What We've Learned from the RLG Partners Metadata Creation Workflows Survey

Leighann Ayers
University of Michigan

Beth Picknally Camden
University of Pennsylvania

Lisa German
Pennsylvania State University

Peggy Johnson
University of Minnesota

Caroline Miller
University of California, Los Angeles

Karen Smith-Yoshimura
OCLC Research

A publication of OCLC Research

What We've Learned from the RLG Partners Metadata Creation Workflows Survey
Leighann Ayers, Beth Picknally Camden, Lisa German, Peggy Johnson, Caroline Miller and Karen Smith-Yoshimura, for OCLC Research

January 2009

OCLC Research
Dublin, Ohio 43017 USA
www.oclc.org

ISBN: 1-55653-409-4, 978-1-55653-409-6
OCLC (WorldCat): 298738330

Please direct correspondence to:
Karen Smith-Yoshimura
Program Officer
karen_smith-yoshmura@oclc.org

Suggested citation:
Ayers, Leighann, Beth Picknally Camden, Lisa German, Peggy Johnson, Caroline Miller and Karen Smith-Yoshimura. 2009. What We've Learned from the RLG Partners Metadata Creation Workflows Survey. Report produced by OCLC Research. Published online at:
http://www.oclc.org/programs/publications/reports/2009-04.pdf

Contents

Tables and Charts

Introduction

RLG partners who have engaged in discussions about renovating descriptive practices have expressed much interest in sharing best practices for streamlining metadata creation workflows for describing a wide range of resources. With staff creating and managing both MARC and non-MARC metadata, administrators are eager to know what workflows work best in different environments that could be applied to their own.

In October-November 2008 we conducted a survey of the RLG partners directed to department heads or directors of units (in a library, archive, museum, institution repository, etc.) who were responsible for creating non-MARC metadata – either solely or in addition to MARC metadata. We received 134 responses from 67 RLG partner institutions (libraries, archives, museums). Appendix 1 comprises tables and charts of the responses and Appendix 2 the list of the RLG partners represented in the survey responses.

Please note that these responses represent individual rather than institutional perspectives. Refer to Appendix 2 for the number of individual responses from each institution. The working group focused on the 121 responses from those who created at least some non-MARC metadata.

What we learned

The working group had hoped that the survey would point to tools and resources for streamlining metadata workflows that might be shared within the RLG Partnership and that could be adapted locally. However, survey responses suggest that the tools being used are very localized, and no one tool kit is being used. Most tools mentioned were generic, i.e., standard software and programming (Java, Perl, XSL-based custom code) or a combination of these that has been locally customized for metadata creation and non-MARC enrichment. The great variation of tools used resembles the responses to the 2007 RLG Programs Descriptive Metadata Practices Survey Results.[1]

1. Smith-Yoshimura, Karen. 2007. RLG Programs Descriptive Metadata Practices Survey Results. Available online at: http://www.oclc.org/programs/publications/reports/2007-03.pdf

We found it encouraging that among those who create both MARC and non-MARC metadata, two-thirds used the **same** staff. Furthermore, 80% reported that creating non-MARC metadata was part of their "routine workflows." Given this characterization of non-MARC metadata creation as routine, we found it surprising that only just over half (56%) have training programs for teaching staff how to create metadata. We speculated that perhaps the word "program" implied something more formal than intended. Only those who said that they had a training program were asked what their primary approach was – and 70% said it was one-to-one rather than group training. Only nine (16%) had training materials that could be shared with others. We deduce that institutional routines are not yet standard enough for inter-institutional collaboration. However, there may be opportunities for cooperative efforts that include training.

The contacts who are willing to share their metadata training documentation and the URLs for materials where available are listed in Appendix 3.

Most respondents (70%) have tools to convert between MARC and non-MARC formats. Respondents named many more cross-walk schemas used than were offered in the survey (which came from the lists on the LC and OCLC Web sites). Only 11 schemas were used by 18% or more of 61 respondents. More people used cross-walk schemas **from MARC** to something else than schemas **to MARC**. The juxtaposition of these results with how institutions choose to expose their metadata would indicate a disconnect between library practices in the units creating non-MARC metadata and user discovery methods. .We suggest that a major hurdle that may need to be overcome is the assumption that people searching for resources will start with the local website, repository, catalog – instead of on the Web (where most people start, according to the research into discovery practices).

We are reassured that libraries are exposing their data in multiple ways. Although the catalog is the first method of exposure (76%) and other methods of local exposure rate high such as the institution's Web site (67%) and the institution's digital repository (68%), a relatively high percentage exposes metadata through union catalogs (63%), crawlers like Google, Yahoo, MSN (56%), and OAI-PMH harvesters (48%). Small percentages expose data to SRU/SRW queries (5%), Flickr (10%), and YouTube (2%). We wonder whether these percentages will increase over the next few years.

We acknowledge that it's difficult to compare staffing for metadata creation across institutions given different organizational patterns, hierarchies, and staffing classifications. However, comparing the percentages for MARC and non-MARC metadata creation indicates some differences in staffing patterns. More professional employees, archivists, and curators are involved with the routine creation of non-MARC metadata than MARC metadata. For non-MARC metadata creation, paraprofessionals or bargaining units were used less often than for MARC metadata, but students or part-time/temporary help was used more. This use of part-time/temporary staff in the "routine" creation of non-MARC metadata may indicate that in fact it is **not** yet routine. Outsourcing or using

contractors is used far less for non-MARC metadata (13%) than MARC metadata (42%). The results for outsourcing MARC metadata are verified by the 2008 Primary Research Group report, Academic Library Cataloging Practices Benchmarks,[2] which states that 41.56% of libraries in the sample outsource authority control, 30% obtain new bibliographic records through outsourcing, and 30% outsource the acquisition of updated headings and bibliographic records. We suggest that libraries and service vendors may wish to explore opportunities to contract out **non-MARC** metadata creation.

