Finding "the Stuff:" Making Standard Bibliographic Data Useful for an Online Environment[†]

Michael Neubert, Sandra J. Bostian

Library of Congress
{mneu, sbos}@loc.gov

Abstract

This paper gives examples of how data developed to describe the original library items (not the digital surrogates) is manipulated at the Library of Congress (LC) in order to make those records more useful for presentation of the digital surrogates. This includes the addition of some data purely for the functionality of the site and some data to help end users with their information seeking activities. The examples are drawn from the experiences of the *Global Gateway* production team and their work in creating collaborative international digital library projects. Most of this work has been done on the *Meeting of Frontiers* web site, an online Russian-American bilingual digital library (http://frontiers.loc.gov/).

One of the challenges for digital libraries is that they must balance the demands of:

- user expectations;
- site programming and functionality; and,
- a highly structured cataloging environment (this is particularly true of the Library of Congress).

The Library of Congress (LC) today presents more than seven million digital items via its *American Memory* and *Global Gateway* sites. The presentation of these digital items is enhanced by accompanying descriptive and subject information, a portion of which is also available to the user (hereafter, "presentation data"). We will be examining how this balance is managed within LC's *Meeting of Frontiers* site, a cooperative bilingual digital library.

American *Memory* and *Global Gateway* are similar in that they are both "digital libraries". *American Memory* is far older, having begun in the mid-1990s. It is described on its home page as a "gateway to rich primary source materials relating to the history and culture of the United States. The site offers more than 7 million digital items from more than 100 historical collections."[6]

Proceedings of the 5th Russian Conference on Digital Libraries RCDL2003, St.-Petersburg, Russia, 2003

Global Gateway is a newer effort to build collaborative bilingual digital library projects with foreign and U.S. partners. These partners are primarily national libraries but also include regional museums and scholarly institutions within those countries. The "flagship" is the *Meeting of Frontiers* (*MoF*) site that, as of May 2003, includes more than 330,000 digital items and over 9,800 library items for which there are bibliographic records. *Frontiers* "is a bilingual, multimedia English-Russian digital library that tells the story of the American exploration and settlement of the West, the parallel exploration and settlement of Siberia and the Russian Far East, and the meeting of the Russian-American frontier in Alaska and the Pacific Northwest."[11]

Meeting of Frontiers differs from *American Memory* projects by the inclusion of a large amount of interpretative text material, more consistent with LC's exhibit approach.[1] The narrative texts provide users with historical background about the materials in the digital library collections. These texts are original to the project, written by historians who have been hired on contract for this purpose.

The digital library portion of the site includes materials–and associated descriptive data–contributed by a dozen partners in the U.S., Germany and Russia. *Meeting of Frontiers* has pushed for uses of this bibliographic data that improve, we hope, the online user's experience.

How is the data organized?

The presentation data for *Global Gateway* projects can be organized in three ways:

- LC's MARC cataloging system;
- Pseudo-MARC records; and
- Non-MARC records.

MARC refers to "MAchine Readable Cataloging", the standard "for the representation and communication of bibliographic and related information in machine-readable form" at the Library of Congress. MARC 21 is the standard format for holding cataloging data used at the Library of Congress and many libraries worldwide. The MARC descriptive data describes the "original" materials (that is, the material before it was converted to digital format). This is also true of the pseudo-MARC and non-MARC data.

Where MARC records do not already exist, either the pseudo-MARC or non-MARC approach is used. For both approaches, the data corresponds to that held in the MARC cataloging fields but is stored in a separate Access database. Pseudo-MARC refers to data that is exported from Access in a MARC-like format (using MARC field identifiers) and joined with LC MARC records for online presentation. Non-MARC data is exported in a simple, tab-delimited format and indexed separately. The fields in this database are then mapped to MARC format in a way that the end user cannot tell the difference online. *Meeting of Frontiers* uses the pseudo-MARC approach but the preferred data format for future Global Gateways projects is the non-MARC database.

Where does the data come from?

