How Much Description is Enough?

A Brief History on the Debate over the Dublin Core Metadata Initiative

Eliot Scott

University of Texas at Austin

How Much Description is Enough?

A Brief History on the Debate over the Dublin Core Metadata Initiative

The description of information resources remains paramount in the quest to catalog the world's knowledge to facilitate simple, straight-forward resource retrieval. Yet contentious issues arise when raising the question of how best to catalog these resources. Some argue that a complex cataloging schema of numerous fields containing detailed descriptions of items are necessary to accommodate the vast array of information resources produced by humanity. Yet others argue that such schema produce needless complexity and confusion, rendering them difficult to manage to all save the devout specialist. Although this conundrum has likely plagued information resource specialists long before the advent of modern electronic information search and retrieval techniques, debates on how best to catalog information have become heated as information specialists grapple with the relatively new and constantly evolving electronic tools at their disposal. A factious debate concerning how best to catalog resources arose between information professionals in the past decade mainly between the more established field of library science and their newer cousins in the computer science field. This debate fully manifests itself in the question of how much resource description is enough, and whether the Dublin Core Metadata Initiative constitutes an adequate metadata schema for the search and retrieval of the vast amounts of physical and electronic information resources currently at humanity's disposal.

During the nascent period of the World Wide Web in the early 1990's, newly minted computer users generated vast amounts of electronic resources none of which contained any cataloging information. Without any description, search and retrieval of these resources remained difficult. In order to remedy the lack of cataloging information of these new electronic resources, the Online Computer Library Center (OCLC) and the National Center for

Supercomputing Applications (NCSA) held a joint conference in Dublin, Ohio in 1995 to discuss "how a core set of semantics for Web-based resources would be extremely useful for categorizing the Web for easier search and retrieval" ("History of the Dublin Core Metadata Initiative," n.d.). As a result of this conference, the Dublin Core (DC) Metadata schema was created, initially containing a set of thirteen core resource description elements with an expansion to fifteen in 1998[1]. These elements made up "Simple Dublin Core" for basic resource description while a modified schema, "Qualified Dublin Core", was developed to enhance the core elements for more precise and detailed description (Sugimoto, Baker, and Weibel, 2002).

Dublin Core was designed as a simplistic resource description format focused on maximum interoperability across resource domains. In addition to containing only fifteen elements of description, DC used techniques such as "the dumb down principle" - which stated that all qualified or sub-element descriptions had to function also as their parent element - to further the simplification and interoperability of the schema (Sugimoto, Baker, and Weibel, 2002). DC's simplification of resource description was designed for those "unschooled" in what DC co-founder Stuart Weibel referred to somewhat derisively as the "cataloging arts" (as cited in Gorman, 1999, p. 19). Thus began the debate between what Weibel referred to as the "Minimalists" and "Structuralists" (as cited in Hopkins, 1999, p. 57). The Minimalists in the DC camp believed that resource description could be simplified thus facilitating easier, faster and cheaper cataloging of the many electronic resources being created. The more traditional Structuralists meanwhile, argued that the world was complex necessitating far more complex description schemes.

[1] A Sixteenth was added in 2004 (Coleman, 2005).

The advent of the world wide web and the buzz surrounding the DC metadata schema as it developed created considerable anxiety for Structuralists as they lamented the disregard by the Minimalists of more detailed and organized information description schemes, particularly the MARC format. As DC was designed to be simple, containing only fifteen elements and 'dumbed-down' cataloging rules for those "unschooled in the cataloging arts", Peter Graham described the discourse surrounding DC as an "anti-MARC discussion" (as cited in Gorman, 1999, p. 19). Michael Gorman likewise (1999) criticized the DC format as merely a "subset of MARC" (p. 18). Meanwhile, non-librarians such as Douglas Winship referred to MARC and AACR as "elitist...fancy coding schemes" (as cited in Hopkins, 1999, p. 56). Gorman and others firmly in the Structuralist camp argued that fifteen elements did not adequately describe the resources available, while Minimalist DC proponents argued that standard library cataloging techniques and metadata formats like AACR and MARC, were needlessly complicated. This debate became increasingly polarized from the end of the 1990's and well into the 2000's with both sides criticizing the other with growing hostility.