Respondents reported that they are focusing on several areas when seeking to streamline metadata creation workflows. The 108 respondents could select numerous aspects. Descriptive metadata elements were selected by 82% of respondents; procedures and standards were the next most frequently selected area (78% and 75% respectively). Comments noted that respondents are looking at reducing duplication of effort through various means, including repurposing data and capturing data from other systems and MARC records.

When asked about the types of materials to which metadata creation tools are applied, 78 respondents identified various material types. By far the most common material type identified was archival materials/finding aids—selected by 73% of the 78 respondents to this question. Electronic texts were next (50%), followed by visual (2-D) materials (45%). Of possible interest is that only 33% are applying the metadata creation tools to electronic dissertations, theses, and reports; this may, however, reflect responses from institutions that do not have these materials in their collections or acquire the metadata from other sources and do not create the metadata locally.

Slightly more than half (54%) of respondents reported that they do not have routine procedures for maintaining and updating non-MARC metadata. We found this surprising, given the historical attention given to maintaining and updating MARC records in the catalog. We did not ask respondents if they have procedures for maintaining and updating MARC records, but can guess that close to 100% of those respondents with catalogs of MARC records would respond "yes." We wonder whether non-MARC metadata maintenance is hindered by the lack of widely available tools or the distributed environment in which non-MARC metadata is created. Of the respondents (46) who do maintenance on non-MARC metadata, 89% reported that they enrich/enhance records, 85% reported they correct typos, and 65% work on authority control or controlled vocabularies.

We were struck by the comments from respondents that indicate that this is still a fluid time in our profession and that organizations are in flux, too. Several noted that they were just developing tools or were in the process of restructuring their workflows.

2. Primary Research Group. 2008. Academic Library Cataloging Practices Benchmarks. Abstract and purchase information available online at:
http://www.researchandmarkets.com/reports/659569/academic_library_cataloging_practices_benchmarks

"We are still in the infancy of figuring out how best to integrate non-MARC metadata into our workflow."

"We are restructuring staff structures in our technical services area at the moment, and streamlining metadata creation is a key requirement (including outsourcing where appropriate and integrating workflows related to the digital repository into routine work)."

"We're still in the 'treat each new collection as special' workflow, but are moving towards more standardization."

"Metadata creation is spread out over 3 different units in the institution with no centralized workflow or central "enforcement" of standards, practices and authority control."

Questions for possible future research

- Will use of MARC metadata decrease?
- Will more libraries expose their metadata to search engines, SRU/SRW queries, Flickr, YouTube, etc?
- Will we see a shift in the assignment of responsibilities parallel to cataloging with MARC data, so that fewer professional librarians will need to be involved in routinely creating non-MARC metadata?
- Will we see more routine maintenance of non-MARC metadata?

Appendix 1: Survey results

Q1: Metadata created

Non-MARC metadata only	22	18.2%
Both MARC and non-MARC metadata	99	81.8%
Total responses	121	

Q2: Do the SAME staff create both MARC and non-MARC metadata?

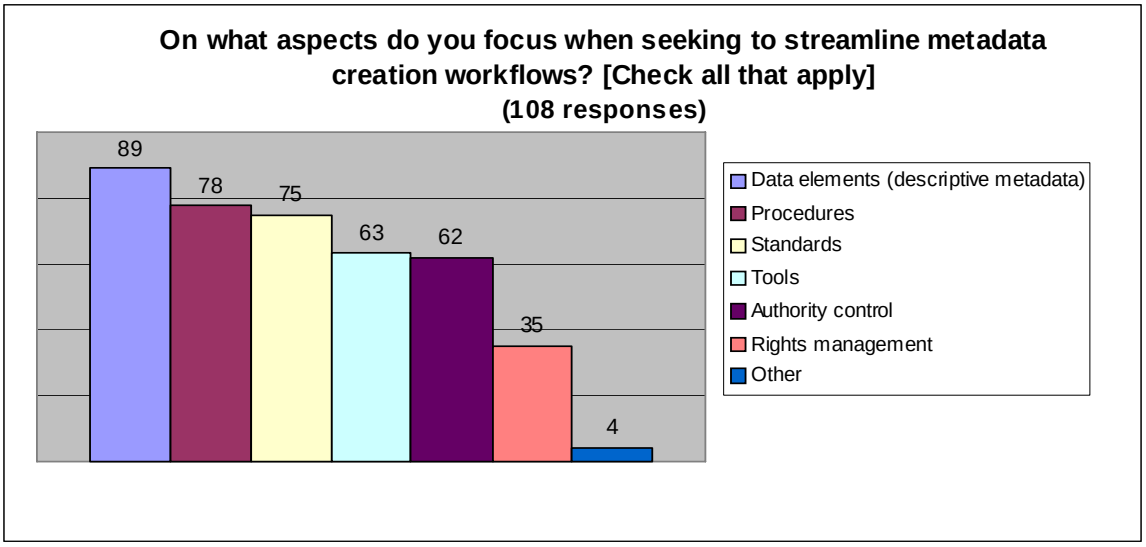
Asked only of those who create both MARC and non-MARC metadata in response to question 1.

