7-1-2017

Best Practices for Mapping Digital Commons Metadata for Harvesting by SHARE

Lisa Palmer
University of Massachusetts Medical School

Joanne Paterson
Western University

Wendy C. Robertson
University of Iowa

Please see article for additional authors.

Comments

Version 1.1.
The report is also available for comments in Google documents

Best Practices for Mapping Digital Commons Metadata for Harvesting by SHARE

Version 1.1[1]
July 2017

Prepared by:

Lisa Palmer, Institutional Repository Librarian, Lamar Soutter Library, University of Massachusetts Medical School, lisa.palmer@umassmed.edu

Joanne Paterson, Head, Metadata Access, Library and Information Resources Management, Western Libraries, Western University, jpater22@uwo.ca

Wendy C Robertson, Institutional Repository Librarian, University of Iowa Libraries, University of Iowa, wendy-robertson@uiowa.edu

Emily Stenberg, Digital Publishing and Preservation Librarian, University Libraries, Washington University in St. Louis, emily.stenberg@wustl.edu

Introduction

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."[2] As of June 26, 2017, 154 repositories and publishers have made metadata available to SHARE for harvesting, and the aggregated data set is available for searching.[3] Many metadata providers are institutional repositories utilizing the bepress Digital Commons platform[4] whose metadata is harvested through the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) protocol[5] for repository interoperability.

It is easy to register your Digital Commons repository and become discoverable in SHARE. Complete the registration form at https://share.osf.io/registration. When you are prompted for your baseURL, provide

[1] Version 1.1 has had feedback from bepress staff incorporated into the document.
[2] http://www.share-research.org/
[3] https://share.osf.io/sources
[4] https://www.bepress.com/products/digital-commons/
[5] https://www.openarchives.org/pmh/

your repository domain followed by do/oai, e.g., http://ir.uiowa.edu/do/oai. Digital Commons repositories are already participating to help increase the exposure of their content. As part of the 2016-17 SHARE Curation Associates program, the authors, who are managers of Digital Commons repositories, began 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.

This document details our findings and recommendations to improve the mapping of Digital Commons metadata to SHARE. It should be emphasized: these are neither requirements of SHARE nor barriers to participation. Our purpose is to gather community feedback from other institutional repositories and to provide bepress with clear recommendations on how their OAI could be enhanced.

Background

OAI-PMH Metadata in Digital Commons

Digital Commons metadata is mapped to Dublin Core elements and is exposed for harvesting through four different OAI-PMH formats:

oai_dc	Default prefix. Mostly 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 Canada (LAC) and for sharing records with Networked Digital Library of Theses and Dissertations (NDLTD).

Detailed information on the mapping and possible customizations is available.[6]

The metadata from each of our four repositories was exposed to SHARE in the default oai_dc format. We discovered some specific problem areas related to the default format which we will discuss in detail below.

SHARE continues to improve their harvesting. This document was almost complete when SHARE posted these new recommendations:[7]

[6] https://www.bepress.com/reference_guide_dc/digital-commons-oai-harvesting/

[7] http://share-research.readthedocs.io/en/latest/harvesters_and_transformers.html?highlight=date#best-practices-for-oai-sources

- Every OAI source supports oai_dc, but they usually also support at least one other format that has richer, more structured data, like oai_datacite or mods.
- Choose the format that seems to have the most useful data for SHARE, especially if a transformer for that format already exists.
- Choose oai_dc only as a last resort.

SHARE Metadata

SHARE has made available their current schema[8], data dictionary[9], and more recently, recommendations for data providers[10]. The recommendations use DataCite 3.X as the guideline for mapping Dublin Core to SHARE. Due to the nature of the data that are collected by SHARE, the schema model is subject to change.

In June 2017 SHARE began a review of their technical architecture and API, during which time they are postposing the development of most harvesters for new metadata providers.[11]

DataCite

We chose to use the DataCite Metadata Schema 4.0[12] as the most current specification. However, SHARE is currently using the 3.X guidelines.[13] In the text below, we note when features are part of the 4.0 schema.

Methods

We began by mapping Digital Commons default Dublin Core mapping for various kinds of collection structures. We then added other vocabularies to the mapping. Some fields require specific names for specific functionality and so tend to be more consistent across institutions, but for other fields bepress allows us great flexibility, even across our own repository.

We then refocused our efforts on Digital Commons to SHARE mapping. The authors looked at our own repository data and how it was mapping to Dublin Core so that we could better understand the gaps and

[8] https://share.osf.io/api/v2/schema

[9] https://docs.google.com/document/d/1OSgsTBNaar8DLHoVvE_Ge0H_5XQKU1OZ8cnA7MMe3lE/edit#heading=h.9gdts46x4zkm

[10] https://docs.google.com/document/d/1nFPg49nQfepAvnpA5o279lM3FYkC0DIjdFGujYMsMrw/edit

[11] http://www.share-research.org/2017/06/share-update-june-2017/

[12] https://schema.datacite.org/meta/kernel-4.0/

[13] https://docs.google.com/document/d/1nFPg49nQfepAvnpA5o279lM3FYkC0DIjdFGujYMsMrw/edit#heading=h.vcpebi1o032y

problems in the context of our own data. We created a spreadsheet where we mapped each SHARE element to DataCite, Dublin Core (both simple and qualified), and Digital Commons.[14] As part of this effort, we looked at documentation for other repositories (not only Digital Commons repositories) to consider the wider repository environment and try to make sure our thoughts went beyond our specific repositories. Throughout this, we asked SHARE programmers some questions to make sure we understood what would work for them. Finally, we pulled together our specific recommendations.