With an attempt at even handedness, Hopkins (1999) noted that OPAC (the Online Public Access Catalog) could "no longer be considered the center of the information world" and that the MARC metadata format would not likely be adopted by non-librarians (p. 66). Hopkins' sentiment acknowledged what Minimalist DC proponents had been arguing - that traditional library metadata formats were overly complex for non-professional catalogers. As one of the goals of DC was to allow electronic content creators to describe their work so that it could be found by web based search engines, complex cataloging rules were seen as a barrier to resource description by the Minimalists. Nevertheless, Hopkins also acknowledged that MARC would continue in its current use stating, "MARC will either continue to develop in response to new

needs...or it will be supplemented by alternate methods of representing that data; in either case MARC will continue to be used by libraries" (p. 66). Thus Hopkins believed that although MARC might not adapt to the ever changing conditions in electronic resource description on the Web, it would continue its dominance in libraries where it had remained a resource description staple for over three decades.

Despite its beginnings in web resource description however, the DC metadata schema quickly spread beyond describing only electronic resources to other materials. By 2005, Coleman notes that Dublin Core metadata was "used widely to describe all types of resources, including books" (p. 154). This expansion of DC usage into areas traditionally held by library catalogers working with the MARC metadata format intensified the firestorm of controversy causing librarians such as Jeffrey Beall to respond even more harshly. In 2004, Beall wrote an obituary for the Dublin Core metadata schema stating that the standard had become "obsolete" and "outlived its usefulness" (p. 40). Beall argued that Dublin Core would be replaced by the MODS metadata format developed by the Library of Congress, as well as by Google and its advanced search algorithms that were unavailable on the early World Wide Web at birth of the Dublin Core metadata format. Beall described the response to his article as "quick...negative and strong" (2006, p. 57). He attributed the reaction to people whose livelihoods depended on the continued existence of the DC format before again promoting the MARC and MODS formats, which he neglected to mention his own livelihood as a librarian was dependent upon (2006). Further bemoaning the simplicity of the DC schema, Gorman noted that the computer professionals that developed metadata schemas to replace MARC believed MARC's "complexity was the result of whims of cataloguers". He continued, "Here is a 'reality check'–the framework standard MARC and the content standard AACR2 are complex because the world is complex–get over it!" (2006,

p. 48). Gorman suggested that Beall's obituary represented the end of the idea that metadata could be created cheaply without the need for trained catalogers (2006). Thus it would seem that the professional catalogers schooled in the "cataloging arts", decried as elitists by Winship half a decade earlier, had begun a very pointed and public defense of their own methods ridiculing DC and its Minimalist proponents in the process.

Yet despite predictions of Dublin Core's demise from Beall and Gorman, the DC schema persisted in a variety of professional information environments. Rettig, Liu, Hunter and Level in describing their experience working with metadata at the Colorado State University Library stated, "DC is flexible and easy to apply. It is particularly useful in addressing cross-collection and cross-discipline searching" (2008, p. 325). Janney documented a collaborative project digitizing and displaying historical images on the web involving the University of Washington Library, the Museum of History and Industry in Seattle and ten local heritage museums which utilized DC metadata to describe the resources (2005). "Dublin Core provided the glue that made it work", Janney declared of the project (2005, p. 33). The DC schema was also used in the medical field. Xu and Okada described the creation of a medical database utilizing the DC schema to facilitate access to medical information in 2007. DC also provided a format which could describe the previously un-cataloged resources contained in the personal archives of University faculty. Rowan Brownlee at the University of Sydney Library documented a process where he mapped individual university research project data to DC for increased access to university resources (2009). Given the success of these DC related projects and the positive assessments of DC, the death knoll that Beall and Gorman surmised appears wildly inaccurate.