Yes	64	66.0%
No	33	34.0%
Total responses	97	

Q3: Is creating non-MARC metadata part of your routine workflows?

Yes	86	79.6%
No	22	20.4%
Total responses	108	

Q4: On what aspects do you focus when seeking to streamline metadata creation workflows?



Other:

Metadata reuse (2)
Preservation metadata (1)
User interface (1)

Procedure comments:

- Defaults within tools
- Outsourcing, capturing metadata from other systems, capturing metadata from resource creators, reducing duplication of effort (esp. rekeying of data)
- Mapping data fields to metadata schema
- Responsibility --who does what when?

Tools comments:

- Development of computer-based workflow tools occupies much development time
- We are moving towards harvesting some metadata from MARC records

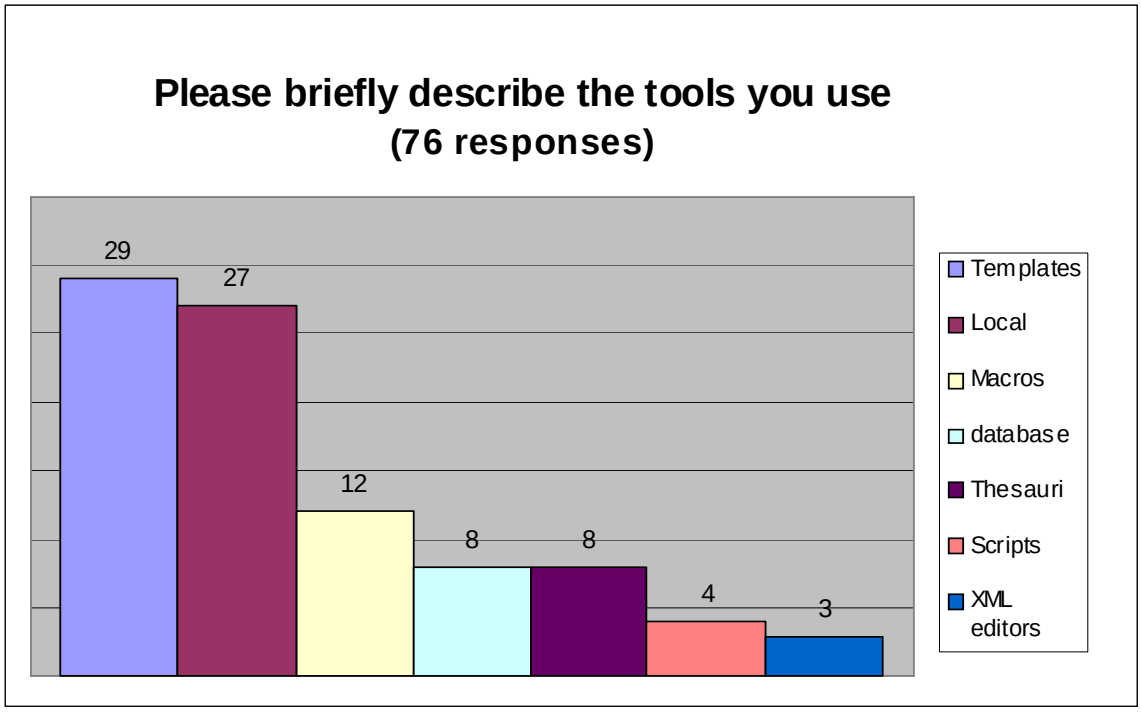
Other comments:

- NOTE: As an institutional archive, determining rights management is not a complicated process, therefore it does not lend itself to streamlining.
- Varies with nature of resource, intended use, etc

Q5. Do you have tools to help in automating the creation of metadata? (automatic metadata extraction, macros, templates, etc.?)

Yes	86	79.6%
No	22	20.4%
Total responses	108	

Q6. Please briefly describe the tools you use



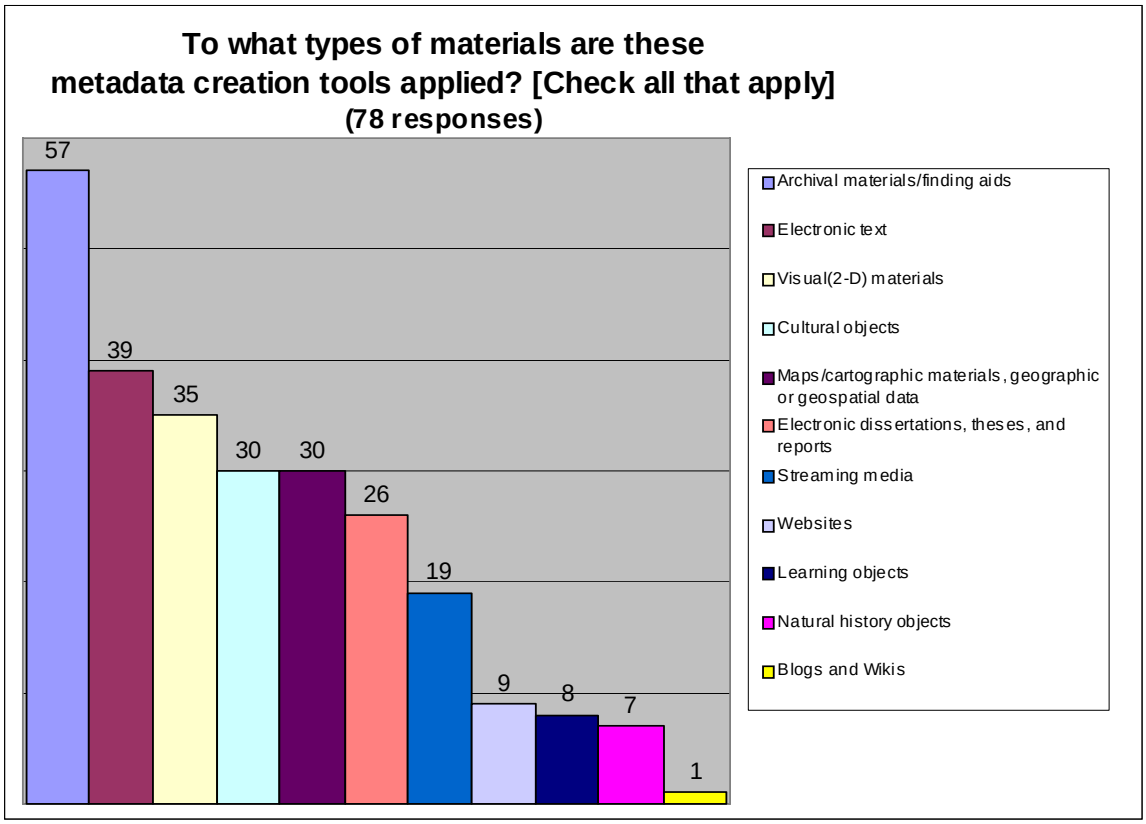
Asked only of those who said they had tools to help in automating the creation of metadata in response to the previous question.

