Repository Metadata: Challenges of Interoperability

Wendy C. Robertson

University of Iowa
Repository Metadata: Challenges of Interoperability

Wendy C. Robertson
The University of Iowa Libraries
This presentation builds on Marisa Ramirez and Nancy Fallgren’s webinar “Making the Most of Your Descriptive Metadata: Planning, Transforming, and Re-using” in the ALCTS Institutional Repositories Webinar Series.

They discussed using various silos as sources for your metadata and gave practical information on how to transform metadata. I highly recommend this webinar, as well as the others in the ALCTS series.

I am going to begin with a quick overview of some general concepts.
NISO’s 2007 publication *A Framework of Guidance for Building Good Digital Collections* includes a useful overview of metadata and provides links to a variety of standards. As with many publications, it is focused on cultural objects rather than scholarly output.

It lists general metadata principles which create a solid foundation for further discussion.

I think of the first principle as: know your audience, know who you are sharing with and know your standards. These themes will come up throughout the presentation.
This second principle regarding interoperability is directly connected to the subject of the presentation.
The third principle is also relevant to interoperability, but it is perhaps one of the hardest to fully implement in a repository.

Metadata Principle 3:
Good metadata uses authority control and content standards to describe objects and collocate related objects.

A Framework of Guidance for Building Good Digital Collections
http://framework.niso.org/
Metadata Principle 4:
Good metadata includes a clear statement of the conditions and terms of use for the digital object.

*A Framework of Guidance for Building Good Digital Collections*
http://framework.niso.org/

Rights and terms of use are particularly important when sharing repository metadata.
These last 2 points are not as relevant, but are still good to review.
There are a few features of repositories that I think make their metadata different. Repository content is a mixture of items not appearing elsewhere (such as working papers), copies of items found elsewhere, and different versions of content, such as an author’s version of a published article.
Content may be submitted by the author. Content may be submitted by the author or have metadata otherwise created outside the library. While self-submittal should be encouraged, it is atypical and has data of variable quality. Content also may be added in a largely automated process.
Typically, repositories include primarily textual materials. These items are frequently PDFs but may be in a document format. The relatively few types of formats and their full text searchability means you may be more flexible on some metadata requirements.

Scholarship and research is not limited to textual items. Repositories can include images, video, learning objects etc. These all have different issues for identification and require more careful metadata. Datasets require much greater metadata expertise. Usually the creator of the data must provide information regarding how the data was created, what it is and how it can be used. This is a developing area in repositories and I look forward to learning from others how best to provide the metadata datasets need.

This breadth of possible formats in a repository creates challenges. The less homogenous the material, the harder it is to have tight standards.
Interoperability
Structured, consistent metadata can be mapped elsewhere. Granularity helps so that mapping can be unambiguous.

Structured, consistent metadata can be mapped elsewhere. This means both data coming to you and data going out. Specificity and granularity allow clear mapping, so try to use Qualified Dublin Core or even MODS. There are plenty of crosswalks you can review and use. Remember, you don’t need to share all fields. Some fields have local use but have little purpose for others.

There have been several excellent articles on shareable metadata and I have included some of them in the bibliography. In particular I’ll mention Shreeves’ and others’ article in First Monday which states “At the most basic level, institutions who contribute metadata through whatever means should consider the content and consistency of their metadata.”
It goes without saying but always look for standards. Unfortunately, there are not yet general repository specific standards. The NISO guide I mentioned previously includes links to some standards. Practices used by other consortia or institutions can be a good starting point. You don’t need to reinvent the wheel and you may interoperate with their repository in the future. Sites provide anything from very short guidelines about their fields to very extensive documentation. The UK, Australia and Germany are ahead of the United States on many repository issues, so don’t be insular when you explore repository metadata.
Examples of Standards & Guidelines

• NCSU:

• IDEALS:
  https://services.ideals.illinois.edu/wiki/bin/view/IDEALS/MetadataPolicy

The following are a few that I happen to be aware of.

