics well enough to think about what is happening inside the machine when we use it. While we know what electrons are and we have an idea about how they behave, making sense of how a complex electronic device such as a laptop works is a baffling task. While the author occasionally wonders about how certain electronic devices work, she generally does not think too much about electronic theory. The more important issues for her are generally whether or not she has the device she needs and if she knows how to use the features.

So, what is the connection between electronics theory and changes to the model of library data? The most significant connection is that these both exist in the realm of theories and that technologies are based on them. Based on what is known about the development of electronics over the past century, we can also expect to see a succession of disruptive technological change as the model of library data is applied over time. The second connection is that the ideas themselves, rather than the technologies that develop from them, are revolutionary in nature. In the column "Libraries, data and the fourth industrial revolution", there was a discussion of an industrial revolution which occurred in the twentieth century and led to our current "digital age". This "revolution" was essentially brought about by the ability to put into action the thoughts behind the theory of electronics. The column also stated that we appear to be at the

close of one industrial age and at the beginning of a new revolution. The author suggests that the discussion of how library data must change in the previous Data Deluge Column "Library data in transition" fits into the larger cluster of changes which is sometimes called "the fourth industrial revolution". Considering, as already discussed, that the average person has a hard time understanding the theoretical underpinnings of the electronic technologies that are behind the third industrial revolution, it is not surprising to find that it is extremely difficult to conceptualize the "thoughts" behind the new model of library data. We know that the fourth industrial revolution is characterized by the increasing use of technologies such as robotics, artificial intelligence, machine learning and cloud computing, etc. But, these are technologies and not models or theories. What is the "big picture" of the change? What ideas are driving it? The author has spent several months paying attention to articles, news reports and videos about the current state of technological change, but she has yet to come across anything that outlines, in a definable way, the overarching model or theory behind the fourth industrial revolution. Sometimes she feels as if she is getting close but then this understanding leads to, yet another, description of specific technologies. Perhaps an understanding of the model exists for those most heavily engaged with making the technologies work together but it has not been given a name yet. Based on her observations, the author feels that the key elements of the new model have to do with a mass of well-ordered data that is accessible on the Web, interconnectivity (of people and devices) and artificial intelligence (AI). The Internet of Things (IoT), technology embedded in the human body, robotics, machine learning and many other emerging technologies seem to largely depend on one or more of these elements. In the end, perhaps our current experience with electronic devices demonstrates that the average person will be able to function in the new information environment despite not understanding how and why everything in that environment works. The bottom line is that, as discussed in previous columns, it appears that society is headed toward a new industrial revolution and that libraries will be profoundly impacted. The role of libraries in the revolution and their place in society will center on the data it has created over the decades, the reconfiguration and optimization of that data and the creation of new high-quality data.

So, if we know that change is coming to libraries and that the change is driven by something that is not a disruptive technology but a revolutionary change in how libraries think about data and how it is organized, and that it is difficult for us as human beings to think about something that abstract, how do we make sense of what is happening and prepare for the changes as they happen? Perhaps there is no simple and straightforward answer. However, in studying the process of change in both the conceptualization of library data and the technique of creating it, the author has identified five particularly significant changes. A practical approach may be that librarians begin to learn about these changes and think about how they relate to, impact upon and could shape their speciality in the near future. The next section will look at these five shifts in how librarians think about and create library data.

1. A new model for cataloguing: adoption of new cataloguing principles (December 2016) IFLA's statement of international cataloguing principles

Up until very recently, all cataloguing was based on what is known as "the Paris Principles" which were formulated after an international meeting of cataloguers in 1961. In 2009, a new statement of principles was released by the International Federation of Library Associations (IFLA) which essentially updated the original principles to reflect the diversity of twenty-first-century media and formats as well as providing guidelines for information searching and retrieval in electronic environments. While the principles were "modernized" eight years ago, the basic ideas behind the where, why and how of cataloguing remained the same as they were in 1961.

The new document can be found at: www.ifla.org/files/assets/cataloguing/icp/icp_2016-en.pdf. These new principles are the result of the extensive international discussion and critical analysis of a draft which eventually led to the dramatic reinvention of cataloguing principles. When fully implemented, they will change the world-view of cataloguing and metadata work. However, while the principles will be transformative, they also reflect shifts in thinking and practice which have occurred over the course of the past two decades or so. For example, since the time Karen Coyle (2000) wrote about the "changing nature of library data" and "data-fying the data", cataloguers and metadata specialists have been gradually shifting their view of cataloguing as creating and maintaining records to creating and maintaining data. The traditional catalogue record structure has been actively falling away in the past 4 or so years and is expected to largely disappear as many libraries transition away from relying primarily on MARC data for discovery. The new principles both reflect the ongoing changes in the field and current realities, while they are also intended to serve as a guide for the development of future library metadata schema, cataloguing guidelines and information search and retrieval systems.