Other tools mentioned once: ERMS, MARC editor, Metadata map, MODS editor, OAI-PMH generator, spreadsheets, Web harvesting

Specific tools reported: MarcEdit (12), MSExcel (6), XMetal (6), Macro Express (5), MSAccess (5), Archivists toolkit (4), ContentDM (4), Dspace (4), PHP (4), XSLT (4), Filemaker Pro (3), JHOVE (3), MSWord (3), oXygen (3), Archon (2), EAD XPress (2), LUNA (2), MARC Report (2), MARCView (2), NoteTab (2), PERL (2), WebGenDB (2), XQuery(2).

Tools mentioned once: Adobe Bridge for XMP, Archive-It, Artesis TEAMS, BePress, Catalogers Desktop, Catalogers Toolkit, dLOC, DROID, EAD2HTML, ePrints, ETD-bd, JEdit, MARC2EAD, RelaxNG, SQL, TMS, USEMARCON, XForms, XMLSpy

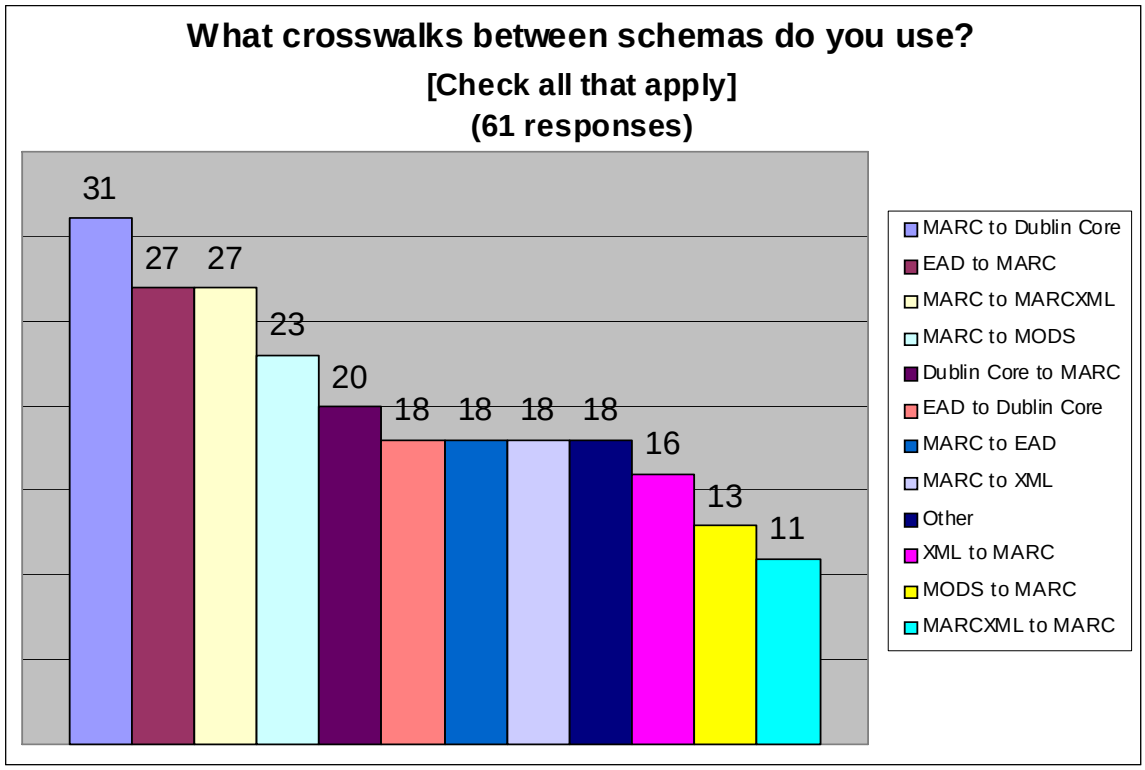
Q7. To what types of materials are these metadata creation tools applied?



Q8. Do you have a tool to convert MARC data into non-MARC formats or vice-versa?

Yes	63	70.0%
No	27	30.0%
Total responses	90	

Q9. What crosswalks between schemas do you use?



