MIND THE GAP

Mind the Gap: Curating Digital Commons Metadata for SHARE

Poster Handout

http://ir.lib.uwo.ca/wlpres/56/

Lisa Palmer, University of Massachusetts Medical School, lisa.palmer@umassmed.edu
Joanne Paterson, Western University jpater22@uwo.ca
Wendy Robertson, University of Iowa, wendy-robertson@uiowa.edu
Emily Stenberg, Washington University in St. Louis, emily.stenberg@wustl.edu

Summary

The goal of the SHARE initiative, a partnership between the Association of Research Libraries (ARL) and the Center for Open Science (COS), is to build a free, open, data set about research and scholarly activities across their life cycle. SHARE provider institutions use a wide variety of repository softwares. As part of the 2016-17 SHARE Curation Associates program, several repository managers who use the bepress Digital Commons platform are collaborating on a gap analysis of the metadata provided by their institutions and harvested by SHARE. Our goals are threefold: to improve institutional metadata curation processes; to provide good and consistent metadata to SHARE; and to develop workflows and recommendations for other Digital Commons institutions to apply.

OAI-PMH Metadata Formats Available in Digital Commons

oai_dc Default prefix. Fixed mappings to select simple Dublin Core elements.
simple-dublin-core Simple Dublin Core, flexible mappings. Alternate format: dcs.
qualified-dublin-core Qualified Dublin Core, flexible mappings. Alternate formats: dcq, qdc.
oai_etdms Generally used by Library and Archives of Canada (LAC) and for sharing records with Networked Digital Library of Theses and Dissertations (NDLTD).

General Recommendations

1. Follow DataCite guidelines for mapping institutional repository metadata to SHARE
2. Until a DataCite format is available, metadata from Digital Commons repositories should be harvested to SHARE using the qualified-dublin-core (qdc, dcq) format rather than the default oai_dc format
3. Map Qualified Dublin Core to DataCite terminology

Recommendations for Digital Commons Repository Managers

1. Consult bepress documentation on metadata options and OAI-PMH
2. Review how records for different collections are exposed in the various bepress OAI-PMH formats
3. Create standard metadata using consistent internal field names for types of series and share your practices publicly
    a. Develop ideal format for each collection type on your demo site.
    b. Document ideal series metadata mapping and make publicly available:
        1. Link to an external site from your repository such as Google Sheets or GitHub
        2. Add a “data dictionary” to your repository at the collection level
        3. Share with Digital Commons user group or Resource Library
    c. Work with bepress consultant to modify and migrate existing collections using this documentation.

Specific Field Recommendations

DOI

Problem: DOI fields are not mapped in bepress oai_dc format and are not mapped in qdc unless specifically requested. Because this unique identifier is unavailable, SHARE is unable to detect possible duplicate records.
Recommendation: Include any and all identifier fields in oai_dc and qdc formats, including DOI (example: <dc:identifier.doi>10.13028/M2301F</dc:identifier.doi>)

Publisher

Problem: Repositories often do not require a publisher field and in many cases the publisher is a different entity. In Digital Commons

oai_dc, publisher defaults to name of the repository. It is also not clear how to include both an institution name and a repository name (or if this is desirable) in the metadata.
Recommendation: Continue discussion with SHARE and the IR community in general to come up with best practices.

Type

Problem: Type is complex because it is used both for a DCMI Type Vocabulary and for something more akin to genre. The situation is more complex for Bepress customers because oai_dc uses ”text” as the default for everything. Digital Commons has a required document_type field that could be mapped to dc:type. However, this same field is used in journals for sections in the table of contents. These variant uses mean the facets cannot be limited to a controlled set of terms.
Recommendation:
Bepress should use a different field for journal display purposes. Bepress mapping should include both Document Type as bepress uses it AND the DCMI type (e.g. text). Repository managers should work closely with SHARE as they continue to develop their vocabulary. Repository managers should look at terms from CASRAI, COAR, etc. to develop more consistent local usage.

Author/Creator

Problem: The “flat” author OAI-PMH metadata from Digital Commons does not expose affiliations, identifiers, or role. This data would be invaluable in helping SHARE to disambiguate author names.
Recommendations: Re-structure author data in Digital Commons like DataCite’s nested structure to accommodate the inclusion of author identifiers such as ORCIDs. Expose author identifier and affiliation for each author in OAI. Expose author first and last name fields as subproperties (<givenName> and <familyName> per DataCite 4.0) in OAI. Ask Bepress to consider incorporating a dropdown menu on the input form to select “role” for each creator, e.g. author, editor, translator.

Next Steps

1. Develop recommendations for specific structures such as journals
2. Contact SHARE with report and requests
3. Contact bepress with report and requests

Resources

- DataCite 4.0 documentation: http://schema.datacite.org/meta/kernel-4.0/
- Metadata Options in Digital Commons: https://www.bepress.com/reference_guide_dc/metadata-options-digital-commons/
- Digital Commons and OAI-PMH: Harvesting Repository Records: https://www.bepress.com/reference_guide_dc/digital-commons-oai-harvesting/
- SHARE search interface: https://share.osf.io/
- SHARE metadata providers: https://share.osf.io/sources
- SHARE Data Dictionary 2017 (work in progress): https://docs.google.com/document/d/1OSgsTBNaar8DLHoVvE_Ge0H_5XQKU1OZ8cnA7MMe3lE/
- SHARE Data Provider Metadata Recommendations Guide (work in progress): https://docs.google.com/document/d/1nFPg49nQfepAvnpA5o279lM3FYkC0DIjdFGujYMsMrw/
- Share_Datacite_Bepress mapping spreadsheet (work in progress): https://docs.google.com/spreadsheets/d/1xPovfi0ateFdMZq6nkduph5jITHU3YJ2VMHLz9Bk_FI/edit?usp=sharing
- Python script to generate a .csv spreadsheet file of institutional metadata from the SHARE API: https://gist.github.com/leb2dg/f061a3af3a390b0a95e0a62490690fe0

Example Record

oai_dc: https://drive.google.com/open?id=1OAvv5UXwm3AI0qBd8pQODj5311pUxCPmYJSfT8snED0
qdc: https://drive.google.com/open?id=1eLQt20-0dvDEIp0miMhJojYK88pVFbI9oHjSqP5BmIw
DataCite: https://drive.google.com/open?id=1cuO-UGO0N05nSnIqC9z7inJ2HPW4s9PbSdCST5QIKCc