NC State has fairly brief but clear descriptions of their metadata fields.

IDEALS has a longer, very helpful listing of fields. Their list includes required and optional elements, fields available only to collection administrators and thesis specific fields. This is the most helpful site I have seen for my needs.
UKOLN is a research organization concerned in part with digital libraries. They have quite a bit of general information about metadata as well as some repository specific information.

Metadata Advisory Committee for Australian Repositories was established to provide recommendations and advice on metadata requirements for digital repositories.

Max Planck Society’s eDoc Server has several technical documents that are interesting when you start needing schemas and scripts.

As with all things on the web, guidelines are apt to change. I have found proposed standards that were never really followed. I have found good guidelines pop-up where they were lacking before. Hopefully there will be open repository standards in general use soon.
More of a model that standard is the Scholar Works Application Profile which maps Dublin Core fields for scholarly works into a FRBR model. It is a UKOLN project. I am unsure how it is being used at this time.

SWAP

Scholar Works Application Profile

- http://www.ukoln.ac.uk/repositories/digirep/index/Scholarly_Works_Application_Profile
- http://dublincore.org/scholarwiki/SWAPD_SP
As you look at your metadata and interoperability, ask yourself: Who are the users of the collection? Are they in a specific discipline or from all areas? Do subsets of the collection have specialized users or special interoperability needs? Assuming the repository is primarily open access content, most of your use will probably come from outside your institution. However you should pay particular attention to the needs of your funding institution. Be sure you understand the purpose of your repository.

You probably get some information from elsewhere and also share your data with others. These two sides of interoperability have somewhat different issues. When your data is going elsewhere, you have the ability to make it meet the needs of that community. Depending on the importance of the group, you may or may not alter your practices. For example, if the University administration requests something, you will probably do it, but if you are voluntarily contributing to a collaborative project you may just want to meet the minimum standards. For data coming in, it often is completely up to you to map, normalize and enhance the metadata to meet your needs.
Metadata in repositories has to come from somewhere, so this is the real starting point.
Our goal in repositories has been to have self-submitted content and this is one of the identifying characteristics of repositories. In reality the “if you build it, they will come” philosophy hasn’t been successful. Most of your submittals may instead be mediated by library staff. If you want to encourage self-submittal, don’t set the bar too high. Allow people to submit with only basic information.
For example, arXiv, a very successful disciplinary repository, requires only authors, titles, abstract and report number.

Theses frequently are self-submitted, either directly to the repository or with metadata from ProQuest. Because submittal is required for students to graduate, you can require additional information and they will add it, although you can’t count on the accuracy of their data. The lesson is people will add extra data when it is in their self-interest to do so.
Create Efficiencies

Where can you get data so that you don’t need to rekey?

We all know it is important to create efficiencies where you can, so look for sources for your data. If existing data can be repurposed you won’t need to spend time re-entering it. You may be able to get data from an individual or department in a spreadsheet, database or from a website. These sources are other forms of user submitted metadata and may have inaccuracies, lack information you need and may even include more specific information than you expected.

If you get data from a commercial source, like Web of Science, there may be limitations, such as not being able to use the abstract.

MARC records are another common source of metadata. Using MARC records and local data sources was covered in Ramirez & Fallgren’s webinar, so I won’t say much more about it.

Each of these sources will have its own issues. You will need to determine how to map it to your fields, what to do with extra information and how much missing information you should take the time to include.

In some cases, the items themselves may have embedded metadata which might be adequate for some information, so there is no reason to pull this data to populate fields, let alone rekeying it.
Automated Metadata Sharing

Goal for new items is to create metadata once and then share to all appropriate places.

Ideally new items would have metadata created once and then shared. Faculty members want to input metadata at most once and have it go to their CV, an administrative database, the institutional repository and a disciplinary repository, such as arXiv or PubMed. Automated deposit is really what should be happening everywhere as much as possible.