General Recommendations for bepress and Digital Commons administrators

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 Specific Fields

Creator

Element[15]	**dc:creator**
dcmi-terms[16]	creator
Digital Commons	Author field. Author information is structured in subfields (first name, middle name, last name, suffix) that are combined for the creator element. Each author is in a separate field, in inverted form. Not a mandatory field.
SHARE	Creators.creator Author disambiguation aided by an identifier (ORCID, Researcher ID, Scopus ID) and affiliation (name, Ringold Number, Name authority file number). Be certain these identifiers are clearly connected to their entities.

[14] https://docs.google.com/spreadsheets/d/1xPovfi0ateFdMZq6nkduph5jITHU3YJ2VMHLz9Bk_FI/edit#gid=193831059

[15] For this and all other items from the Dublin Core Element Set, see http://dublincore.org/documents/dces/

[16] For this and all other dcmi-terms sections, see http://dublincore.org/documents/dcmi-terms/ for definitions

DataCite	Mandatory. Includes optional properties for given name, family name, identifiers, and affiliation.
Problem(s)	The "flat" author OAI-PMH metadata does not expose affiliations, identifiers, or role. This data would be invaluable in helping SHARE to disambiguate author names.
Recommendation(s)	1. Re-structure author data in Digital Commons like DataCite's nested structure to accommodate the inclusion of author identifiers such as ORCIDs. 2. Expose author identifier(s) and affiliation for each author in OAI. 3. Expose author first and last name fields as subproperties (<givenName> and <familyName> per DataCite 4.0) in OAI. 4. Incorporate a dropdown menu on the input form to select "role" for each creator, e.g. author, editor, translator.
Example	This is an example of how this element looks in DataCite and ideally how it might look in Digital Commons OAI. <creator> <creatorName>Robertson, Wendy C.</creatorName> <givenName>Wendy C.</givenName> <familyName>Robertson</familyName> <affiliation>University of Iowa</affiliation> <nameIdentifier schemeURI="http://orcid.org/" nameIdentifierScheme="ORCID">0000-0002-3368-5080</nameIdentifier> </creator>

Contributor

Element	**dc:contributor**
dcmi-terms	contributor
Digital Commons	Free text field. Generally implemented as separate fields for each contributor (e.g. for thesis/dissertation advisors) or as a single field with contributors separated by a semicolon or other delimiter. A field labeled "advisor" would likely map to the Dublin Core element <contributor>. Information stored in the field would appear with a <dc:contributor> tag when exposed for harvesting.
SHARE	creators.creator Contributor type not in SHARE

DataCite	Contributor (with type, name identifier, and affiliation subproperties) is recommended.
Problem(s)	Not used in a standard or consistent way in Digital Commons repositories. As a free text field, it is not possible to include type, identifier, and affiliation information in a structured way.
Recommendation(s)	1. Re-structure contributor field in Digital Commons so that it is not a free text field but is instead structured like author information. See recommendations for dc:creator element. 2. Incorporate a dropdown menu to select "role" of each contributor, e.g. advisor, editor, sponsor, etc. (See DataCite's list in Appendix 1 of the 4.0 schema.)[17] 3. Ask SHARE to add "advisor" to their mappings for contributor roles as that is not in DataCite, yet is commonly used in repositories that publish theses and dissertations. 4. In the short term, when possible, contributor field should include a qualifier if mapping is obvious, e.g. dc:contributor.editor, dc.contributor.advisor 5. Employ consistent best practices for data entry. Per DataCite, contributor personal name should be entered in inverted format (FamilyName, GivenName), e.g. Patel, Emily 6. Repository name could map to dc.contributor (with qualifier dc:contributor.distributor) especially when working with data.
Example	This is an example of how this element looks in DataCite and ideally how it might look in Digital Commons OAI. <contributors> <contributor contributorType="HostingInstitution"> <contributorName> IFM-GEOMAR Leibniz-Institute of Marine Sciences, Kiel University </contributorName> </contributor> <contributor contributorType="ProjectLeader"> <contributorName>Starr, Joan</contributorName> <nameIdentifier nameIdentifierScheme="ORCID" schemeURI="http://orcid.org/">0000-0002-7285-027X</nameIdentifier> <affiliation>California Digital Library</affiliation> </contributor>

[17] https://schema.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf

	<contributor contributorType="Distributor"> <contributorName>eScholarship@UMMS</contributorName> ></contributor> <contributor contributorType="Editor"> <contributorName>Federal Institute for Population Research</contributorName> <nameIdentifier schemeURI="http://isni.org/isni/" nameIdentifierScheme="ISNI">0000000094455866</nameIdentifier> </contributor> </contributors>