The MARC records presented in *Meeting of Frontiers* are "pulled" (exported) from the Library of Congress Integrated Library System (ILS), a *Voyager* system from Endeavor. As of the May 2003 release, *MoF* had 9,848 bibliographic records, of which 5,256 were pulled from the ILS.

ILS record sets are distributed to other institutions across the US and internationally. This means that any changes made to them must adhere to strict national standards, the *Anglo-American Cataloguing Rules*, 2nd edition (AACR2), for all fields other than those designated for local institution use. Only professional cataloging staff across several different Library of Congress divisions can make changes to ILS records and the process for requesting and making these changes is time consuming.

In the case of *Meeting of Frontiers*, most of the pseudo-MARC records are from other institutions. Along with the scanned images, partners also supply bibliographic data, which is then incorporated into the online presentation.

Some additional records are created by project staff for LC items without MARC records or where only a collection level, or bulk, record exists. This is most common with materials from the Manuscripts division. Due to the size of most manuscript collections, they are usually catalogued with only collection-level records. Because the digital library presentations run at the item level, staff must then create a minimal record specific to the digital object.

How do we use the data?

Regardless of source or format, the descriptive data used in *Meeting of Frontiers* describes the original "manifestation" of the "work" and not the digital manifestation. Here a "manifestation" is a "physical embodiment of an expression of a work"–a "work" being "a distinct intellectual or artistic creation".[2] That is, if we were talking about the "work" *War and Peace* then a printed version would be one manifestation; a digitized version would be another.[4] Following is how this data from the original manifestation is augmented for presentation online in *Meeting of Frontiers*.

Search by Format

Users are not offered a view of the complete bibliographic record that lies "behind" the bibliographic presentation on the *Meeting of Frontiers* site. Some fields have only a behind the scenes role that provides enhanced functionality to the online presentation, such as limiting searches by format. Currently, this is not a standard feature of LC web presentations.

This search feature is particularly relevant to *Global Gateway* projects because they are presentations of multiple collections with differing formats. *American Memory* projects usually consist of one collection per presentation, usually in a homogenous format. *Meeting of Frontiers* currently has materials in the following formats: text (manuscripts, printed materials), still images (prints, photographs), moving images, maps, and sound recordings.

The format is indicated via the MARC 985 field, which is a local field used for collection ID codes. *Meeting of Frontiers* adds an additional two-letter code to this, which allows the programming to index items according to type of original object (text, cartographic, still image, etc.).

Current MoF Format Codes (MARC 985a)

ma	manuscripts
pm	printed materials
mp	maps
rs	recorded sound
mi	moving images
si	still images

MODS-based Codes (MARC &Non-MARC datasets)

tx	text (manuscripts and printed materials)
si	still image
mp	cartographic
sr	sound recording
mi	moving image
ob	three-dimensional object
sw	software, multimedia
mx	mixed material
nm	notated music

Table 1. Format codes used to limit searches

The US MARC standards do include a "Type of record" code in the Leader/06 field, which indicates the characteristics of and defines the components of the record. However, LC staff found this code problematic for digital items in several ways. It was used more as an indicator of which AACR2 cataloging rules were followed and also mixes type of resource with mode of issuance and physical carrier. Largely because of this, it was decided to add a separate format code specific to online needs.

