

## DTD variety is the spice of metadata life

---

*April Younglove*

[ayounglo@emporia.edu](mailto:ayounglo@emporia.edu)

*School of Library and Information Management*

*Emporia University*

*January 2007*

---

In Roy Tennant's 2002 article, "MARC must die," Tennant dramatically demands a complete overhaul or the total elimination of MARC, the library world's favorite longstanding schema for cataloging data. He points out that MARC maintains many arcane features that are no longer necessary. Obscure digital coding in record headers is not human readable and is pointlessly redundant since computers are no longer dependant on punch cards. In addition to accusing MARC of being backwards, Tennant also asserts that it lacks the capability to go forwards. Its inability to handle pictures, digital objects or lengthy text blocks like book reviews makes it unsuitable for an online future where such features are commonplace. At a time when XML is quickly becoming the industry standard for data storage and transfer over the Internet, Tennant observes that MARC's incompatibility causes it to remain stagnant. If MARC cannot work with XML, he fears that its data will become increasingly irrelevant in the growing search environment outside of library catalogs. Thus, Tennant concludes that MARC is not robust or granular enough to be successful in a web era. However, in 2004, in an article entitled "Building a New Bibliographic infrastructure," Tennant tempers his previous position on MARC saying, "I decided I had convicted the wrong subject. Let MARC die of old age rather than homicide" (§ 1). This reversal came about in part because newly emerging Document Type Descriptions (DTDs) like Dublin Core, MODs, and METs, to name a few, were making great strides towards peacefully co-existing and even interoperating with MARC.

In his article, "XML and MARC: Which is Right?" Bruce Johnson addressed the idea that perhaps XML would be MARC's successor after its predicted demise. Instead of concluding, as some did at that time, that MARC had to be discarded in favor of XML, he states that "as related to development of standards for digital libraries, XML and MARC do not represent an "either/or" choice but rather an "and/with" choice" (2001, 86). To change catalogs over to an XML format, Johnson was envisioning a nightmarish scenario where libraries would attempt mass conversion of the data already in MARC records, which at that time was very difficult. He observed that libraries had invested so much both emotionally and physically in MARC that it would be foolish not to investigate and promote solutions that could bridge the gap between AACR2's standards and XML's flexibility (87). While it is now much easier to translate between MARC and other formats in 2007, Johnson's statement that "it is likely that the huge investment in data creation and software will guarantee its [MARC's] use long into the distant future," is still valid (88). Although there are some recent attempts to update MARC's AACR2

standards, currently librarians still prefer and rely on cataloging with both AACR2 and MARC. Because of this, alternate DTDs and XML upgrades have all been made to seamlessly integrate with existing MARC software. OCLC's Connexion client is an example of how that MARC can co-exist and even interoperate with XML and alternate DTD schemas such as Dublin Core (DC). On the surface Connexion may appear to some users to be MARC driven, but actually it utilizes XML on its server side. OCLC's infrastructure allows users to not only upload and download MARC records, but also to catalog in DC or extract records in MARCXML (Reese, 2007a, slide 12). However, this is not common knowledge to the majority of catalogers who still depend primarily on MARC. As another example of how a catalog can accommodate MARC but still retain the advantages of XML, consider the software Syndetics. Syndetics allows III users to customize their MARC OPACs with an XML wrapper that can add-on book jacket pictures and allow interactive features without actually altering any MARC records. The disadvantage of not demanding that MARC change to accommodate any of these new features, though, is that it gives more leverage to those vendors who can create elaborate work-arounds. If MARC itself were able to evolve then perhaps catalogers and local library staff would be more able to freely experiment without having to pay for upgrades from a company one predetermined feature at a time.

While MARC is not being actively replaced by any DTDs, there are a great many DTDs that now co-exist with MARC. It is worthwhile to examine a few in order to understand how new bibliographic descriptive formats are affecting the way that libraries store and exchange metadata. Dublin Core (DC) is the simplest DTD with only 15 fixed fields. This means that even non-catalogers can quickly organize metadata by using DC. Its simplicity requires MARC users to sacrifice a great deal of granularity though if users intend to translate records from MARC into DC for sharing over the web. Other formats, like MODS, can more accurately translate MARC data into XML, and for this reason are more popular with libraries (Coyle, 2005, ¶ 15). The Library of Congress' Network Development and the MARC Standards Office created MODS, or Metadata Object Description Schema, in order to "provide an alternative between a simple metadata format with a minimum of fields and little or no substructure such as DC and a very detailed format with many data elements having various structural complexities such as MARC 21" (Guenther & McCallum, 2002, ¶ 3-5). For the most part, MODS is a subset of MARC, translating its AACR records into XML friendly format. It substitutes MARC numbers with human-readable tags -- so that, for example, the MARC 245 field becomes "title" in MODS (Coyle, 2005, ¶ 15). Unfortunately MODS does not actually add any additional capabilities to MARC beyond enabling XML file sharing, and, in fact, collapses a few of its data fields to make the records more exchangeable (Guenther & McCallum, 2002, ¶ 5).