Coverage

Element	**dc:coverage**
dcmi-terms	coverage spatial temporal
Digital Commons	Digital Commons maps the place of publication of a book to coverage. When geographic coordinates have been added to an item, Digital Commons uses dc.coverage.spatial.lat and dc.coverage.spatial.long.
SHARE	not in SHARE
DataCite	The date type “collected” is used for the time period covered by a data set. GeoLocation (with point, box and polygon sub-properties) is recommended. Geographic coordinates can include both a name (geoLocationPlace) and coordinates (pointLongitude and pointLatitude)
Problem(s)	This field should be used for things about a place or time period. Dublin Core scope notes for coverage state: “The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.”
Recommendation(s)	1. Publisher field should map to dc:publisher and not to dc:coverage on book structures. 2. When a time period is relevant (as for a dataset), repository managers should request that field be mapped to dc:coverage.temporal and in DataCite to dateType=”collected”

	3. Continue to map geographic coordinates to dc:coverage.spatial.lat and dc:coverage.spatial.long.
Example	<dc:coverage.temporal>2016-08-01T07:00:00Z</dc:coverage.temporal> <dc:coverage.spatial.lat>33.7801493</dc:coverage.spatial.lat> <dc:coverage.spatial.long>-115.90649940000003</dc:coverage.spatial.long>

Date

Element	**dc:date**
dcmi-terms	available created date dateAccepted dateCopyrighted dateSubmitted issued modified valid Guidance on using these properties is available in the DCMI User Guide.[18]
Digital Commons	The Publication Date appears in Dublin Core in ISO 8601 format (YYYY-MM-DD) with time, e.g. simple dublin core <dc:date>2012-08-21T21:00:00Z</dc:date> Date is NOT a required element. Events often do not include a date. Qualified Dublin Core may lack a date by default in some structures (e.g. events). If there is a date, the default is: <dc:date.created>2008-10-08T07:00:00Z</dc:date.created> There are not separate fields for date submitted / accepted / issued / modified / valid. Date available is used for embargo dates. However, if a series has an embargo option but the item is not under an embargo, it may get a default date of today or 1970. Date issued is used for dataset series in qualified Dublin Core. e.g. <dc:date.issued>2016-10-29T07:00:00Z</dc:date.issued> The <record> header contains the date and time the metadata was last updated, which may correspond to the date of upload into Digital Commons Repository. <header>

[18] http://wiki.dublincore.org/index.php/User_Guide/Creating_Metadata#Dates

	<identifier>oai:ir.lib.uwo.ca:biophysicspub-1029</identifier> <datestamp>2015-01-28T20:55:40Z</datestamp> <setSpec>publication:biophysics</setSpec> <setSpec>publication:biophysicspub</setSpec> </header> The upload date appears on the download button and in the metadata page source (meta name="bepress_citation_online_date") but does not appear currently in the OAI.
SHARE	dates.date date_published date_updated free_to_read_date (see rights section) date_created date_modified Note: Date_created and date_modified are listed at the top of https://share.osf.io/api/v2/schema/Publication but are not included in the data dictionary or data provider guide. We are unclear how they are being utilized.
DataCite	PublicationYear is mandatory and not repeatable. This is the year when the data was or will be made publicly available. If the DOI is for a digitized item, supply the year for the digitized version and not the original item. If there is no standard value, use the date that would be preferred from a citation perspective. Date (with type sub-property) is recommended. ● Accepted [The date that the publisher accepted the resource into their system. To indicate the start of an embargo period, use Submitted or Accepted, as appropriate.] ● Available [The date the resource is made publicly available. May be a range. To indicate the end of an embargo period, use Available.] ● Copyrighted ● Collected [The date or date range in which the resource content was collected. Typically used for datasets.] ● Created [The date the resource itself was put together; this could be a date range or a single date for a final component, e.g., the finalized file with all of the data. Recommended for discovery.] ● Issued [The date that the resource is published or distributed e.g. to a data center] ● Submitted [The date the creator submits the resource to the publisher. This could be different from Accepted if the publisher then applies a selection process. Recommended for discovery.] ● Updated ● Valid
Problem(s)	1. Digital Commons has one Publication Date field and it is currently not

	possible to distinguish between different sorts of dates (created, available, etc.) or have multiple date fields (with the exception of Embargo Date). The SHARE, DCMI, and DataCite schemas all allow for multiple date fields with a choice of date types, to various extents. 2. Distinguishing between different sorts of dates can be very challenging to apply consistently. What is the difference between issued and accepted and created? Which should we use? If something was published in 1995, digitized in 2016, and not posted in the repository until 2017, what do we call the 1995 and 2017 dates? 3. See rights regarding dateCopyrighted. 4. Do we need a date submitted and date accepted, especially for theses, journals we publish, and preprints? We don't have a mechanism to show the date for when scholarship was made public vs. when it passed through peer review. 5. If the item in the repository has been updated (not the metadata) there is not a place for this date to indicate the date updated/modified. 6. NISO's recommended practice Access License and Indicators (NISO RP-22-2015) is beginning to be used and we should consider how to incorporate into our IRs (e.g. <free_to_read start_date="2015-02-03"/>).
Recommendation(s)	1. Continue to map embargo date to dc:date.available. 2. Map the posted date to date available when there is no embargo date so that date.available can reliably map to a free to read date. 3. For Qualified Dublin Core, the publication date should be mapped to dc.date.issued rather than dc.date.created. 4. Request that bepress accommodate multiple date fields in a record, perhaps with a dropdown menu to be able to indicate which type of date. 5. Consider adding a field for date modified if the item itself (not the metadata) has changed 6. See Coverage for dateCollected. See Rights for dateCopyrighted. See also Rights for embargo / free to read date.
Example	<dc:date.available>2017-06-20T07:00:00Z</dc:date.available> <dc:date.issued>2014-12-01T08:00:00Z</dc:date.issued> Thesis: <dc:date.accepted>2017-05-15T09:00:00Z</dc:date.accepted> <dc:date.available>2017-06-27T11:00:00Z</dc:date.available>