Even if a fully automated process is not currently possible, you also may be able to pull in data in a semi-automated way using an identifier such as a PubMed ID or by receiving a regular export of items.

I think of automated submittal as parallel to shelf ready. You need to decide how much tweaking you need to do and what can you accept. Spend your time on fields that really matter to your users and partners and only fix the rest in an automated way if at all.
Automated deposit into repositories using SWORD is a reality, just not widely adopted, at least in the United States. I think some of these initiatives are moving faster in places with mandates.

Even if your repository doesn’t support SWORD or if your researchers don’t publish in publications that offer SWORD deposit, look at how the process works to ensure your metadata fields will be structured to receive publisher data.

An interesting project in the UK is the RePosit Project, funded by JISC, the Joint Information Systems Committee. It is trying to increase use of repositories by embedding deposit in a researcher publications management system, or CRIS (Current Research Information Systems). They are working towards interoperability with Romeo data, repositories and research information systems. The Repositories Support Project recently had a workshop on RoMEO and CRIS in practice and the presentations are available online. I am very excited to see what these groups come up with since I think this is our future.

Efforts to automate deposit are already very successful in some places. I learned in a message on the JISC repositories list that the University of Auckland receives 95% of their deposits either via customized SWORD interfaces, or from their CRIS (Symplectic). I am really impressed by this.
Perfect metadata for few objects doesn’t make sense. Repositories need content and we all have limited resources, so you need to find a happy balance between a pile of unfindable stuff and beautiful metadata. If you are responsible both for recruiting content and cleaning up metadata, getting content will undoubtedly take the lead, because without content a repository is nothing.

Each sub-collection in your repository may have its own needs, but you should also consider the needs for your collection as a whole, as well as for those with whom you are sharing. If you have a specialized metadata scheme for a collection, you need to be certain it will still work well within your whole collection. You should allow for the growth of your repository so be careful about being too rigid in your requirements.

For metadata created any way other than manual entry by library staff, you will need to determine how extensive a review it needs. You may simply accept metadata as it is, with no review. You may decide to review only a couple of fields you determine are critical, and ignore the rest. You may decide to forgo controlled vocabulary for full text items. You may accept items with no keywords or abstracts. However, as soon as you share your metadata with others, some of these quality control issues become more noticeable.

As more deposits are automated, our expertise can be focused on QC that impacts interoperability and adds value, such as controlled subjects.
There are several ways you can enforce quality control on your fields and you should make use of them whenever possible. Place an automated check on the format or value of fields when possible. Having form fields for dates is useful so people can enter them one way and you can store them in ISO format.

Fields that you determine are essential, such as title, should be set as required. Default values are also helpful. For example, you may decide that language is important for interoperability, but since most of your content is in English, you can set that as a default. When content in other languages is added, this field can be easily edited.

Choosing options from a list is generally faster and easier for submittal and will result in consistent data. Remember, consistency is critical for mapping and interoperability.

Other types of corrections may be made in an automated fashion. For example, the metadata for some of our theses and dissertations have titles in all capital letters. We already transform the XML file, so we included the case change as part of the transformation. The changes are not perfect but were deemed acceptable for our repository and we would have ignored the problem completely if we couldn’t have made changes without additional work.
The most critical fields are those that relate to the user tasks of finding, collocating, identifying, evaluating and obtaining content.

Unless you manage a large disciplinary repository (such as arXiv) it is unlikely most visitors will find your content by directly searching the repository. This makes sharing your data with other collections critical.
Integrating your repository content with your other library materials is obvious, but sometimes repositories are not fully incorporated into the rest of the library. There is quite a bit of information about crosswalking Dublin Core to MARC and there are tools like Terry Reese’s MarcEdit to help.
Like us, you may have integrated your content not in the MARC catalog but in a resource discovery system, using OAI-PMH. I think all repository software supports OAI-PMH to share structured metadata in Dublin Core and other formats. Mapping your data well to OAI is an essential part of interoperability since it is the way most repositories share their metadata. If your data will be shared with a regional or consortial collection of repositories, remember to map to meet their needs as well.

