RDA, Linked Data, and the End of Average

Philip E. Schreur

Stanford University

May 5, 2017

In the 1940s, the United States Air Force had a serious problem. They were moving rapidly into the jet age but still were having many crashes, on their worst day 17 crashes in a single day. At first these crashes were assumed to be attributable to pilot error because there had been no equipment malfunctions. However, the first ever cockpit had been designed back in 1926. Engineers measured hundreds of men and came up with the perfect, standard measurement for cockpit size: measurements included seat shape, the distance to the pedals, even helmet shape.

In the 1950s, they re-measured 4,000 pilots to see if those dimensions changed. They took into account ten different parameters to come up with a new average. However, when they looked at the measurements of individual men, 0% had those average dimensions. And when they took into account only three dimensions, just 3.5% of the men had those new dimensions. The solution was to realize that there was no average man and to make everything in the cockpit flexible enough to adjust to the pilots' individual sizes.

About the same time, Dr. Robert Dickinson and Abram Belskie measured 15,000 women to come up with the perfect average, or "ideal" dimensions of a woman. She was given the name of Norma, and she became so popular that the Cleveland Health Museum began selling small statues of her, she appeared in Time magazine, her dimensions were read allowed on the TV show "This American Look" so women could measure themselves to see how well they matched the ideal. In 1945, there was a contest to see which woman best approximated this

standard. Less than forty women out of 3,864 contestants matched even five of the nine dimensions. Apparently, there is no ideal, or averaged size woman eithr.

In 1967, ALA published the Anglo-American Cataloging Rules. According to the introduction, description should be based on a perfect, or ideal, copy. Chapter 1 is on Entry: the chapter first gives general rules for what is most common and they are followed by special rules for exceptions such as law, and religious publications. Chapter 2 is on Headings for Persons: once again, the chapter gives general rules for what is most common and they are followed by special rules for foreign languages. But what is perhaps most telling is that the rules are divided into three parts: Entry and Heading, Description, and **Non-Book Materials.** The ideal format is the book, everything else is, well, just something else. And everything is assumed to have a physical dimension, a standard number of authors, a standard title placement. Any deviations from the standard must be noted with information as to the exception such as "cover title." Even the style of the catalog card was dictated:

> "A uniform style is adopted for all catalog entries, covering spelling, capitalization, punctuation, abbreviations, use of numerals, indentations, and for printed cards, [even] type faces."

In 2015, Todd Rose published a book called *The End of Average: How We Succeed in a World that Values Sameness*. In his book, Rose documents the invention of the idea of average as the ideal, starting from the writings of Quetelet in the early 19th century. For the first time, Quetelet used the scientific technique of taking multiple measurements to establish the ideal average measurement, to Social Sciences. By taking advantage of the large number of data sets of measurements of human dimensions made during the middle of the 19th century, Quetelet established the idea of the "average" size of the human male, in this case, average meaning "ideal." We still use his index today in the form of the body mass index, or BMI. All variations from this average were considered aberrations from the ideal. This concept of average as the ideal became an epidemic in the 20th century:

- We standardized and automated production lines in factories
- We standardized educational instruction for the average student

- We created standardized entrance exams for higher education
- We created standardized intelligence exams
- We created normal development patterns for infants, any variation from which was considered an indication of a possibly abnormal child
- And we created ideal metadata descriptions of standard library materials

In his brilliant book, through a number of detailed analyses, Rose demonstrates that there is no average for complex patterns of behavior or, by extension, complex objects. These complex objects or behaviors are made up of individual elements that vary widely.

- There is no average intelligence
- There is no average way to learn
- There is no ideal body shape
- And there is no standard library resource

From October to December in 2010, Stanford enthusiastically became part of the early adopter testing of RDA. By that time, we had become committed to linked data as our path forward. We were convinced that linked data would be the way to bring our data to the web in an intelligible way by:

- allowing us to integrate our metadata across institutions in a way we hadn't been able to in the past
- and allowing us to make use of the growing masses of information on the web

And RDA, in conjunction with linked data, looked to be the key to making this happen. As opposed to other testing institutions, we trained our entire department in the use of RDA from the very beginning. And we made two early decisions in the context of RDA and linked data:

- first, we did authority work and made added entries for all creators associated with a resource no matter how many, 5, 10, 15 … we knew that eventually we would need identifiers for all those entities
- and second, we added roles for all persons and corporate bodies so that when the data was converted to linked data from MARC, we would know in what context the creator was related to the resource

There are two other aspects of RDA that I'd like to stress that make it so well adapted to our new environment, and the first is …

Flexibility.

My greatest challenges have always come from our university librarian, Michael Keller. Not only did he want to see the libraries make a transition to linked open data, he wanted us to capture ALL THE INFORMATION THE UNIVERSITY CREATES. Just let that sink in a minute: this would include art, music, data sets, reading lists, perhaps even football scores. Although linked data opens the possibility of linking any sort of data, it needs a context to be put in for discovery, and AACR2 was not going to be that construct. What I needed was something flexible enough to accommodate an every widening family of resources and data types.