2. Library Reference Model (November 2016)

In the era shaped by the Paris Principles, cataloguing was largely a "rule based" activity. The expert cataloguer could essentially memorize rules and apply them in a somewhat mechanical way. Of course, there are exceptions to this characterization in areas such as the cataloguing of serials, music and manuscripts and rare books. However, for the most part, cataloguing a published monograph involved processes which could largely be

memorized and involved relatively little creative or critical thinking. Over the past 20 years or so, there has been an increasing recognition that an overly simplified rule-based approach to cataloguing, while highly efficient and cost-effective, was not leading to the creation of the type and quality of metadata required in today's information environment. Several non-MARC library metadata containers such as Dublin Core and MODS also have also been tried and largely failed in terms of achieving revolutionary change. In light of a number of widely discussed papers, many in the cataloguing community came to believe that a theoretical model of bibliographic information needed to be created and used by cataloguers in place of the traditional rules. According to this line of thinking, a model needed to be in place before schema, standards or guidelines could be built. Rather than doing rote work, cataloguers would have a basis upon which to make effective decisions about how to create the best-quality bibliographic data regardless of the schema they use. This model would not be limited to use in, for example, MARC or Dublin Core, but is intended for any situation where libraries create metadata for discovery purposes.

The earliest attempt at the creation of a new theoretical model was the Functional Requirements for Bibliographic Data (FRBR), which was, in turn, adopted by the international body which developed Resource Description and Access (RDA). RDA became the preferred and predominant descriptive cataloguing approach for academic and national libraries as of April 2013.

As more libraries adopted and applied RDA, they soon began to bump up against its limitations. Even with twice annual changes and updates to RDA, it was evident that there were limitations that could not be overcome because the theoretical model upon which it was based has limitations. Therefore, in 2015, a process to update FRBR was sought. The Library Reference Model (LRM) was presented to the cataloguing community for consideration and review. In the fall of 2016, LRM was accepted by IFLA as the replacement for the original FRBR model (see www.ifla.org/files/assets/cataloguing/frbr-lrm/frbr-lrm_the replacement for the original FR20160225.pdf). However, it is also recognized in the cataloguing community that LRM is still likely not robust enough to handle all types of information and resources which require the creation of metadata. For example, the PRESSoo model will likely be required for continuing resources (serials and integrating resources) and FRBRoo for three-dimensional objects, digital files and artworks. Therefore, the acceptance of LRM is seen as just a step in the evolution of library data/metadata.

While RDA cataloguers now generally understand and apply FRBR in their work, LRM has presented us with new concepts and terminology. Because theoretical models are abstract, the learning curve tends to be slow and non-linear. Therefore, it is recognized that with the acceptance of both LRM and the new cataloguing principles, cataloguers are faced with yet another significant amount of learning which is not unlike, but much more far-reaching than, the task we faced in 2013 when RDA was put into general use.

3. 3R Project (began April 2017)

The key tool that cataloguers use to create RDA data (i.e. to catalogue) is called the "RDA Toolkit", which is a subscription service containing RDA instructions, community-specific guidelines (e.g. Program for Cooperative Cataloguing policy guidelines, various national library guidelines, music cataloguing options), examples and links to related resources such as the Metadata Registry. Because RDA is much more complex than AACR2 and is updated twice annually, it is not practical for cataloguers to attempt to learn RDA in the way that they learned (memorized) AACR2. In addition, libraries which attempted to create local policy manuals which include RDA instructions or created "cheat sheets" for paraprofessional staff often found that it was nearly impossible to keep their documentation up-to-date because of the ongoing changes in RDA. This reality has more or less forced that majority of cataloguers to depend on the RDA Toolkit. Unfortunately, it is very difficult to navigate the RDA Toolkit. It seemed that with each revision to the RDA guidelines and with each addition of new community guidelines (e.g. for music librarians, audiovisual cataloguers, serials librarians, etc.), the complexity and confusion experienced with the Toolkit was getting worse. By early 2016, there was no question that the cataloguing community was feeling considerable pain because of the general inability to keep up with all of the changes and to use the RDA Toolkit in an effective way.