Description

Element	dc:description
dcmi-terms	abstract description tableOfContents
Digital Commons	Digital Commons maps comments, peer_reviewed, abstract, link to thumbnail image for books to description. Peer reviewed is mapped as a 0 or 1, e.g. <dc:description>1</dc:description> Qualified Dublin Core includes non-Dublin Core fields on thesis records which map to NDLTD (thesis.degree.name ; thesis.degree.level ; thesis.degree.discipline ; thesis.degree.grantor).[19] They are also mapped in the thesis specific etd-ms OAI format.
SHARE	description version
DataCite	Description (with type sub-property) is recommended. descriptionType is mandatory if description is used: Abstract, Methods, SeriesInformation, TableOfContents, TechnicalInfo (new in version 4.0), Other The Version element is optional and should probably map to dc:description. FundingReference (with name, identifier, and award related subproperties) is optional. New with version 4.0 (had been a type of contributor)
Problem(s)	In Digital Commons, description maps to four different fields by default with no qualifier. The image jpg and peer reviewed don't map well to DataCite. They would be better with a qualifier.
Recommendation(s)	1. Continue to map abstract to dc:description.abstract 2. Separate fields should be made as needed for the following and mapped as specified: ● Table of Contents - dc:description.tableOfContents ● Methods (for data) - dc:description.methods ● Technical Information - dc:description.technicalInfo ● Map the comments field to dc:description.other 3. Funder name and award number should map in a nested structure in a specialized datacite OAI format.
Example	<dc:description.abstract>This study replication data resource includes information referencing specific items in the Alzheimer's Disease

[19] http://www.ndltd.org/standards/metadata#thesis.degree-caja-guest-0

	Neuroimaging Initiative (ADNI) database.</dc:description.abstract> <dc:description.tableOfContents>Description of quartz-window calorimeter -- Operation of the calorimeter -- Calibration of the calorimeter -- Description of boiler and furnace</dc:description.tableOfContents> <dc:description.methods>Utilizing the ADNI database, we identified 41 individuals who remained stable for 48-months (NC) and 16 who converted to MCI (CNV). Of these 57 subjects, all had available baseline clinical and MRI data, but only 16 NC and 11 CNV had available FDG-PET data. </dc:description.methods> <dc:description.other>Unpublished fieldwork reports (Grey Literature Library)</dc:description.other> <dc:description.funding>This study was supported by an investigator-initiated research grant to Benjamin U. Nwosu from Novo Nordisk, Inc., grant number H-13938. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.</dc:description.funding> This is an example of how funding information displays in DataCite 4.0 and ideally how it might look in Digital Commons OAI. <fundingReferences> <fundingReference> <funderName>Novo Nordisk</funderName> <funderIdentifier funderIdentifierType="Crossref Funder ID">http://doi.org/10.13039/501100004191</funderIdentifier> <awardNumber> H-13938</awardNumber> </fundingReference> <fundingReference> <funderName>National Institute of Allergy and Infectious Diseases</funderName> <funderIdentifier funderIdentifierType="Crossref Funder ID">http://doi.org/10.13039/100000060</funderIdentifier> <awardNumber awardURI="https://projectreporter.nih.gov/project_info_description.cfm?aid=9109531&icde=0">R01AI106725</awardNumber> <awardTitle>REGULATION OF CD8+ T CELL IMMUNITY TO TUBERCULOSIS</awardTitle> </fundingReference> </fundingReferences>

Format

Element	dc:format
dcmi-terms	extent format medium
Digital Commons	[derived, not in record]
SHARE	formats.format size
DataCite	Format is optional. Use file extension or MIME type where possible, e.g., PDF, XML, MPG or application/pdf, text/xml, video/mpeg Size is optional and can map to dc.format.extent
Problem(s)	1. Digital Commons returns the MIME type of the item in simple OAI. This value is derived and is not in the record. It cannot be mapped in qdc unless the value is added to a different field manually. This value would map to DataCite Format. 2. If controlled terms are used for the format, there is no way to indicate the vocabulary (IMT[20] etc.) 3. DataCite also has the field Size ("Unstructured size information about the resource"), which can map to pagination in other schemas and to dc.format.extent. The DataCite element can also include the file size which ought to be able to be derived from the file, but is not. 4. In qualified Dublin Core, some institutions have opted to map first page and last page to format.extent, but with no context for what the number means: <dc:format.extent>297</dc:format.extent> <dc:format.extent>304</dc:format.extent>
Recommendation(s)	1. Provide the ability to generate the MIME type from the file type for qdc as well as for oai_dc. 2. Meanwhile, repository managers should consider adding the value themselves to each item manually and map to qdc. This has the added benefit of being in the quarterly backup so that non-pdf primary file types can be identified properly. 3. Map total number of pages, file size, or duration to dc:format.extent 4. First page and last page should not be mapped to dc:format.extent. If