Despite the success of DC, it did not replace MARC as the metadata format of choice for librarians, just as Hopkins predicted. In a survey responded to by 68 libraries conducted by Jin

Ma in 2007, 78% reported using DC metadata. However 91% reported using MARC. Rather than advocating for one or the other, Ma concludes "libraries have stepped out of the MARC box" supporting a variety of projects using a variety of formats (2009, p. 12). Similarly, Coleman argued several years earlier that libraries needed to integrate DC metadata with older metadata formats such as MARC (2005). In fact numerous projects have utilized both DC and MARC, as well as MODS, EAD and others quite effectively. Documenting a digital preservation project at the University of Houston Digital Library, Chen and Reilly note that the library used DC for its access metadata, but required a more robust set of elements for its preservation metadata. They proceeded to demonstrate how the library converted DC metadata to METS utilizing open source tools (2011). Likewise, the University of Illinois Library provided MARC, METS and DC metadata for its digitized book collection (Han, 2011). Ohio State University even converted its entire MARC library catalog to DC so that it could be loaded into its institutional repository for greater library patron access (Walsh, 2011). Thus it would seem, contrary to the "Tower of Babble of metadata schemas" predicted by Beall if multiple schemas were implemented (2006 p. 61), DC and MARC have both thrived in their co-existence, creating a multi-lingual metadata environment.

Nevertheless, problems in the DC schema predicted by Gorman and Beall have plagued many projects utilizing the DC schema as well. While creating new metadata standards for the Mountain West Digital Library, the Utah Academic Library Consortium Task Force "wrestled with the problems within Dublin Core metadata" noting that while DC "metadata is a universal metadata schema, lack of consistency in metadata was an issue that confronted Task Force members when examining various projects" (Toy-Smith, 2010, p. 8). Likewise, in an analysis of six digital video repositories' DC metadata records, Weagley, Gelches and Park identify "barriers

to interoperability...most notably the confusion caused by semantic ambiguity among DC elements" (2010, p. 37). In both these projects, the failure of DC was not a lack of a sufficient number of elements, but rather a lack of consistent cataloging rules and a controlled vocabulary for the elements.

Yasser identifies five basic problems with metadata, "Incorrect Values, Incorrect Elements, Missing Information, Information Loss, and Inconsistent Value Representation." (2011, p. 51). Incorrect values within elements, missing information and information loss can occur in any record, regardless of the metadata scheme utilized. However, the semantic ambiguities among DC elements can result in the population of incorrect elements[2], while the lack of a controlled vocabulary in DC often results in inconsistent value representation. In a survey of information professionals conducted by Park and Childress, participants commented that both "conceptual ambiguities" and "semantic overlaps" in DC led to confusion and multiple interpretations of DC elements (2009, p. 727). Park and Childress point out that these problems create inaccurate and inconsistent metadata across institutions that limit interoperability between catalogs (2009). Other surveys conducted more recently have also demonstrated increasing frustration with the DC format due to these inconsistencies and difficulties in integrating with other catalogs (Zeng, 2009), and have shown declines in DC usage among libraries from the Ma survey in 2007 (Park & Tosaka, 2010).

The frustration on the part of librarians and other information professionals with the DC schema is not surprising. As Park and Childress (2009) point out:

[2] This is particularly true for the "format" and "type" element (Weagley et.al., 2010; Yasser, 2011)

> The DC metadata scheme was developed primarily for one very important purpose – to facilitate cross-domain information resource description. Accordingly, DC semantics underscores the lowest common denominator for resource description. It is expected that document authors without an educational and professional background in cataloguing will be able to use the scheme. (p. 728).

DC was designed to be a schema for the masses and not for cataloging professionals. Thus its rules for resource description are somewhat vague and open to interpretation. Although this may facilitate DC's use among non-professionals, it also creates problems when merging resource descriptions from a variety of repositories.