With IFLA's decision to adopt LRM as the primary conceptual model for cataloguing, it soon became apparent that this would have an impact on RDA. Essentially RDA would need to be reviewed from the bottom up (i.e. what happens if you replace FRBR with LRM) and then revised. Knowing that the RDA Toolkit requires rethinking and revision to make it more usable for cataloguers, the need to review and potentially rewrite many existing RDA instructions seemed to be the ideal time for a major overhaul of essentially everything related to RDA. This work came to be known as the 3R Project (RDA Toolkit Restructure and Redesign Project). For more details about the project, see www.rda-rsc.org/3Rprojectupdate

The plan for the 3R project is to freeze the development and updates in the Toolkit between April 2017 and April 2018. This will give the publisher of the Toolkit time to completely rebuild the service while the RSC (the international body which governs RDA) can essentially rewrite RDA. In the meantime it is expected that libraries will learn LRM so that they will be ready to learn the new RDA when it is expected to be released along with the new RDA Toolkit in 2018.

The whole issue of the RDA Toolkit and its importance to cataloguers is significant. Never before have cataloguers been so dependent upon a service such as this. In addition, the fact that the 3R Project is so badly needed is an indication of the nature of the shifting sands and the reality is that it is nearly impossible for the average librarian to become an "expert" in the

new cataloguing practice because it is so complex and fluid.

4. Virtual library data (cloud-based)

There is an increasing movement within the cataloguing and metadata community to prefer a new model of creating, storing and sharing metadata which practically eliminates the need to store and manage records locally. In this model, metadata is created by an "expert", validated according to various international standards and stored in a virtual location. The metadata can be corrected or enriched by authorized persons or organizations, but once it is validated and released into the cloud, it essentially becomes the official version of that information which is shared by the entire world. An increased use of authority data which can link multiple controlled vocabularies and handle multiple languages and scripts or characters nearly eliminates the need to duplicate data for different audiences.

In this environment, libraries would not manage "records" locally but would contribute and enrich data in the virtual environment. While this means that the type of cataloguing libraries typically do today will disappear as will the type of local records we are accustomed to creating and managing, it also means that a greater responsibility will be placed on local libraries to create the most accurate and highest quality metadata possible because what they create will be used by the entire world. Even with the smaller scale experiments that OCLC is carrying out, it is showing that cataloguing done by someone who is not fully trained and/or does not adhere to standards does not perform as desired in virtual environments. Metadata that looks and performs perfectly in a local system, and may have even been "tweaked" to do so, can become a profound failure in these large environments which require a higher level of compliance with standards. This is why organizations such as OCLC have recently implemented a required validation process for libraries who contribute data to WorldCat, for example.

Thus, virtual data requires two levels of expertise. The first level of expertise is with regard to the subject itself. Metadata creators must be experts in the subject for which they are creating data (i.e. person, place, thing, publication, etc). Or, the cataloguers must have the ability to do the necessary research to build or gather the required expertise to create authoritative data (metadata). Library of Congress' Name Authority Cooperative (NACO) has an existing manual which helps cataloguers who need to do this sort of research. The second level of expertise is with regard to international cataloguing standards and community-specific guidelines. The latter is particularly significant because there is an increasingly recognition that certain types of resources or information can be made discoverable only when they are described in a way that is appropriate to their category. In the past three years, various "communities" within the larger cataloguing community have been creating guidelines which define what is appropriate to do when creating metadata for specific categories of resources. For example, a map needs to be described as a map and not as if it were a book or musical recording. Many libraries, for example, have applied traditional generic monograph cataloguing to diverse resources. However, this approach does not scale well. Libraries who have taken a very generic approach to cataloguing have found that as their collection grows, they are unable to make key differentiations during their search processes. Their discovery interfaces may technically have the capacity to refine searches to retrieve different types of resources, but those features will not work if the data to support them are not in existence or are not formatted according to a generally recognized and accepted standard. The problem gets even worse in virtual discovery environments where multiple metadata formats and controlled vocabularies may be in use. This means that the ability to effectively apply community-specific guidelines (e.g. music, serials, manuscripts, etc.) will become increasingly important in the near future. Cataloguers must, first, be able to identify when a community-specific guideline should be used and, second, must have the training to know how to apply it properly. While large and specialized libraries have used guidelines such as these for many years, it has recently become important for all academic libraries to learn and apply them. This is part of the reason why the RDA Toolkit has become important and many libraries are gradually letting go of their local cataloguing policies in favor of internationally accepted guidelines.