[20] http://www.iana.org/assignments/media-types/media-types.xhtml

	mapping is desired, add something that makes it clear what the numbers are. 5. If the total pages are mapped to dc:format.extent, recommend using the word "pages" to make it clear what the value means, e.g. <dc:format.extent>12 pages</dc:format.extent>
Example	1. <dc:format.extent>256 pages</dc:format.extent> <dc:format>application/pdf</dc:format> 2. <dc:format.extent>3KB</dc:format.extent> <dc:format>application/xml</dc:format> 3. <dc:format.extent>28:24 minutes</dc:format.extent> <dc:format>video/mp4</dc:format>

Identifier

Element	**dc:identifier**
dcmi-terms	bibliographicCitation[21] identifier
Digital Commons	There can be two instances of the <dc:identifier> tag in an "oai_dc" record: ● URL of the article index page (*e.g.,* <dc:identifier>http://ir.lib.uwo.ca/oncpub/54</dc:identifier>). ● A direct link to the full-text version of the primary file. (*e.g.,*<dc:identifier>http://doi.org/10.1016/S0360-3016(00)00484-3 </dc:identifier>) Note that elements of a citation that are used in creating an open URL are in individual fields (volnum, issnum, fpage, lpage, issn) but typically not mapped to Dublin Core.[22] Year, article title, first author, journal title, and doi are also used in an OpenURL but these often are mapped to Dublin Core. If elements of a citation are mapped to qualified Dublin Core, it looks like this (note that the journal title is mapped to a different element): <dc:source>PloS one</dc:source> <dc:identifier>1932-6203 (Linking)</dc:identifier> <dc:identifier.bibliographicCitation>10</dc:identifier.bibliographicCitation>

[21] See http://dublincore.org/documents/dc-citation-guidelines/ and http://wiki.dublincore.org/index.php/User_Guide/Creating_Metadata#BibliographicCitation

[22] See also http://www.ukoln.ac.uk/repositories/digirep/index/Scholarly_Works_Application_Profile#Bibliographic_Citation

	<dc:identifier.bibliographicCitation>2</dc:identifier.bibliographicCitation> <dc:identifier.bibliographicCitation>e0115671</dc:identifier.bibliographicCitation>
SHARE	identifier. Type of identifier not in SHARE.
DataCite	Mandatory, non-repeatable field (with mandatory type sub-property). The only allowed value is a DOI. DataCite uses Digital Object Identifiers (DOIs) at the present time and is considering the use of additional identifier schemes in the future. <identifier identifierType="DOI"> http://doi.org/10.5061/dryad.09d0k</identifier> AlternateIdentifier (with type sub-property) is optional. <alternateIdentifiers> <alternateIdentifier alternateIdentifierType="PMID">26289232 </alternateIdentifier> </alternateIdentifiers>
Problem(s)	1. URLs of additional files in Digital Commons are not exposed through OAI. 2. DOI and other identifiers are not exposed in oai_dc. 3. PubMed ID "a special identifier," while it brings metadata in, is not retained, unless manually added to a separate field. 4. If elements of a citation are mapped, it is not clear if the content represents a volume number, issue number, etc.
Recommendation(s)	1. Dublin Core: Recommended best practice is to identify the resource by means of a string conforming to a formal identification system. 2. All item specific identifiers for the version in the repository should be mapped to dc:identifier. In qualified Dublin Core, these should be qualified to identify the specific identifier type. 3. Map PubMed ID to dc:identifier.pmid. 4. Do not map elements of bibliographic citations to dc:identifier.
Example	<dc:identifier>http://ir.uiowa.edu/wwqr/vol18/iss1/15</dc:identifier> <dc:identifier> http://ir.uiowa.edu/cgi/viewcontent.cgi?article=1644&context=wwqr </dc:identifier> <dc:identifier.doi>10.13008/2153-3695.1644</dc:identifier.doi> <dc:identifier.bibliographicCitation>Walt Whitman Quarterly Review 18(1-2) (2000).</dc:identifier.bibliographicCitation>[23] <dc:identifier.pmid>26289232</dc:identifier.pmid>

[23] Note that this example is for an article in a journal published by the repository.

Language

Element	dc:language
dcmi-terms	language
Digital Commons	Language is not a standard field in Digital Commons sites. It will map in oai_dc and qdc if the field is present.
SHARE	language, not currently exposed SHARE can use any standard terminology such as ISO 639-1, 639-2 or 639-3.
DataCite	Language is optional. DataCite uses IETF BCP 47, ISO 639-1 language codes.
Problem(s)	The lack of a language field reduces the usefulness of repository metadata in a global setting. Its lack makes non-English content harder to locate within a predominantly English language collection. Its lack is particularly noteworthy in countries with multiple official languages. Consistently including a language will make the data far more interoperable.
Recommendation(s)	1. ISO 639-2/B uses bibliographic terminology (i.e. fre vs fra) which largely corresponds to MARC[24] so its use will allow repository metadata to mesh smoothly with traditional materials in library discovery systems. 2. Bepress should include language as a standard field. Most series will have content in one language, so a default value can be used. 3. Series administrators should request this field be added to existing series. 4. If there are multiple languages, recommend separating the values with a semicolon (e.g. eng; lat; grc)
Example	<dc:language>eng</dc:language>