Asked only of those who said they had a tool to convert between MARC and non-MARC formats in response to the previous question.

Other: MODS to Dublin Core (3), CDWA to VRA (2), EAD to MODS (2), ISAD to MARC (2), MODS to CDWA (2), MS Access to MARC (2).

Crosswalks mentioned once: Dublin Core to MODS; Dublin Core to VRA; EAD to MODS; EAD to XML; FGDC to MODS; FGDC to Dublin Core; ISAD(G) to Dublin Core; MARC to MADS (authority files); MARC to tab-limited; MARCXML to Dublin Core; MARCXML to ONIX; MARCXML to Solr-ready XML; METS to FOXML; MODS to Dublin Core; MODS to METS; MODS to MIDAS; MODS to VRA; MODS to XML; MS Access to XML; MS Excel to XML; tab-limited to MARC;TEI to Dublin Core; VRA to CDWA: VRA to Dublin Core

Comments:

- Lots of custom formats to MODS, Dublin Core, and EAD. Starting to look at CDWA Lite ‹-› MODS. We're also starting to look at SKOS and other formats for authority data, and mapping between source formats for things like TGM and the Getty vocabs into SKOS for a local terminologies service.

- Much of the work we do, especially relating to digital object metadata, is still done with locally developed crosswalks, since our inputting/CMS for these types of materials does not use XML-based encoding.

Q10. Do you have tools that systematically enhance existing non-MARC metadata?

Yes	24	24.0%
No	62	62.0%
Don't know	14	14.0%
Total responses	100	

Q11. Briefly describe the tools you use – include any URLs that provide more detailed information.

Asked only of those who said they had tools that systematically enhance non-MARC metadata in response to the previous question.

Only 21 responses. One noted that the same tools are used as those for metadata creation. Three are in the process of developing tools internally; almost all of the rest use some combination of software and standard programming (Java, Perl, XSL-based custom code) to create customized tools in-house.

Q12. Do you have training programs for teaching staff how to create metadata?

Yes	56	56.0%
No	44	44.0%
Total responses	100	

Q13. Describe your *primary* approach to metadata training

One-on-one	39	69.6%
Group training	17	30.4%
Total responses	56	

Asked only of those who said they had training programs for teaching staff how to create metadata in response to the previous question.

Q14. Have you developed training materials that you can share with others?

Yes	9	16.1%
No	47	83.9%
Total responses	56	

Asked only of those who said they had training programs for teaching staff how to create metadata in response to the previous question.

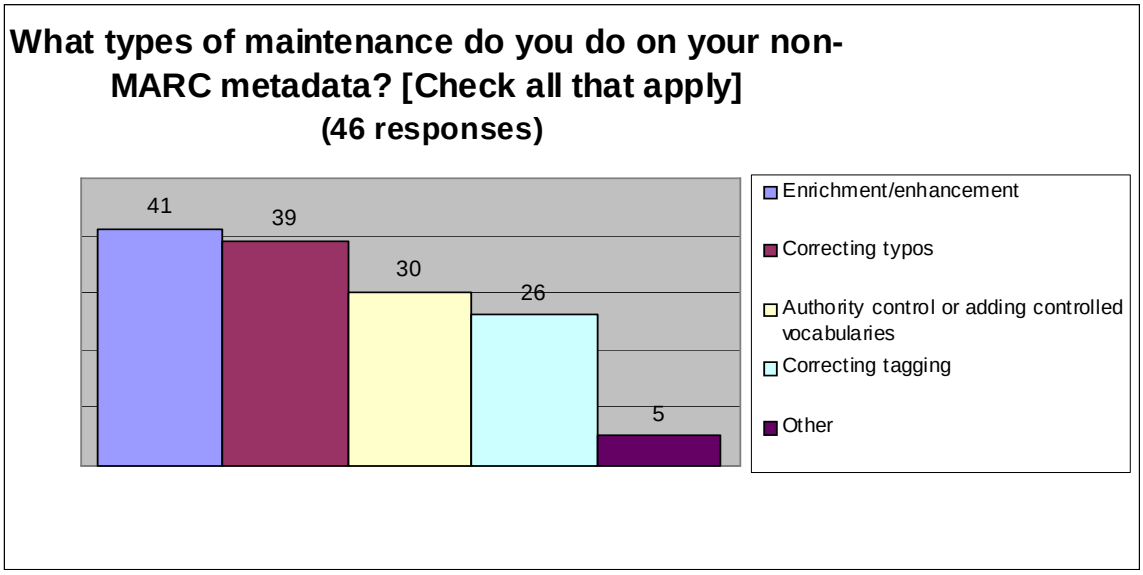
Q15 Please give the URL for the training materials or the contact person.

The resources and contact persons we received permission to share are listed in Appendix 3.

Q16. Do you have routine procedures for maintaining and updating non-MARC metadata?

Yes	46	46.0%
No	54	54.0%
Total responses	100	

Q17. What types of maintenance do you do on your non-MARC metadata?