METS was created on a different principle entirely than either DC or MODS. It can either hold its own internal descriptive metadata, or point within its records to external data created in other standards like DC, MODS and MARC (Tennant, 2002a, ¶ 8). Librarian Karen Coyle describes METS as being like a digital binding and cover for the computerized files that make up a virtual book. Within this binding, METS has a sort of copyright page that provides the technical data about "file formats, the technology used in

scanning if the item began its life on paper, and the digital transformations and compression that have been used on the files” (Coyle, 2005, ¶ 16). METS, the Metadata Encoding and Transmission Standard, does not sacrifice any exchanged data because it not a true metadata container. Instead, METS is a wrapper that uses place holders to refer to foreign material. For internally contained data, METS prefers XML. It consists of six elements that contain or point to the digital object’s metadata: a header, descriptive metadata, administrative metadata, file section, structural map and behavior section. With the exception of the header and structural map, all elements are optional and can be stored externally in any format. That it has so few required can be seen as either an advantage or a disadvantage. It is highly flexible and can contain almost any type of data imaginable (pictures, text, music files etc.), but because there is no overarching standard, two METS records from two different sources could conceivably be incompatible.

Tennant realized that if libraries wish to participate in the larger web forum and keep their cataloging methods from becoming marginalized, they must not insist that MARC be the only DTD platform that their catalogs use for storing and retrieving metadata. Library metadata that cannot be harvested by search engines will be invisible to patrons who increasingly look for data online. OCLC’s open Worldcat initiatives are attempting to create new access points between existing data sets trapped in catalogs and what people see on the web. Open Worldcat allows Internet users to simultaneously search all participating OACs with a single search box. OCLC also uses XML on its server side to make publications available to Google Scholar and Google Books (Reese, T., 2007d). A key feature of open access is to allow harvesting from one system to another. If librarians wish to remain relevant, they should invest more time in learning about emerging DTDs and the XML structures that are beginning to drive MARC software “under the hood.” While MARC is a time-proven format for metadata storage, insisting that it remain the only format is both backwards and naive.

In the future, hopefully specialized DTDs will allow users catalogers to enhance metadata by tailoring it to the needs of specific user groups without sacrificing bibliographic authority control. In his 2006 piece, “The New Cataloger,” Tennant writes optimistically that “the modern cataloger will one day be a software-enabled specialist who can gather, subset, normalize, and enrich piles of records for a specific audience or purpose.” For instance, “libraries would standardize on known metadata schemas for particular data objects. IE, cartographic data would be done in FGDC [used for cartographic and geographic information], general text in MODS or TEI, Finding Aids in EAD [used by Archives and Special Collections to define finding aids]” (Reese, 2007b). All of these document standards can be translated into MARC’s AACR format or adapted into other industry standards if need be. This means that catalogs can use a variety of DTDs internally while maintaining data integrity and bibliographic consistency. According to Oregon State librarian Terry Reese, technology is coming (like his Rosetta stone, which translates between 10 different metadata storage languages) that will make the translation, or crosswalking process, between different schemas even more of a reality and less labor intensive than it currently is (Reese, 2007c). In other words, in the future, MARC will not disappear so much as become a jumping off point for future metadata schemas in libraries.

## References

- Coyle, Karen. (2005). Understanding Metadata and its purpose. *Journal of Academic Librarianship*, 31(2), 160-163. Preprint. Retrieved January 15, 2007 from [http://www.kcoyle.net/jal2\\_Metadata.html](http://www.kcoyle.net/jal2_Metadata.html)
- Guenther, R. & McCallum S. (2002). New Metadata standards for digital resources: MODS and METS. *Bulletin of the American Society for Information Science and Technology*, 29(2), 12-14. Retrieved July 9, 2006 from [http://www.findarticles.com/p/articles/mi\\_qu3991/is\\_200212/ai\\_n9150534](http://www.findarticles.com/p/articles/mi_qu3991/is_200212/ai_n9150534)
- Johnson, B. (2001). XML and MARC: Which is right? *Cataloging & Classification Quarterly*, 32(1), 81-90.
- Reese, T. (2007a). The History and Future of XML in Libraries. Posted in Blackboard's Course Documents: Lectures for Emporia University LI862XK students.
- Reese, T. (2007b, January 17). RE:On the importance of shared standards. Online posting. *Blackboard Discussion Board*. Accessed January 25, 2007 from LI862XK/XL METADATA FRAMEWORKS (SPRING 2007) > DISCUSSION BOARD > LI862XK/XL SPRING 2007 CLASS DISCUSSION BOARD Week 2: Discussion topic
- Reese, T. (2007c, January 18). RE:RE:Week 2: Discussion topic. Online posting. *Blackboard Discussion Board*. Accessed January 25, 2007 from LI862XK/XL METADATA FRAMEWORKS (SPRING 2007) > DISCUSSION BOARD > LI862XK/XL SPRING 2007 CLASS DISCUSSION BOARD Week 2: Discussion topic
- Reese, T. (2007d, Jan. 25). RE:RE:RE:RE:RE:Emerging Standards. Online posting. *Blackboard Discussion Board*. Accessed January 25, 2007 from LI862XK/XL METADATA FRAMEWORKS (SPRING 2007) > DISCUSSION BOARD > LI862XK/XL SPRING 2007 CLASS DISCUSSION BOARD Week 2: Discussion topic
- Tennant, R. (2002a, April 15). Digital libraries – Metadata as if libraries depended on it. *Library Journal*, np. Retrieved January 25, 2006 from <http://www.libraryjournal.com/article/CA206408.html>
- Tennant, R. (2002b, Oct. 15). MARC Must Die. *Library Journal*, 127 (17) 26-28. np on PDF from Emporia Course Reserves.
- Tennant, R. (2003). Building a New Bibliographic Infrastructure. *Library Journal*, 129(1), 38.
- Tennant, R. (2006, April 15). The new cataloger. *Library Journal*, 32. Retrieved January 20, 2006 from InfoTrac.