Publisher

Element	dc:publisher
dcmi-terms	publisher
Digital Commons	Name of repository by default. This default can be replaced with a specific publisher or multiple publishers at the repository's request. Users can also create a blank text field for a value to be entered, or use the value from a

[24] See "Relationship to ISO 639-2" in *MARC Code list for Languages* 2007 Edition https://www.loc.gov/marc/languages/introduction.pdf

	particular metadata field.
SHARE	publisher
DataCite	Mandatory field that is not repeatable. Defined as "The name of the entity that holds, archives, publishes prints, distributes, releases, issues, or produces the resource. This property will be used to formulate the citation, so consider the prominence of the role."[25]
Problem(s)	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. Some institutions publish original journals under their institution, not the repository. Republished materials, such as article reprints, will have a different original publisher. It is not clear how to include both an institution name and a repository name (or if this is desirable) in the metadata.
Recommendation(s)	1. Continue discussion with SHARE and the IR community in general to come up with best practices. 2. Bepress should include repository as a separate field. It should not map to publisher as the default, as publisher is not repeatable in DataCite. Options could include dc.relation.isPartOf, dc.contributor.distributor (if it is data), or dc.source. 3. Repository managers should always include a publisher field. Preprints and postprints should use the institution's default publisher (whether the name of the institution or repository). Publisher PDFs should have the name of the publisher. 4. When there is no publisher, there should be a default listed, either the institution or repository. If a DOI has been assigned, the publisher field should be consistent with the metadata submitted for the DOI minting.
Example	<publisher>Springer</publisher> <publisher>Western University</publisher> <publisher>eScholarship@UMMS</publisher>

Relation

Element	**dc:relation**

[25] See DataCite Metadata Scheme V 4.0, https://schema.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf (page 12).

dcmi-terms	conformsTo hasFormat hasPart hasVersion isFormatOf isPartOf isReferencedBy isReplacedBy isRequiredBy isVersionOf references relation replaces requires
Digital Commons	This field is not included in default record configurations in Digital Commons. It will map in oai_dc and qdc if the field is present. A single title field is required.
SHARE	relatedidentifiers.relatedidentifier
DataCite	RelatedIdentifier (with type and relation type sub-properties) is recommended. DataCite expects that relations will be made using identifiers.[26] ● IsCitedBy ● Cites ● IsSupplementTo ● IsSupplementedBy ● IsContinuedBy ● Continues ● HasMetadata ● IsMetadataFor ● IsNewVersionOf ● IsPreviousVersionOf ● IsPartOf ● HasPart ● IsReferencedBy ● References ● IsDocumentedBy ● Documents ● IsCompiledBy ● Compiles

[26] See https://schema.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf p. 37-40 for details.

	● IsVariantFormOf ● IsOriginalFormOf ● IsIdenticalTo ● IsReviewedBy ● Reviews ● IsDerivedFrom ● IsSourceOf
Problem(s)	1. Dublin Core and DataCite schemas handle this differently in that DataCite includes a "relatedIdentifierType" property that is not present in Dublin Core. 2. Because the field in not included in default Digital Commons configurations there is much variability where it has been implemented in repositories. Relationships are made most often with titles or descriptive information rather than solely with identifiers. Related identifiers often end up being mapped as dc:identifier. 3. Many libraries are now providing DOI minting services. In some cases they will need to make a distinction in a record between a DOI that they have minted and related DOIs, e.g. a DOI minted for a dataset and the DOI of the related published journal article.
Recommendation(s)	1. Continue discussion with bepress, SHARE and the IR community in general to come up with best practices.
Example	<dc:relation.isPartOf>Physiological Reports</dc:relation.isPartOf> <dc:relation.isFormatOf>10.1002/phy2.255</dc:relation.isFormatOf> <dc:relation.isPartOf>2051-817X</dc:relation.isPartOf> <dc:relation.isPartOf>Iowa Research Online, the University of Iowa</dc:relation.isPartOf> <dc:relation.isFormatOf>9780877455387</dc:relation.isFormatOf> Dataset linking to published paper: <dc.relation.IsReferencedBy>http://doi.org/10.1371/journal.pone.0137525</dc.relation.IsReferencedBy> Published paper referring to its dataset: <dc.relation.References>http://doi.org/10.1371/journal.pone.0137525</dc.relation.References>