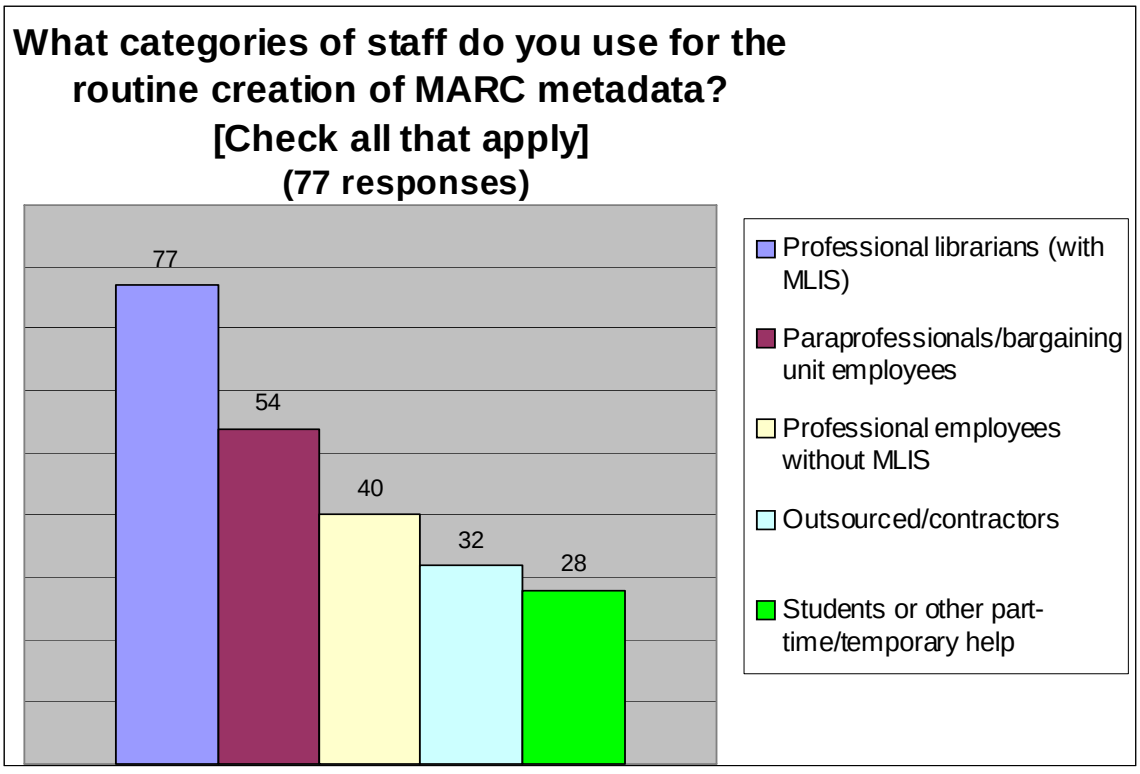


Asked only of those who said they have routine procedures for maintaining and updating non-MARC metadata in response to the previous question.

Other:

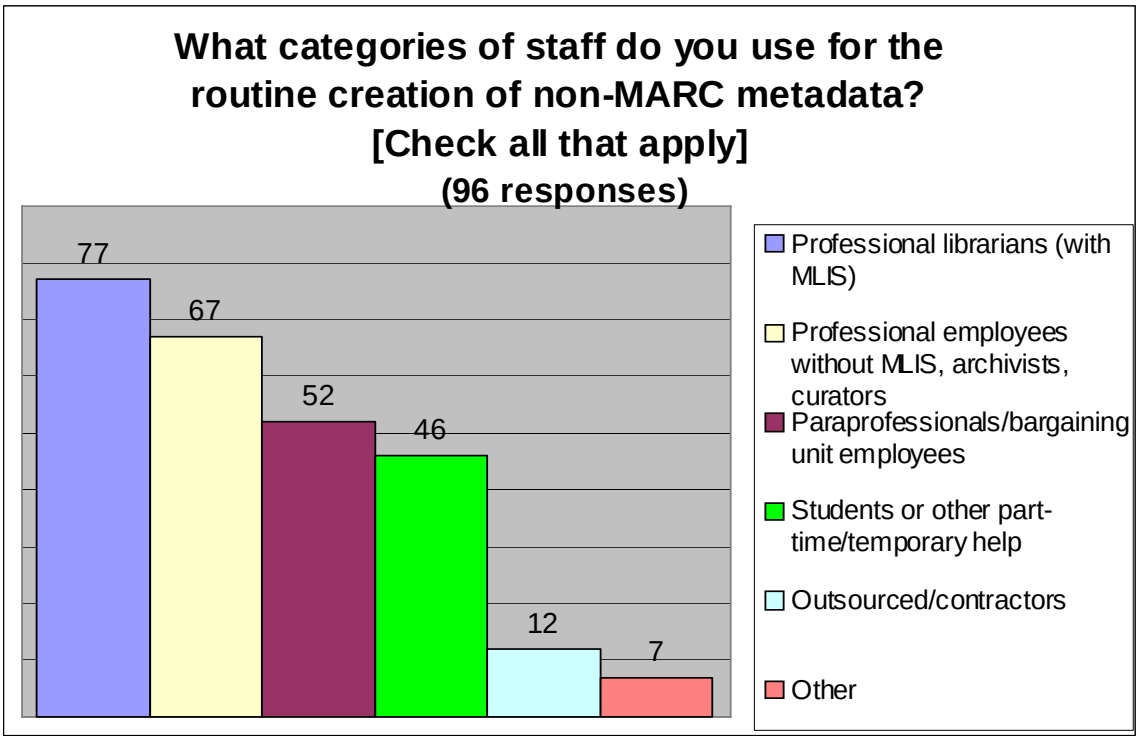
- Currently our process for enrichment involves updating the original source of metadata (usually MARC) and re-extracting to update the non-MARC metadata. This can be batch or record by record
- Version migration (EAD 1.0 to EAD2002; MODS version updates)
- External validation against local best practices
- Crosswalking locally created legacy schemas to standard schemas
- Web templates permitting curatorial staff to vet and approve publicly-displaying online captions for digital images of collection material.