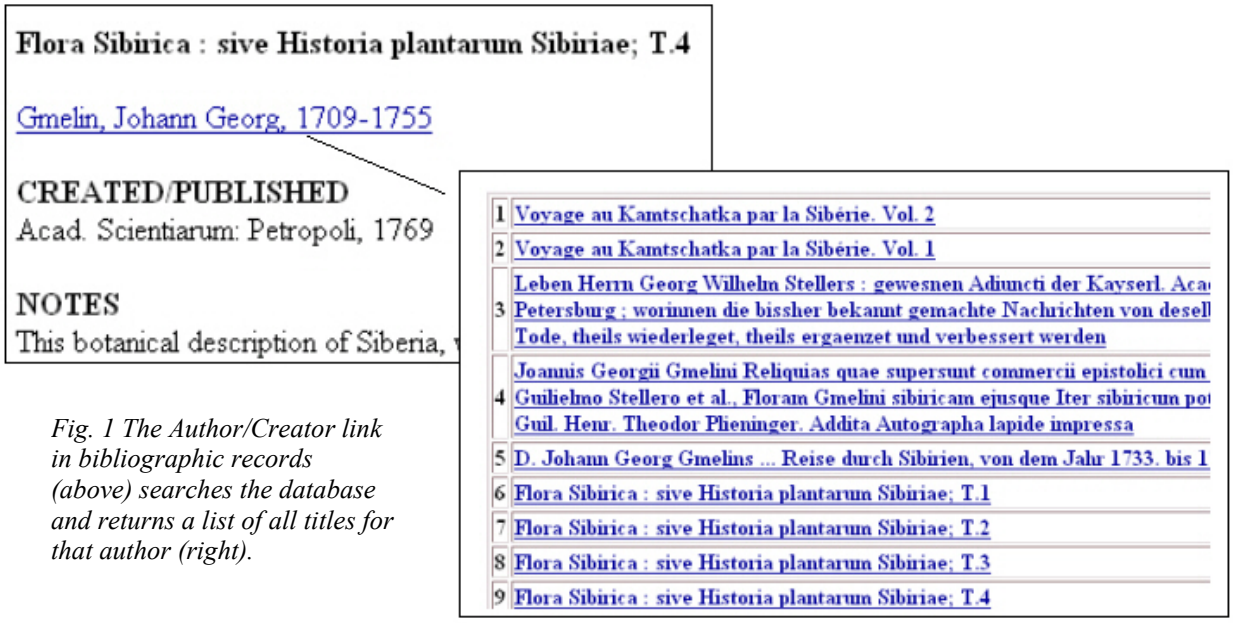

Fig. 1 The Author/Creator link in bibliographic records (above) searches the database and returns a list of all titles for that author (right).

This collection ID plus format code is added to both MARC and pseudo-MARC records. Originally, the programmers established these two-letter indicators.

Because other projects find this functionality desirable, the codes are being modified to more closely match the <typeOfResource> terms in the Metadata Object Description Schema (MODS) (*Table 1*). The format codes will also be added to new non-MARC recordsets to allow combined searching of MARC, pseudo-MARC and non-MARC data.

Name authority "normalization"

Cooperative projects like *Meeting of Frontiers* have a wide variety of bibliographic records taken from many different kinds of sources. These include both records created at the Library of Congress and records received from partner institutions.

Early in *MoF's* development, it was decided to let each institution catalog their contributed materials according to their own standards. This worked well when the project was small but as the size of the site and number of partners scaled up, problems arose with the functionality of the Author/Creator (MARC 100 $a) and Related Names (MARC 700 $a) links.

Online presentations in *American Memory* and *Global Gateways* have a feature where the author/creator and related names fields are a link that searches the project's records and will return a list of all works by that author. This function works well when all the records are from the same source and cataloged according to the same rules, as in *American Memory*.

However, cross-institutional differences in transliteration, cataloging and naming conventions were causing incomplete results to be returned by the search query. One example of the range of variation possible is the famous explorer Ivan Fedorovich Kruzenshtern, who was known as:

- Ivan Fedorovich Kruzenshtern;
- I.F. Kruzenshtern;
- Adam Johann von Kruzenshtern; and
- Adam Johann Krusenstern.

Within LC's ILS system, such discrepancies are solved by referring to an "authority file," which gives the authorized name heading for a particular person. In deciding how to normalize names between *Meeting of Frontiers'* ILS and partner records, it became clear that the one set of records we couldn't easily change were the ILS records. Partner records live in a separate database and are never entered into LC's ILS so a larger range of alterations is possible with that data. Because ILS records match the authority files available online (some exceptions exist for very old records that have not been upgraded), it was decided to normalize all incoming Author/Creator and Related Names data to the authority files online. This allows us to standardize our example above to "Kruzenshtern, Ivan Fedorovich, 1770-1846" for all materials.