Rights

Element	**dc:rights**

dcmi-terms	accessRights license rights
Digital Commons	The rights field is mapped to dc:rights. If a Creative Commons license has been selected, it will map to dc: rights in oai_dc and to dc:rights.license in qdc
SHARE	Rights, not currently exposed "free to read type" (URI to a rights statement for the work) and "free to read date" (the date when the work becomes free to read), neither of which are currently exposed.
DataCite	Rights is optional. Provide a rights management statement for the resource or reference a service providing such information. Include embargo information if applicable. rightsURI (which can link to creative commons) RightsHolder is a contributor type. Copyrighted is a date type.
Problem(s)	1. Rights is a free text field, which typically includes a human readable copyright statement with all the elements required for a complete copyright statement. Institutions may also have included the citation for the published version in this field. Some institutions/systems have been splitting the rights holder and the date copyrighted into specific fields for improved mapping. 2. URIs are growing in use for rights (e.g. RightsStatements.org and http://vocabularies.coar-repositories.org/documentation/access_rights/) and Digital Commons needs to accommodate such URIs. They are not a license on the work but a URL specifying the rights, so they should not be combined with the Creative Commons licenses. 3. Embargoed content is identified in the qdc dc:date.available field, but not in oai_dc. Content with other restrictions, such as limited to campus use or subscribers could be mapped if the institution has a field indicating such an access restriction. These mappings will not appear in oai_dc. NDLTD requires an indication if something is Not publicly accessible; Limited public access ; or Publicly accessible. These map well to the Eprints AccessRights Vocabulary Encoding Scheme.
Recommendation(s)	1. Continue to map rights to dc:rights and Creative Commons licenses to dc:rights.license.

	2. Bepress should add a field for rights URIs mapped to dc:rights.uri. The COAR rights[27] URIs should be used. 3. All restrictions (campus use, subscribers, etc.) should be made clear and mapped to dc:rights.accessRights. 4. Institutions should consider splitting the non-rights data out of their rights field for improved mapping overall. 5. Institutions should consider splitting their rights statements into date and rights holder for even better metadata mapping.
Example	<dc:rights>© 2015 Wendy C Robertson</dc:rights> <dc:rights.license>http://creativecommons.org/licenses/by/3.0/ </dc:rights.license> <dc:rights>Material in the public domain. No restrictions on use.</dc:rights> <dc:rights.license>https://creativecommons.org/publicdomain/mark/1.0/ </dc:rights.license> <dc:rights.accessRights>Access restricted until 02/23/2019 </dc:rights.accessRights> <dc:rights.accessRights>Access restricted to UI faculty, staff and students.</dc:rights.accessRights> <dc:rights.accessRights>Full text restricted to subscribers. </dc:rights.accessRights> <dc:rights.accessRights>http://purl.org/eprint/accessRights/OpenAccess </dc:rights.accessRights> <dc:rights.accessRights>http://purl.org/eprint/accessRights/RestrictedAccess </dc:rights.accessRights> <dc:rightsHolder>American Institute of Physics</dc:rightsHolder> <dc:date.dateCopyrighted>2011</dc:date.dateCopyrighted>

Source

Element	**dc:source**
dcmi-terms	source[28]
Digital Commons	This is the series that contains the item; it may be a repository invented series, conference, journal or monographic series. Exports name of series (or other Digital Commons publication type) by default. oai_dc typically maps custom_citation to dc:source in standard series and source_publication to dc:source in journals.

[27] Controlled Vocabulary for Access Rights (Draft V1) http://vocabularies.coar-repositories.org/documentation/access_rights/

[28] see also http://wiki.dublincore.org/index.php/User_Guide/Creating_Metadata#Source_and_Relations

	fpage may be mapped to dc:source as well.
SHARE	Used to identify the source of the data, e.g. CrossRef
DataCite	No corresponding value
Problem(s)	1. The purpose of this field seems fuzzy in general. The default mapping seems confused as well. 2. In qualified Dublin Core, some institutions have opted to map elements of the citation (journal title, volume, issue), but with no context for what the values mean: <dc:source>Neonatology</dc:source> <dc:source>98</dc:source> <dc:source>4</dc:source> 3. Golden Rules for Repository Managers recommend "The source or suggested citation of an item (e.g. journal's name, volume and issue of an journal article) is provided in <dc:source>. We are uncertain if this is what we should do or if relation would be a better place.
Recommendation(s)	1. Continue discussion with bepress, SHARE and the IR community in general to come up with best practices. 2. The repository series should not map here by default. 3. The fpage should not map to dc:source.[29] 4. Repository managers should review their mappings because the default mapping to dc:source has changed over time and IR managers may not wish to retain this mapping. 5. Elements of the citation should not map here with insufficient context for what the values means. 6. Repository name could map to dc:source.
Example	<dc:source>Physiological Reports 2:3 (2014) pp. 1-10. https://doi.org/10.1002/phy2.255</dc:source> <dc:source>eScholarship@UMMS</dc:source>

Subject

Element	**dc:subject**
dcmi-terms	subject

[29] There is no place in Dublin Core for fpage. It is part of dc:identifier.bibliographicCitation but as a number by itself it is meaningless.

Digital Commons	Disciplines (bepress taxonomy), keywords (free text), and subject_area (controlled vocabulary terms) map to subject.
SHARE	Subjects uses the bepress controlled vocabulary. Tags are keywords.
DataCite	Subject (with scheme sub-property) is recommended. May include URI of term. Could include a language term.
Problem(s)	1. There is no way to indicate if a specific vocabulary is being used in the a keyword or subject_area field. 2. Bepress has a well developed subject list, but the terms do not have URIs so it is not built for linked data.
Recommendation(s)	1. Bepress should assign URIs to their schema so that it may be used more effectively in a linked data environment. 2. A URI for disciplines would make it possible to distinguish between "subjects" and "tags" in SHARE. Until URIs exist, we should consider mapping keywords to dc:subject.keyword. 3. Bepress should work with libraries using LCSH terms, FAST and MeSH so that these can also make use of URIs.
Example	<dc:subject.keyword>link resolver</dc:subject.keyword> <dc:subject.keyword>SFX</dc:subject.keyword> <dc:subject.keyword>OpenURL</dc:subject.keyword> <dc:subject.keyword>context-sensitive linking</dc:subject.keyword> <dc:subject>Library and Information Science</dc:subject> <dc:subject>Transcription, Genetic</dc:subject>