Examples include:

- OCLC's experiments: www.oclc.org/research/publications/2015/oclcresearch-library-linked-data-in-the-cloud.html
- VIAF and ISNI: www.loc.gov/aba/pcc/documents/PoCo-2016/VIAF-ISNI-position-paper.pdf

5. Linked (Open) data for libraries

This is essentially a philosophy for sharing library data which embodies the practices described in the previous section. The idea is that information discovery and retrieval can be revolutionized through this approach. If data is hosted in the cloud and that data conforms to international library standards or other recognized standards, it can be retrieved and reused in various ways by various applications. This is where the power of the Web brings long-standing principles and practices of information science to life. Libraries can create richer and more responsive discovery environments; they can make use of high-quality information supplied by organizations outside of the library community; and libraries can make their high-quality data available to organizations and services globally.

For the dream of this new reality in information discovery to fully materialize, the ideas discussed in the previous four points must fall into place. In fact, points one and two were largely created to support the movement of library data from existing local databases and into a new environment on the Web. Services such as the National Library of Medicine's subject headings or MeSH (https://id.nlm.nih.gov/mesh/) and OCLC's WorldCat (www.oclc.org/developer/news/2014/learn-more-about-worldcat-works.en. html_) are already heavily invested in the creation and hosting of what is called linked open data. Large

national, academic and specialized libraries around the world have been gradually following this movement by upgrading their data to the new international standards and releasing it to the Web for use.

What has been learned about linked data in recent years is that organizations must be identified and recognized as the official providers of certain types of data. Individuals or groups outside of that organization can have a way to add or update that data, but the official provider must ultimately remain responsible for its quality and accuracy. To avoid muddles, non-authorized persons or organizations should not recreate the data elsewhere on the Web. To avoid unnecessary work and the endless treadmill of updating local data, the idea of "cleaning" and "loading" data locally also should be abandoned. The idea of "reusing" metadata needs to essentially fall away and be replaced with a new model where libraries use the data as it is found in an official source. If corrections or enrichments need to be made, that should be done through the official provider or one of its authorized partners. The official providers must use an approach to creating and making the data accessible which includes documentation that is openly accessible in a recognized location and that documentation should include data which can readily be processed by machines. Putting this sort of coordination into effect among libraries is a considerable challenge. Libraries have recognized that there are non-library organizations whose cooperation is required as well. The diversity of organizations which ideally should work together makes the challenge even greater. For example, for the new vision to be most effective, ONIX (publishing industry) metadata, ISSN data, ISBN data, D.O.I. metadata and ISNI identifiers all need to meet certain minimum standards. In some cases, such as ONIX, the industry that creates and primarily uses the metadata may not have any immediate need for upgrading, standardizing or enriching their data to the same level of "perfection" as libraries require. International library bodies are currently working with a number of non-library groups to promote the value of putting in the extra effort. In some cases such as ISSN information, the data is currently not freely available on the Web. Therefore, it is recognized that considerable discussion and negotiation needs to occur before libraries can attain the ideal conditions for fully realizing the vision. That being said, there are an increasing number of libraries and services which are making use of open linked data to create new discovery experiences and to reach new audiences.

Finally, national libraries, the Library of Congress and international library cataloguing governance bodies have made various statements to the effect that, unlike RDA which had an official date after which it took effect, the implementation and use of linked open data in libraries will be gradual and uneven. At this point, it is essential for academic libraries in particular to understand the bigger picture of the linked open data movement and also start to define their areas of expertise and their role in creating and maintaining information in this new environment.

In conclusion, cataloguing and metadata librarians have said in recent years that their discipline is being reinvented. Given the interest that many in the library profession have in shifting traditional library data and siloed discovery environments to a new reality that makes use of linked open data on the Web, it is not surprising that a "reinvention" process is underway. While it is difficult to grasp and discuss the full significance of the change that is ongoing in libraries, the author hopes that her introduction to five of the major events in the area of cataloguing and metadata creation helps readers begin to appreciate the nature of the revolution that is still in its early stages.

REFERENCES

Christensen, C. (1997), *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*, Harvard Business School Press, Boston, MA.

Coyle, K. (2000), "Changing the nature of library data", Library Technology Reports, pp. 14-29.

Lucas, H. (2012), *The Search for Survival: Lessons from Disruptive Technologies*, Praeger, Denver.

Donna Ellen Frederick (*donna.frederick@usask.ca*) is Metadata Librarian at University of Saskatchewan, Canada.