Staff "massaging" the data must make sure that the person listed in the authority file is, in fact, the person who is the author of the item. This is done either by referencing birth/death dates or viewing the full record and examining aliases or the source of the authority listing. Below is an example of possible name variants in an authority record. Many are not as detailed and, in the case of more obscure authors, there may be no authority record at all. In case of no match, the team member confirms that, if there are other instances of the same author within the non-ILS data, that they are at least consistent with each other.

HEADING: Benyowsky, Maurice Auguste, comte de, 1746-1786
100 10 |***a*** Benyowsky, Maurice Auguste, |***c*** comte de, |***d*** 1746-1786
400 10 |***a*** Benyowsky, Mauritius Augustus, |***c*** Count de, |***d*** 1746-1786
400 10 |***a*** De Benyowsky, Maurice Auguste, |***c*** comte, |***d*** 1746-1786

400 10 |***a*** Beniowski, Maurycy, |***c*** hrabia, |***d*** 1746-1786
400 10 |***a*** Beniowski, Maurice Auguste de, |***d*** 1746-1786
400 10 |***a*** Beniowski, |***c*** hrabia, |***d*** 1746-1786
400 10 |***a*** Beniowski, Moritz August, |***c*** graf von, |***d*** 1746-1786
400 10 |***a*** Benyovsky, Myricz, |***c*** gryf, |***d*** 1746-1786
400 10 |***w*** nnaa |***a*** Benyovszky, Myric, |***c*** gryf, |***d*** 1741-1786
400 10 |***a*** Benyovszky, Myric, |***c*** gryf, |***d*** 1746-1786
400 10 |***a*** Benyovszky, Myricz, |***c*** gryf, |***d*** 1746-1786
400 10 |***a*** Benyowzky, Maurice Auguste, |***c*** comte, |***d*** 1746-1786
400 10 |***a*** Benyovszki, Myric, |***c*** gryf, |***d*** 1746-1786
400 10 |***a*** Benovskэ, Moric August Aladar, |***c*** hrabe, |***d*** 1746-1786
400 10 |***a*** Benjowsky, Moritz August, |***c*** graf von, |***d*** 1746-1786

Table 2. Variants from Name Authority Record

Another aspect of authority normalization that presents a problem is where the authorized heading is not the most common name for the author. An example would be the German explorer Gerhard Friedrich Müller who came to Russia in 1725 to join the newly founded Academy of Sciences, participated in the Great Northern Expedition and subsequently became secretary of the Academy. Within the scholarly community he is known as Gerhard Friedrich Müller (sometimes Mueller or Muller due to diacritics issues) but the authorized heading for him is Gerard Fridrikh Miller.

Because an ILS item as Miller was already part of the presentation, the new partner records were normalized to that authority heading to make the search function work. The ILS record includes the more common spelling in an Author/Creator subfield (100 $c) so searches for the more likely Müller or Muller would find the ILS record. However, this subfield is not available within *Meeting of Frontiers'* database of non-ILS records. To facilitate searching of partner materials for the more common variants, a bibliographic note (MARC 500) was added to partner records giving the more common form of his name.

Related Digital Items

One of the drawbacks of Meeting of Frontier's digital library/exhibit hybrid approach is that there is not a seamless integration of the two sides. While the interpretative text makes extensive use of linking to the digital library side, the converse is not true. Currently, the only link from the bibliographic records to the interpretative materials is a home page link at the top.

Within *MoF*, the Related Digital Items field (MARC 859) has traditionally been used to relate child records back to the parent record. An example of this is a photo album where the individual photos within the album have been catalogued separately as well as the album as a whole. The individual photos records would use a persistent URL (handle) to link back to the parent album record.

With the May 2003 release, the *MoF* team is beginning to use this field more extensively to link to collection scope notes and related interpretative text. Persistent URLs are being registered for static html pages, and even particular scope notes within pages, and then added to bibliographic data. By making the links persistent URLs, maintenance issues such as link rot can be minimized, particularly in harder to change ILS records.