Title

Element	**dc:title**
dcmi-terms	alternative title
Digital Commons	Required and not repeatable.
SHARE	titles.title
DataCite	Mandatory, with optional type sub-properties. DateCite includes Title-Alternative Title, Title-Subtitle, Title-TranslatedTitle, Title-Other (Other is new with version 4.0)

	Language may be specified in the title field.
Problem(s)	1. If there are translated titles, there is no good way to include a language code to the title field in Digital Commons. 2. Subtitles are with the main title. There is not a completely standard means to split them if needed. This is a very minor issue. Guidelines for subtitles are variable, but generally have recommended a colon (possibly with no spaces, colon space, or space colon space).
Recommendation(s)	1. If a repository includes additional title fields, they should be routinely mapped as appropriate to: o dc:title.alternative o dc:title.subtitle o dc:title.translated o dc:title.other 2. IR managers should try to be internally consistent with subtitles, but at least set them off with a colon. The current Scholarly Works Application Profile recommendation is 'space-colon-space'.
Example	1. <dc:title>Этюд о ревности</dc:title> <dc:title.translated>Study in Jealousy</dc:title.translated> 2. <dc:title>The Finality of the Image : Lévi-Strauss, Pop Art, and the Snapshot</dc:title>

Type

Element	**dc:type**
dcmi-terms	type
Digital Commons	By default, simple oai_dc includes the value “text” mapped to dc:type. This can be configured to use document_type upon request. Digital Commons does not impose a standard list of document_types, so each series/structure and each institution can use their own list. The document_type field is used by journals for journal- or issue-specific display in the table of contents, conflating uses of a single field. Default terms: Article ; Book ; Book Review ; Conference Proceeding ; Dissertation ; Editorial ; Letter to the Editor ; Response or Comment ; News Article
SHARE	resourceType

	resourceTypeGeneral See the Share Data Dictionary 2017 for a detailed list. Note that their list includes preprint. Controlled vocabularies, especially with URIs, are preferred
DataCite	ResourceType (in version 4.0 the general type became mandatory, with an optional sub-property). The preferred format is a pairing of the resourceTypeGeneral value with a subproperty (CASRAI Outputs Type list is recommended). Examples: Dataset/Census Data; Text/Conference Abstract ResourceTypeGeneral values[30] are very similar to DCMITYPE. DataCite uses Audiovisual instead of MovingImage and Still Image and it has added Model, Workflow, and Other.
Problem(s)	1. dc:type is used for both a broad term from a limited list and more detailed terms. Both uses are important and both types should appear in records. 2. There are too many possible vocabularies for type and none that seem to fully meet repository needs. The IR community as a whole need to work on this. COAR's vocabulary is designed for IRs, but this is not the same as CASRAI, used by DataCite. The SHARE type list is shorter than either of these, but is being developed. See Type Mapping for details. 3. Bepress is using the document_type field for two very different purposes, which adds additional complexity.
Recommendation(s)	1. Bepress should not use the same field for controlled terminology and for the ordering and display of contents in journals. 2. There should be two dc:type fields, one for a short controlled list (either DCMITYPE or DataCite's list) and a more detailed term. 3. Repository managers should look at terms from SHARE, CASRAI, COAR, etc. to develop more consistent usage. Ideally this will be a controlled lists of types and not locally developed. 4. SHARE should map this standard list of types for all repositories. 5. A specific preprint type should be included. A postprint type could be considered as well. 6. In the future, URIs should be included with types.
Example	1. <dc:type>Text</dc:type> <dc:type>Article</dc:type> 2. <dc:type>Text</dc:type> <dc:type>Preprint</dc:type> 3. <dc:type>MovingImage</dc:type>

[30] See https://schema.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf p.32-33 for details

	<dc:type>Lecture</dc:type> 4. <dc:type>Text</dc:type> <dc:type>Dissertation</dc:type>

Additional Recommendations for Digital Commons Repository Managers

1. Consult bepress documentation on metadata options and OAI-PMH[31]
2. Review how records for different collections are exposed in the various bepress OAI-PMH formats. Are custom fields mapping as expected/desired?
3. Create standard metadata using consistent internal field names. Ways to do this include:
    a. Develop ideal format for each collection type on your demo site and use as a reference going forward
    b. Add a "data dictionary" to your repository at the collection level
4. Work with bepress consultant to modify and migrate existing collections using this documentation.
5. Share your practices publicly
    a. Link to your data dictionary from your repository to an external site such as Google Sheets or GitHub
    b. Share with Digital Commons user group or Resource Library

Next Steps

We plan to ask for input from the community, including Digital Commons customers, bepress, OSF SHARE programmers, SHARE Curation Associates, and more broadly among other interested repository managers.

[31] https://www.bepress.com/reference_guide_dc/digital-commons-oai-harvesting/