Q18. What categories of staff do you use for the routine creation of MARC metadata?



Comment: We don't create MARC metadata. We create records in our own schema, and then "express" them in the MARC format for licensees who want records that way.

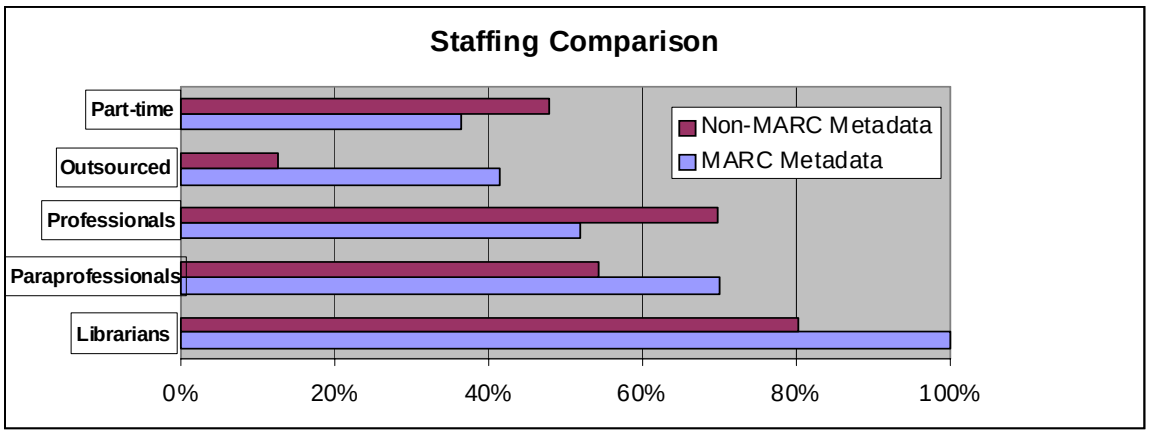
Q19. What categories of staff do you use for the routine creation of non-MARC metadata?



Other:

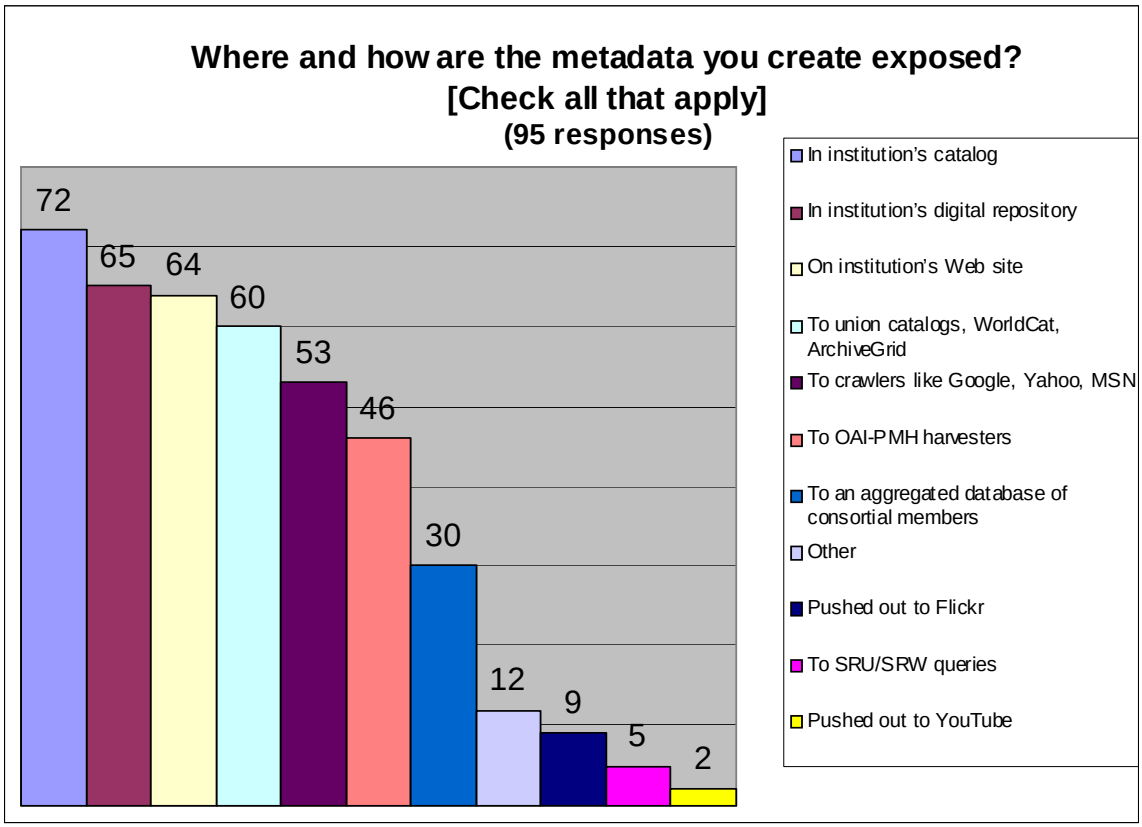
- Volunteers and unpaid "helpers" including faculty and scholars with [subject] expertise (5)
- [Content] contributors to the IR (1)
- External project partner (1)

Comparative staffing for routine creation of MARC vs. non-MARC metadata



This chart compares the staffing categories by percentage of those responding to the previous two questions.

Q20. Where and how are the metadata you create exposed?