Fig. 2 Example of links to static pages in bibliographic record via the 859 field in a pseudo-MARC manuscript record

These links are easy to add to the pseudo-MARC records and staff plan to retrofit old records as time goes by. For ILS records, more planning ahead is required. As new materials are selected for digitization and their records upgraded for online presentation, the related digital items (859) fields will be added but retrofitting is not a high priority because of the cataloging resources involved.

Keywords

Another area where we are expanding efforts, particularly for partner data is the uncontrolled subject terms (MARC 653). These do not adhere to standard LC Subject Heading (LCSH) rules or thesaurus conventions– what would be better termed keywords in a web environment. Several of our foreign partners have been providing either keyword or their own subject heading terms.

This data is indexed along with LCSH and presented under the Subjects heading in the display. A lot of this data has been in foreign languages or transliterated. It is the intention of the *MoF* team to make translation of this terminology a higher priority to aid in the search process.

The regular process for LC ILS records includes subject cataloging so we do not anticipate a need to add

keyword entries at this point. However, an area where this is of major concern for LC materials is for manuscript items.

As noted before, LC manuscripts are catalogued at the collection level with bulk records. For online presentation, *MoF* staff prepare separate item level records. These have generally been very minimalist, consisting of author, title, and basic data needed to run the presentation. While this is sufficient to make the display work, it does not help the user find materials via the search engine. It is now the thinking among the *MoF* team to add keywords to the MARC 653 field and also take relevant LCSH from the bulk record and apply them to the item level record.

Short Descriptions

At various times several of our partners have supplied 2-3 sentence summaries or descriptions of the contents of an item. This is something that we would like to incorporate as a standard feature. Books, particularly those in a foreign language, can be somewhat mystifying to the user with nothing more than a title, even when the title is translated.

For non-ILS data, this information has been included as a bibliographic note (MARC 500) because the current database lacks a specific Summary (MARC 520) field. Within the ILS, the MARC 520 field has been primarily used for summaries of children's literature. We would like to see this done for *MoF* records as well. This is not currently done but would add to the user's understanding of the material, particularly for foreign language items.

Multi-lingual Data

A limitation of the current InQuery search engine is that multi-lingual stemming, word use, and use of encoded text are not currently supported. Although bibliographic data is presented online in Latin-1 encoding, the search engine cannot handle diacritics. As a result of this, in pseudo-MARC records, Cyrillic data must be transliterated and diacritics must either be dropped from European languages or specially encoded.

The ILS supports entry but not presentation of seven foreign character sets—the seven JACKPHY languages (Japanese, Arabic, Chinese, Korean, Persian, Hebrew, and Yiddish). Currently, Cyrillic data in the ILS is all in transliterated form. Initial user testing has shown that transliterated titles present problems for both English speakers (who don't know what to make of them) and foreign users (who may not be familiar enough with the Latin alphabet to make sense of the transliteration).

One way to address this issue for English speakers is to include translated titles in the alternative title field (MARC 242). This field is indexed for searching and displayed in bibliographic records. It is not shown in search results or the collection browse lists but that is an area where additional benefit could be gained. An English language translation of the title is not generally done as part of the ILS cataloging process so this needs to be done as part of the MARC record upgrade process. *MoF* team members are actively involved in translating titles for partner materials to aid both searchability and usability.

The ILS is moving toward Unicode implementation in 2005 or so. Policy decisions about whether to enter Cyrillic data when the new Unicode version of Voyager is installed and how/what to retrofit are still far down the road. When coupled with the uncertainty of how the present search engine can be adapted to Unicode text, including vernacular data in online presentations is still some time down the road.

However, *MoF* staff are taking the long view on this and, with an eye to the future, beginning to include Cyrillic data in the pseudo-MARC database. This is in addition to the transliterated and translated title and author data. This information is not in exportable fields but could easily be mapped to them when it does become possible to present this data online. It is the team's outlook that showing titles in the original Cyrillic script, as well as in transliterated and translated forms would provide additional benefit to multi-lingual users.