I rarely quote from RDA but I'd like to take the time for a few here:

- 0.0 Purpose and Scope: RDA provides a set of guidelines and instructions on recording data to support **resource discovery**
- 0.1 Key features
  - RDA provides a flexible and extensible framework for the description of resources …
  - RDA is designed to take advantages of the efficiencies and flexibility in data capture, storage, retrieval, and display made possible with new database technologies

- 0.2 Conceptual models
    - The use of FRBR, FRAD, and FRSAD for "the flexibility and extensibility needed to accommodate newly emerging resource characteristics
- 0.4.2.3 Flexibility
    - The data should function independently of the format, medium or system used to store or communicate the data. They should be amenable to use in a variety of environments

And I could go on … This emphasis on extensibility and flexibility should provide us with the key elements we need in moving forward in a rapidly evolving, international information space.

At the same time, Stanford is trying to outsource the creation of metadata for its traditional resources. We have outsourced cataloging for Italian materials, French, US/UK publications, South American imprints, Arabic materials … the more sources we can find the more we will purchase. And this will leave us with time for an ever-growing amount of non-traditional materials in need of metadata, at least non-traditional to our cataloging department. Not all will need full level cataloging, but they will all need descriptive metadata for discovery and the more consistent a model we can provide the more consistent discovery will be. I'm not quite sure how flexible RDA can be, but we will certainly be pushing at the edges in our work over the next number of years.

And the second great area is RDA's emphasis on relationships. Linked data itself is all about relationships. I think we often conceive of linked data as simply linking things together, but it is how those things relate to each other that is key. In RDA, an amazing number of chapters (chapters 17-37) are devoted to relationships:

Relationships between

- A work, expression, manifestation and item
- Relationships that are used to find works
- Relationships used to find related works

- Relationships used to find related person, families, or corporate bodies
- Relationships used to find relate concepts

RDF may seem like a non-library standard of the W3C that we are adapting for our work but I think that this is a very narrow viewpoint on our part. One of the many areas librarians have a brilliant understanding of is relationships and how they affect discovery. For instance, there is the issue of works. Early on in Stanford's work with BIBFRAME, we had a lengthy disagreement with the National Library of Medicine on how to define a work. OCLC has created work records from their database. For their next BIBFRAME pilot, the Library of Congress will be extracting BIBFRAME work records from their legacy data. Currently, the Biblioteque nationale de France is working out how to extract work records from their InterMARC data. The German libraries have a consistent representation of their data in RDF, including works. Do I think we will ever agree on the definition of a work, BIBFRAME of otherwise? It seems very unlikely. However, If we all can define what we mean by a work, and describe the relationships between these various conceptions, we can relate the data in a useful way for discovery. This understanding of the key position of relationships and how they should be defined is a true forte of librarians. And as we look to how we can make linked data our own, how we can give back to the broader linked data community, it's in this area of relationships that we can make our most dramatic impact.

Stanford's main foray into linked data is a Mellon funded grant called Linked Data for Production, or LD4P. LD4P is a collaboration between six institutions (Columbia, Cornell, Harvard, Library of Congress, Princeton, and Stanford) to begin the transition of technical services production workflows to ones based in linked data. This first phase of the transition has three broad areas of development:

- first, the ability to produce metadata as linked data communally,
- second, the enhancement of BIBFRAME to encompass the multiple formats that libraries process,
- and third, the engagement of the broader academic library community.

As the LD4P team looks forward, we see five major outcomes of this project. First will be the development of the ability for libraries to work in an open, networked environment in the construction of their metadata, allowing for more immediate and simple exchange of data. Up until now, we have worked in the isolation of our ILS systems and uploaded our data to a single monolithic data store to share. The new world based in linked data will be both more transparent and immediate. As we work through what it means to do our jobs in this new networked environment, we anticipate raising a number of interesting question such as:

- the use of traditional authority records versus identifiers, are both necessary? In what circumstances?
- the definition of what it means to share metadata for commonly held resources in a recordless environment.
- the assurance of data quality and provenance that has been supported in the past by the Program for Cooperative Cataloging.

A second major outcome will be the extension of the BIBFRAME ontology in four key areas: performed music, cartographic materials, archival film, and rare books. A third major outcome will be multiple open source tool development for use in metadata creation and transformation in a linked open data environment. Fourth, a key part of the LD4P Program will be the engagement of LD4P with other strategic linked-data projects through the LD4P Program Manager. And last, LD4P will proactively reach out to the international community to help develop a cohesive library perspective on this transition to linked data.