One particularly important collection of repository metadata is OAISTER. This used to be at the University of Michigan, but since 2009 it has been part of WorldCat. Content is no longer harvested simply by OAI. You now map your data using the WorldCat Digital Collection Gateway. OCLC has detailed information on contributing repository data.
You can look at how other repositories map their data in OAI by adding commands after the repository’s base URL. If you don’t know the base URL, you can search in the Registry of Open Access Repositories and look at the record details. It can be particularly helpful to search by repository software to see examples that you know are possible locally. I picked this example at random, in case anyone is wondering. Click on ListMetadataFormats to see what metadata formats the repository supports. You can see items in the whole collection by adding the ListRecords and metadataPrefix commands, like the last URL displayed without the set information. If you want to look at a specific collection, you can figure out what their collections are with ListSets. You can then combine the commands to see the records for a specific set in a specific format.
Remember that subsets of your collection may have special interoperability needs. The Networked Digital Library of Theses and Dissertations focuses on a specific type of resource. They provide a standard set of metadata elements used to describe an electronic thesis or dissertation.
It is mostly Qualified Dublin Core, but also includes the additional elements degree name, level, discipline and grantor.

Even if you don’t plan to contribute your theses and dissertations at this time, review the fields that are important for them and include the fields in your metadata if possible. Out of NDLTD’s 22 elements, only title, creator, subject, date, type and identifier are required, so while encouraging more specific information, they also want to encourage participation.
Other Institutional Sharing

Institutions often have other projects parallel to the repository and you should coordinate with them as early as possible to foster interoperability.

There are probably extremely important local initiatives you should try to work with on campus. Many institutions have or are developing a centralized place for faculty information for promotion and tenure needs. If the official place for citations is the repository you should have a much easier time convincing faculty it is in their interest to contribute. However, Administrations need other information, such as editorships, prizes, committees and teaching information. This means your repository either needs to accommodate all those needs, or more practically, the two systems need to interoperate. The model of two interoperating systems is what is happening at the University of Auckland and with the RePosit Project.

Departments or Colleges may want customized displays of data on their sites. If they can pull from the repository, they can repurpose citations to highlight their department, create an expertise index, or simply feature the most noteworthy items. In order to meet their needs, there may be new fields that you need to include, such as information on grant funding or a way to mark items to be featured. Conversely, if you will be getting the data from them, try to have your critical fields included in their database.

Depending on the importance of these partnerships, you may commit to additional quality control. Don’t trust that just because the data matters for tenure that there won’t be errors! If you have built a place to house data centrally for other services to be built on, you have a greater responsibility for quality data, but because of the corresponding increase in the repository’s importance to the institution, hopefully you will receive sufficient support to ensure the metadata meets local needs. This quality control may also take place in the RIS which feeds the repository and other services.
Too often librarians don’t think about the world at large and truly integrate with non-library products, projects and services. Not to dismiss the library searching tools, but the reality is that most of your use will come from search engines, and probably mostly from Google. Make sure your site is optimized for search engines. This includes making sure your site is fully browsable with static links that aren’t redirects. You should be able to get to all your content in less than 10 clicks from the home page. Making a sitemap can help if you don’t have an easily browsed structure.
Google Scholar

“If you're using repository or journal management software, such as Eprints, DSpace, Digital Commons or OJS, please configure it to export bibliographic data in HTML "<meta>" tags.”


Google Scholar has specific information about what it looks for when indexing. Their information explicitly states repositories should use html meta tags. The fields in their example are title, author, date, journal title, volume, issue, first page, and PDF URL, in essence the basic elements of a citation.