Conclusion

Cooperative digital library projects, particularly international ones where metadata standards can vary greatly, are in a unique position to highlight the balancing act of online presentation. There are good, not-so-good, and essentially bad aspects to implementing the presentation of digital surrogates in a highly structured record environment.

One bad feature is the difficulty reconciling cross-national, cross-institutional, and even cross-division (that is, even within the same institution) differences in metadata generation that affect site functionality. The name authority issues discussed here are more about how metadata is generated in various contexts than the container used for exchange. Another bad feature is technical limitations that hinder full presentation of existing data, such as the InQuery problems with encoded languages.

A not-so-good aspect of highly structured bibliographic records is that the initial cataloging has been done for the offline world. This means that records must be upgraded to meet purely functional needs, such as the format codes, and help users find what they want—via translated titles and summaries. With cataloging resources always at a premium for ILS records and limited staff for pseudo-MARC massaging, these requests have to be balanced against the value added. User-focused upgrades sometimes fall by the wayside in deference to functional upgrades.

A highly structured metadata environment is good because it's an established and tested construct that allows repurposing of vast quantities of offline data. It also provides a framework for mapping pseudo-MARC and non-MARC data, as well as value-added information such as summaries and keywords. At the same time it can help address needs not met by the site programming, such as links to static pages in the framework.

The authors are primarily concerned with day-to-day aspects of building an operational site that will scale upwards successfully as additional materials are added. In the digital library "real world," there will always be trade-offs between site functionality, technical limitations, user expectations, and a highly structured metadata system. Our goal is to make the data serve the reasonable information seeking expectations that users bring to the site.

References

[1] Added Entry Fields (70X-75X), 2003. http://www.loc.gov/marc/bibliographic/ecbdadde.html#mrcb700

[2] Bennett, Rick, Brian F. Lavoie, and Edward T. O'Neill. "The Concept of a Work in WorldCat: An Application of FRBR" *Library Collections, Acquisitions, and Technical Services* 27(1) (Spring, 2003), and available online at http://www.oclc.org/research/publications/archive/2003/lavoie_frbr.pdf

[3] Exhibitions (Library of Congress), 2003 http://www.loc.gov/exhibits/

[4] Functional Analysis of MARC 21 (Library of Congress), 2003. http://www.loc.gov/marc/marc-functional-analysis/functional-analysis.html

[5] Global Gateway: World Culture & Resources (Library of Congress), 2003. http://international.loc.gov

[6] Home Page: American Memory from the Library of Congress, 2003. http://memory.loc.gov

[7] Library of Congress Authorities, 2003. http://authorities.loc.gov/

[8] Main Entry Fields (1XX), 2003. http://www.loc.gov/marc/bibliographic/ecbdmain.html#mrcb100

[9] MARC Standards, 2003. http://www.loc.gov/marc/

[10] Meeting of Frontiers: About the Site, 2003. http://frontiers.loc.gov/intldl/mtfhtml/mfabout/mfabout.html

[11] Meeting of Frontiers: Home Page, 2003. http://frontiers.loc.gov

[12] MODS User Guidelines: Elements, 2003. http://www.loc.gov/standards/mods/mods-userguide-elements.html#typeofresource

[13] Note Fields (Part 1: 50X-53X), 2003. http://www.loc.gov/marc/bibliographic/ecbdnot1.html#mrcb520

[14] Subject Access Fields (6XX), 2003. http://www.loc.gov/marc/bibliographic/ecbdsubj.html#mrcb653

[15] Title and Title-Related Fields (20X-24X), 2003. http://www.loc.gov/marc/bibliographic/ecbdtils.html#mrcb242

† The views and opinions expressed herein do not necessarily state or reflect those of the Library of Congress, United States Government or any agency thereof.

[1] http://frontiers.loc.gov/intldl/mtfhtml/mfabout/mfabout.html where it states, among other things, that the "design of the site combines elements of two approaches used by the Library of Congress in presenting educational material in electronic form: the collections-based approach of the National Digital Library's American Memory program, and the method of integrating items from many collections to tell a single story that is used in Exhibitions: An Online Gallery."