Other: ArtStor (2), Z39.50 queries (2). Internal use only (2), Working on exposing metadata (3), Selling metadata (1), Link to Wikipedia (1), Part of aggregated interface (1)

Additional comments:

- Our metadata librarian position is vacant and we are still in the infancy of figuring out how best to integrate non-MARC metadata into our workflow. Aside from EAD, we do not have very sophisticated tools and would be interested in hearing about more.
- We plan to expose to OAI-PMH harvesters, but have not yet.
- An OAI harvesting capability is being developed.
- Our DAM system and collection management system are for internal users only.

- We are restructuring staff structures in our technical services area at the moment, and streamlining metadata creation is a key requirement (including outsourcing where appropriate and integrating workflows related to the digital repository into routine work).

- We're still in the "treat each new collection as special" workflow, but are moving towards more standardization. We're looking for the right balance - putting everything in CONTENTdm with the default metadata template is too far in that direction in our opinion. We have more and more standard procedures, but this is an ongoing process. I also forgot to mention earlier we do a great deal of mapping to and automatic generation of technical metadata, especially for still images, audio, and video.

- We are in the process of building our digital repository. Once built, the metadata we are creating will be exposed through the repository and our website. Tools are also being constructed to help streamline metadata creation.

- Not yet created but is being designed.

- Really wanted to note data enhancement for geographic data. We also record some relational data for 3D objects, essentially planetary information referencing object orientations.

- Metadata creation is spread out over 3 different units in the institution with no centralized workflow or central "enforcement" of standards, practices and authority control.

Appendix 2: RLG Partners represented in the survey responses

Multiple responses were encouraged from within the same institution if they represented different units creating metadata. If more than one response was received, the number is indicated within parentheses. A total of 134 responses were received from 67 RLG Partners during the survey period 2008-10-06 to 2008-11-07.

American Museum of Natural History	New York University
Arizona State University	Newberry Library
Art Institute of Chicago	Oregon State University (2)
Athenaeum of Philadelphia	Pennsylvania State University
Autry National Center of the American West	Princeton University (2)
Bibliothèque nationale de France	Rutgers University
Boston Public Library (2)	Smithsonian Institution (2)
Brigham Young University (3)	Stanford University
British Library (9)	Swiss National Library
California Institute of Technology	Syracuse University
Chemical Heritage Foundation (2)	Temple University
Sterling and Francine Clark Art Institute	Trinity College Dublin (2)
Columbia University	University of Aberdeen (6)
Cornell University (2)	University of Alberta (2)
Duke University	University of Arizona
Emory University	University of California, Berkeley (5)
Folger Shakespeare Library	University of California, Los Angeles (6)
The Frick Collection and Frick Art Reference Library	University of Cambridge (3)
Getty Research Institute (4)	University of Chicago (2)
Huntington Library, Art Collections, and Botanical Gardens	University of Edinburgh (6)
Imperial College of Science, Technology, and Medicine	University of Florida (4)
Indiana University Bloomington (2)	University of Leeds
International Institute of Social History (2)	University of Maryland (2)
Library Company of Philadelphia	University of Miami (2)
Library of Congress (3)	University of Michigan
Linda Hall Library of Science, Engineering & Technology	University of Minnesota (4)
Metropolitan Museum of Art	University of Oxford (2)
Minnesota Historical Society	University of Pennsylvania (3)
National Archives and Records Administration	University of Texas at Austin (3)
National Library of Australia (4)	University of Toronto
National Library of New Zealand	University of Warwick
National Library of Scotland	University of Washington
The New School (2)	Yale University (3)
New York Public Library (2)	

Appendix 3: Contacts and URLs for metadata training materials

Archives Florida publishes EAD training materials under "Learning Resources" at http://www.fcla.edu/dlini/OpeningArchives/. Please contact John Nemmers at jnemmers@ufl.edu for more information or questions.

Kayla Willey (kayla_willey@byu.edu) may be contacted about the training materials Brigham Young University has developed.

For information about training materials used at the Bibliothèque nationale de France, contact Aline Bouchard at aline.bouchard@bnf.fr.

The Digital Library of the Caribbean (http://www.dloc.com/?c=dloc&m=hitdigit) has a training manual describing the tools used by its Caribbean partnership at http://www.dloc.com/?c=dloc&m=hitmanual. The University of Florida documents its expanded mapping from METS to MARC at http://www.uflib.ufl.edu/ufdc2/technical/Metadata/metadata.htm. Contact Laurie Taylor at Laurien@ufl.edu for more information.

The Getty Research Institute publishes its training materials for Getty vocabularies at http://www.getty.edu/research/conducting_research/vocabularies/training.html. The PowerPoint presentations the Getty uses for training are available at http://www.getty.edu/research/conducting_research/vocabularies/training.html.

The Getty's editorial manuals are available at http://www.getty.edu/research/conducting_research/vocabularies/editorial_guidelines.html. For more information, or to learn about licensing and/or contributing to the Getty vocabularies, contact vocab@getty.edu.

University of Maryland publishes its training documentation at http://www.lib.umd.edu/dcr/publication. Jennie Levine may be contacted at levjen@umd.edu.

Lara Friedman-Shedlov at the University of Minnesota has developed training materials for EAD and XSLT and you can contact her directly via e-mail at ldfs@umn.edu.

Training materials used for the University of California, Los Angeles' Digital Library Project are available at http://unitproj.library.ucla.edu/cataloging/policies/index.cfm#diglib.

Contact Claudia Horning at chorning@library.ucla.edu for more information.