Early on, the partners decided to explore multiple domains simultaneously in our transition to linked data. Columbia is looking at the intersection between the museum and library communities. This sub-project will focus on testing BIBFRAME's suitability for the description of art objects, both two-dimensional and three-dimensional. Cornell will be doing two projects. First is the community development of a library ontology extension for the rare materials community focusing on the instance and item level because data such as provenance and binding are currently not well defined in library ontologies. And second is the original metadata creation for noncommercial LPs from their Hip Hop collection. Harvard's sub-project for LD4P explores best practices for creating native Linked Data descriptions for library

cartographic resources including printed maps, atlases, and digital geospatial datasets. The Library of Congress will be working on four projects. For their first project, LC will focus on metadata creation for its archival film and recorded sound collections, folding in recommendations from the AV data modeling study they commissioned in 2014. LC's next project explores best practices for creating Linked Data descriptions for print and photograph resources. LC's third project is BIBFRAME 2.0 vocabulary development. And LC's last project will explore the BIBFRAME and RDA data models and best practices for creating Linked Data descriptions for resources in monographic, serial, notated music, and cartographic formats, as these resource types most actively make use of RDA descriptive cataloging standard. In March 2015, Princeton acquired the personal library of Algerian-born French philosopher Jacques Derrida. Taking this collection, the overarching goal of Princeton's LD4P project is to explore, develop, and implement linked data standards for the description of special collections materials and the annotations they contain.

Stanford will be working on two projects. The first is the Performed Music Ontology Project. This project aims to develop a BIBFRAME-based ontology for performed music in all formats, with a particular emphasis on the modelling of works, events, and their contributors. The second project is called the Tracer Bullets, the conversion of four of our traditional technical services workflows to linked data. The idea is to reimagine an entire cataloging workflow as RDF based, from acquisitions to discovery, and to do it as a thin red line, something that can be fleshed out more fully over time. Our four pathways are Traditional Vendor-supplied Cataloging, Original Cataloging, Self-deposit of a single tem to the Digital Repository, and Ingestion of a collection into the Digital Repository.

As the grant approaches the end of its first year, we are finishing up Pathway 1, copy-cataloging making use of vendor copy, and will be moving on to original cataloging starting this summer. Since 2010, Stanford has taken advantage of RDA and what we could express in MARC as a preparation for our transition to linked data. But Workflow 2, that is, original cataloging directly in RDF, will be our first great test of RDA explicitly in the context of linked data.

Over the course of this morning we have heard five excellent presentations.

Tiziana Possemato spoke on "How RDA is Essential in the Reconciliation and Conversion Processes for Quality Linked Data." Tiziana spoke about Casalini's SHARE-Virtual Discovery Environment and its ability to draw together multiple representations of an entity into a single entity. This is a truly new form of "authority control" accomplished through the clustering of an entity's attributes in order to create a profile which can be used to draw together these multiple representations across institutions.

Magda El-Sherbini spoke on "RDA Implementation and the emergence of BIBFRAME." Magda very helpfully started out by focusing our attention on this transition's effect on discovery and users' needs, beginning with two opportunities (searching the catalog by subjects in a users' preferred language and the representation of attributes by URIs so that we can link related concepts). She then clearly showed the good, the bad, and the ugly of trying to encode our RDA data in MARC and the opportunities that BIBFRAME provides for being able to better reflect the goals of RDA and make that data available to the Web in a coherent, understandable way. She's also quick to note, however, that BIBFRAME is not a completely arbitrary carrier of data and that RDA has definitely had an effect on its evolution.

Gordon Dunsire spoke on "RDA and Practical Linked Open Data." He elucidated the fourfold path for capturing and recording data. Also how we can link to external non-FRBR properties through unconstrained properties. He closed with a number of strategic challenges to move us into this new world.

Anita Goldberga spoke on "Identification of entities in the Linked Data Collection Rainis and Aspazija." This pilot project focused on creating a linked data resource for the Latvian National Awakening. She noted the lack of sufficiency in the current NLL authority database especially in the areas of events and places. And also RDA's strengths in supplying language independent relationship designators and controlled vocabularies. However, RDA will also need to expand current relationship designators to include concepts such as "is mentioned."

Jackie Shieh gave a "report from the PCC Task Groups on URIs in MARC and BIBFRAME mapping. Jackie stressed important steps that we can take now as a community to move towards linked data readiness. By approaching a standardized way of including identifiers in

MARC data, whether they are for Real World Objects or Descriptions, we prepare our data for smoother conversion to linked data. The mapping from MARC to BIBFRAME will be an ongoing process. Far from being dead, the MARC formats continue to evolve as we try to capture more refined data or even remove ISBD punctuation. BIBFRAME as well will continue to evolve as it is tested in real life production settings and RDA itself will continue to evolve as well. How do we manage to keep these three in synch as we move forward? Do we wait for the perfect time to convert our legacy data? Reconvert as standards evolve?

The papers concluded with my own presentation on "RDA, Lined Data, and the End of Average." These talks are wonderful examples of the growing interactions between RDA and linked data:

- From RDA's focus on data construction for discovery to BIBFRAME's enhanced ability to capture that data in a Web accessible way
- From BIBFRAME's evolution to better reflect RDA to RDA's design to be integratable with the Web
- From the need for better entity identification and reconciliation to the complexity of synchronizing three developing standards, MARC, BIBFRAME, and RDA

One of the main drivers is clearly RDA. Because of RDA's

- extensible and flexible framework
- its stressing of relationships
- its focus on discovery

it's a natural representation of our data on the Web. And RDA will continue to drive the development of both MARC and BIBFRAME as libraries make a fundamental shift to linked data and look to RDA as a model for how best to represent their complex data for international discovery and reuse.