Repositories generally have a metadata page which could include meta tags. Google Scholar can also use the structure of the PDF to identify the title from the top of the first page and expects the authors below the title.
Collocating Content
One specific issue for repositories is the relationship between the post-print and the published item. Make the relationship clear; describe the content you actually have but also allow them to come together. This is somewhat similar to the issues of describing the actual photograph taken in 1920 or the digitized file made in 2011. A user typically wants the 1920 date not the 2011 date. So, do you include the volume, issue and first page of the published item in individual fields or do you just include it in a citation field? To meet Google Scholar’s needs the citation data will need be in individual fields.
Using standardized terminology so results can cluster is a long standing practice in libraries. Controlled subjects are of course traditional way to collocate resources. Since repositories typically have very current articles and papers, existing subject vocabularies may be inadequate, as well as taking staff time you can’t afford. Keywords are great, and if you can have a prompted list based on collection, it can help keep terms consistent and reduce errors.
For repository interoperability, look particularly at the lists of format types used by others. Some repositories have a clear list with notes, others allow you to browse by type, and others have types in their schema. Some of the sources that we looked at when developing our list of format types are Georgia Tech, the Scholar Works Application Profile, Metadata Advisory Committee for Australian Repositories, and Max Planck Society’s eDoc Server, as well as OpenURL structure.
Many authors do not have a Library of Congress name authority record because they publish articles, not books. Even if an author has a record, your repository may not be able to include dates or parenthetical information. Collecting all of an author’s content is also of concern outside libraries. For example, faculty has an interest in pulling these variants together to help with citation analysis. I think librarians should be involved with and shape those efforts and not try to build a separate library solution.

You may be able to get faculty to confirm that the content you pulled in through batch ingest is correct. However, when content is aggregated with others, there will frequently be conflicts in names, which probably will require an institutional affiliation or author ID to disambiguate.
There are several library projects and studies trying to develop recommendations for authority control in repositories. The PubMed Author ID Project is one of the big projects that will have a major impact on a large portion of repository content. There are also commercial products, like ResearcherID, that are trying to unify authors, which demonstrate a need for this beyond libraries.

These issues may be solved through linked data, microformats, or some other solution. At this time, all I can suggest you do is be consistent locally, using the name the faculty member currently publishes under, and possibly even instituting a local author identifier.
Identifying Content
Complete metadata in large part is the classic “on a horse” issue. Be sure there is sufficient information in the metadata to provide context when it is merged with other records.
If you follow the Google Scholar recommendations, you will have citation information broken out into individual fields for use in meta tags. However, a plain text, human readable citation is important because your metadata needs to serve not only machines but humans. Ideally this could be created by the repository automatically, based on the individual fields. The Dublin Core Metadata Initiative Citation Working Group has recommendations about encoding citation information.
Anytime you have full content available, people may bypass your metadata page and go directly to the item. In order to reduce clicks, search engines try to link directly to the full content and not to your metadata page. People also may save an item to their desktop and so content should have sufficient information to be cited properly when it is out of the repository. For a PDF, this probably means having a title page with complete metadata, including citation, URL, rights and version. Embedded metadata may help with other items.
Selection is often not included in interoperability discussions, but the more your data is shared elsewhere the more you should be concerned that all user needs are being met.
It is very important to include rights information. You need this information both to support getting content and to support use of the content. Rights information is particularly important for items whose copyright is with the publisher. But it is also important to identify Creative Commons work, in part to encourage its application, and public domain content so that readers understand what they can and cannot do.
When posting content that also appears in a journal, you should make it clear what version it is, that is, before or after peer review, or in the final publication format. There are two guidelines that give similar but somewhat different language to make this clear. Following their recommendations is particularly good when your repository metadata has been merged with other data because different versions of the article may be aggregated together.
CrossRef Guidelines

Author Posting of author’s version of the work:
"Author Posting. (c) {Publisher}, YYYY. This is the author’s version of the work. It is posted here by permission of {Publisher} for personal use, not for redistribution. The definitive version was published in PUBLICATION, {VOL#, ISS#, (DATE)} doi:{nnnnnn} http://dx.doi.org/{nnnnnn}."