Nonetheless, despite DC's shortcomings, it has aided in resource description for a variety of institutions. As an XML based format, it is also highly extensible and flexible and the format has evolved and will continue to evolve to compensate for these weaknesses. Martin (2011) notes that the Dublin Core Application Profiles (DCAP) and the DCMI Abstract Model continues to provide guidance and structure while defining terms and the rules for their use within the DC schema for wider use of data within the linked DC community of catalogs. Thus, efforts such as the DCAP and the Abstract Model will undoubtedly assist in remedying some of DC's imperfections.

It is perhaps not the DC format that is the primary problem however, but the schema's implementations and the notion that non information professionals can adequately catalog resources. In an analysis of the DC metadata contained in a DSpace repository, Kurtz discovered that " One repository used librarians to oversee the archiving process, while the other two employed two different strategies as part of the self archiving process. The librarian-overseen

archive had the most complete and accurate records for DSpace entries" (2010, p. 40). That a librarian created the most complete and accurate metadata is hardly surprising, yet it underscores the fact that any resource description done properly by trained professionals can be far more useful than resource description by non professionals, regardless of format. Weagley et. al. reiterate this point in their conclusion of a DC based digital video repository lamenting "collections that are not under any sort of library type jurisdiction by which quality standards are maintained and monitored" contain consistency problems "no matter what metadata scheme is used" (2010, p. 54). Accordingly, trained catalog librarians will remain the best suited to describe resources regardless of the metadata format.

What becomes obvious from the history of the debate over Dublin Core is not that a certain number of descriptive elements are necessary when performing resource description, but rather that the quality of the descriptions and their interoperability with one another remain paramount in creating a worldwide network of interlinked libraries and resources. What is also plain, is that a single metadata format to describe every information resource is neither necessary nor desirable. Although a detailed MARC record is indeed highly useful for resource retrieval, the creation of a MARC record is difficult, time consuming and thus expensive. Though less detailed, a DC record is also very useful for information description and retrieval, and can be created more quickly and cheaply than a MARC record. Moreover, if created by a professional cataloger, the DC record can be converted to MARC at a later date. Any resource description is ultimately better than none. With constant budget cuts and reduced funding for information resource institutions, a DC record, although less detailed than a MARC record, may be the best metadata format for a budget strapped institution to create a standards compliant catalog.

References

Beall, J. (2004). Dublin Core: An Obituary. *Library Hi Tech News, 8*, 40-41. doi:10.1108/07419050410567399

Beall, J. (2006). The Death of Metadata. *The Serials Librarian, 51*(2), 55-74. doi:10.1300/J123v51n02_05

Brownlee, R. (2009). Research Data and Repository Metadata: Policy and Technical Issues at the University of Sydney Library. *Cataloging & Classification Quarterly, 47*(3), 370-379. doi:10.1080/01639370802714182

Chen, M., & Reilly, M. (2011). Implementing METS, MIX, and DC for Sustaining Digital Preservation at the University of Houston Libraries. *Journal of Library Metadata, 11*(2), 83-99. doi:10.1080/19386389.2011.570662

Coleman, A. S. (2005). From Cataloging to Metadata: Dublin Core Records for the Library Catalog. *Cataloging & Classification Quarterly, 40*(3), 153-181. doi:10.1300/J104v40n03_08

Gorman, M. (1999). Metadata or Cataloguing? A False Choice. *Journal of Internet Cataloging, 2*(1), 5-22. doi:10.1300/J141v02n01_03

Gorman, M. (2006). Metadata Dreaming: The Keynote Speech at the Canadian Metadata Forum, September 2005. *The Serials Librarian, 51*(2), 47-54. doi:10.1300/J123v51n02_04

Han, M.-J. (2011). Creating Metadata for Digitized Books: Implementing XML and OAI-PMH in Cataloging Workflow. *Journal of Library Metadata, 11*(1), 19-32. doi:10.1080/19386389.2011.545001

History of the Dublin Core Metadata Initiative. (n.d.). Retrieved June 18, 2011, from DCMI website: http://dublincore.org/about/history/

Hopkins, J. (1999). USMARC as a Metadata Shell. *Journal of Internet Cataloging, 2*(1), 55-68. doi:10.1300/J141v02n01_06

Janney, K. (2005). Collaborative Cataloging: Using Dublin Core to Unite Local Heritage Organizations. *Journal of Internet Cataloging, 7*(1), 33-48. doi:10.1300/J141v07n01_03

Kurtz, M. (2010, March). Dublin Core, DSpace, and a Brief Analysis of Three University Repositories. *Information Technology and Libraries, 29*(1), 40-46. Retrieved from http://ezproxy.lib.utexas.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=tfh&AN=48049878&site=ehost-live

Ma, J. (2009). Metadata in ARL Libraries: A Survey of Metadata Practices. *Journal of Library Metadata, 9*(1), 1-14. doi:10.1080/19386380903094977

Martin, K. E. (2011). Marrying Local Metadata Needs With Accepted Standards: The Creation of a Data Dictionary at the University of Illinois at Chicago. *Journal of Library Metadata, 11*(1), 33-50. doi:10.1080/19386389.2011.545006

Park, J.-R., & Childress, E. (2009, December). Dublin Core metadata semantics: an analysis of the perspectives of information professionals. *Journal of Information Science, 35*(6), 727-739. doi:10.1177/0165551509337871

Park, J.-R., & Tosaka, Y. (2010, September). Metadata Creation Practices in Digital Repositories and Collections: Schemata, Selection Criteria, and Interoperability. *Information Technology and Libraries, 29*(3), 104-116. Retrieved from http://www.ala.org/ala/mgrps/divs/lita/ital/292010/2903sep/sept_pdf.cfm

Rettig, P. J., Liu, S., Hunter, N., & Level, A. V. (2008). Developing a Metadata Best Practices Model: The Experience of the Colorado State University Libraries. *Journal of Library Metadata, 8*(4), 315-339. doi:10.1080/19386380802656371

Sugimoto, S., Thomas, B., & Stuart, W. (2002). Dublin Core: Process and Principles. In E. Lim, S. Foo, C. Khoo, H. Chen, E. Fox, S. Urs, & T. Costantino (Eds.), *Lecture Notes in Computer Science: Vol. 2555. Digital Libraries: People, Knowledge, and Technology* (pp. 25-35). doi:10.1007/3-540-36227-4_3

Toy-Smith, V. (2010). UALC Best Practices Metadata Guidelines: A Consortial Approach. *Journal of Library Metadata*, *10*(1), 1-12. doi:10.1080/19386380903543734

Walsh, M. P. (2011, January). Repurposing MARC Metadata for an Institutional Repository: Working with Special Collections and University Press Monographs. *Library Resources & Technical Services, 55*(1), 33-44. Retrieved from http://ezproxy.lib.utexas.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=tfh&AN=57194030&site=ehost-live

Weagley, J., Gelches, E., & Park, J.-R. (2010). Interoperability and Metadata Quality in Digital Video Repositories: A Study of Dublin Core. *Journal of Library Metadata*, *10*(1), 37-57. doi:10.1080/19386380903546984

Xu, W., & Okada, M. (2007). EBM Metadata Based on Dublin Core Better Presenting Validity of Clinical Trials. *Journal of Medical Systems, 31*(5), 337-343. doi:10.1007/s10916-007-9073-0

Yasser, C. M. (2011, May). An Analysis of Problems in Metadata Records. *Journal of Library Metadata, 11*(2), 51-62. doi:10.1080/19386389.2011.570654

Zeng, M. L., Lee, J., & Hayes, A. F. (2009). Metadata Decisions for Digital Libraries: A Survey Report. *Journal of Library Metadata, 9*(3), 173-193. doi:10.1080/19386380903405074