Author posting of the final definitive version (e.g. the PDF) of the work:
"Author Posting (c) {Publisher}, YYYY. This article is posted here by permission of {Publisher} for personal use, not for redistribution. The article was published in PUBLICATION, {VOL#, ISS#, (DATE)}, doi:{nnnnnn} http://dx.doi.org/{nnnnnn}.

http://crossref.org/08downloads/author_guidelines.pdf

CrossRef Guidelines for Standard Citations in Author Postings recommends that “the author version of a document should contain a full citation to the formally published work and a persistent DOI link to enable users to locate the definitive version of the work. CrossRef is also working to have repositories retrieve DOIs automatically in cases where the DOI isn’t provided by the author.”

Their guidelines combine citation and rights. At the University of Iowa, we have chosen to split the content into two fields so that we can map it more effectively.
Journal Article Versions uses somewhat different terminology than CrossRef and different than Sherpa/Romeo. While it is important to make it clear what version you have, at this point there doesn’t seem to be a complete consensus on terminology.

Identify content as peer reviewed.

It is important to identify content as peer reviewed not only in the context of your repository but also when the data is shared. This is an important feature for faculty and can alleviate some of their fears regarding submitting content. If other services are built onto a repository for or by a department, they will want this information. Ideally peer reviewed will both display and be a search limiter.
Identify content as peer reviewed.
Look at your repository from the perspective of a user. You want people to use your content so it should be as easy as possible for people to access it. Ranganathan’s 4th law of library science, “save the time of the user”, is always good to keep in mind.
Integrate with Other Services

“3. Embed library content within researchers’ regular workflows, integrating with services others provide.”


Your repository content should be in all the places people interact with scholarly materials. OCLC Research produced a document called *Support for the Research Process: An Academic Library Manifesto*. It includes 10 call to action points for libraries to play a central role in support of scholarly research and publishing. All points are good to keep in mind, but point 3 about embedding content within regular workflows and integrating with services is particularly relevant for repository interoperability.

The sixth point is also important: “Recognize that discovery of content will happen outside of libraries—but that libraries are uniquely suited to providing the organization and metadata that make content discoverable.”
Repositories often have a recommended style for citation, which is helpful, especially when people are citing in non-academic sources, such as Wikipedia. However, your repository probably has content across various disciplines, all with their own citation style. The best option is to make sure your repository communicates effectively with citation management software.
RIS is a standardized tag format which allows citation programs to exchange data. Information about it is in several places including on the Reference Manager site. BibTex is another standard format for citations.

As long as your metadata can be mapped clearly into a standard citation format, it ought to work. Nevertheless, test various formats to ensure they go correctly into different software such as RefWorks, EndNote, and Zotero. You may also be able to support pulling data into your repository from these sources. Even if you can’t support exporting your data to citation management software yet, designing your metadata fields with this interoperability in mind will allow added functionality in the future.
Your repository of course supports linking directly to content, and may have handles to keep the link stable. However, the sort of automatic linking we have in link resolvers would be best. This is particularly important for content not hosted elsewhere, especially an entire journal or working paper series. I am unsure if OpenURL linking is possible for miscellaneous working papers and post prints. I don’t know how a link resolver knowledgebase could rationally account for all these separate items, but perhaps there will be a way in the future. Even so, you should make sure your repository elements can be mapped to OpenURL elements, which are largely the same as citation fields.
Finally, allow your data to be reused in ways you haven’t anticipated. It is possible there will be mash-ups coming from unexpected directions. People are doing great things with open data and you want to encourage people to make use of your repository in all ways.
Bibliography


• Dublin Core Metadata Initiative Citation Working Group. “Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata”, n.d. http://dublincore.org/documents/dc-citation-guidelines/

  http://www.oclc.org/gateway/support/best_practices.pdf


• Salo, Dorothea. “Name Authority Control in Institutional Repositories.” Cataloging & Classification Quarterly 47, no. 3 (2009): 249.


Thank you!
Questions?

wendy-robertson@uiowa.edu
http://ir.uiowa.edu/lib_pubs/76/