

Institutional Identifiers in Repositories: A Survey Report for the NISO I2 Workgroup

Overview

The NISO I2 Working Group surveyed repository managers and developers to determine the current practices and needs of the repository community regarding institutional identifiers. Results from the survey will inform a set of use cases that will be shared with the community, and that are expected to drive the development of a new standard for institutional identifiers.

Executive Summary

Clear trends

The survey showed that standardized institutional identifiers are seen as important and it was agreed there is a need for them in the repository community. The need for identifiers is underscored by the ways in which repository content is shared. A clear majority of repositories include identifiers for the repository itself and many include institutional identifiers. Those that include the latter generally also include identifiers for subordinate units within the identified institution. Most of these identifiers are not used in other usage contexts -- e.g., Inter-Library Loan, electronic resource management systems, etc. -- but there is some agreement that it would be important for a single identifier to be used for all organizational purposes. The majority of respondents would be willing to participate in a registry of institutional identifiers provided that participation is voluntary and cost-free.

Institutional identifiers already in usage are largely based upon the Uniform Resource Identifier (URI) standard, whether they take the form of Hypertext Transfer Protocol (HTTP) URIs, Uniform Resource Names (URNs), CNRI Handles, or OCLC PURLs. An overwhelming majority of respondents consider resolvability of institutional identifiers important.

Metadata elements

The core metadata associated with an institutional identifier should require the Institution Name element, the Parent Institution element, and the Uniform Resource Locator (URL) element. The Region element is largely considered unnecessary, and pluralities consider Address and State/Province unnecessary. Most repositories are

already collecting some or all of the core metadata elements considered required or preferred. There is little agreement on the necessity of the following core metadata elements: Related Institution; Variant Name; City; and Country.

Areas with little agreement

Institutional identifiers are assigned in various ways: some are handled manually; others via automated processes; and others via a combination of manual and automated processes. A third of respondents would prefer to reflect institutional hierarchy in the identifiers, with nearly as many preferring to have non-hierarchical identifiers. There were a range of answers to the question of which organization would be best-suited to manage a registry of institutional identifiers.

Audience and Distribution

The intended audience of the survey was repository managers and developers. In order to increase the diversity of respondents, the group decided to take two tacks.

First, the group nominated a number of repositories considered prominent, and this short list was augmented with repositories identified via OpenDOAR¹, a directory of open access repositories. The directory allowed the group to associate potential survey respondents with repositories, and to choose repositories that are diverse with regard to geography, type of repository, software platform, and industry. The group decided that one-hundred was a good number of potential respondents.

Second, acknowledging that any such list would be incomplete, the group identified a number of mailing lists that were likely to be followed by the repository community. These lists are enumerated in Appendix A.

The survey was distributed via the Survey Monkey website² on Thursday, June 18th, 2009 to the e-mail addresses of the one-hundred individually-chosen repository contacts and via the group to the identified mailing lists, as well as to group members' personal blogs³ and microblogs⁴. Survey Monkey generated one link for each of these purposes so that results from individually-chosen contacts and those from listservs and blogs could be kept distinct, which was useful for group members to gauge how successful each tack was. The survey remained open until Monday, July 6th, 2009, a period of seventeen days.

1. <http://www.opendoar.org/>

2. <http://www.surveymonkey.com/>

3. <http://lackoftalent.org/michael/blog/2009/06/20/i2-survey/>

<http://namesproject.wordpress.com/2009/06/19/institutional-identifiers-repositories/>

4. <http://twitter.com/mjgiarlo/status/2230486784>

It is likely that repositories from academic and research libraries may have been over-represented in the survey results. The group will make an effort to include repository communities from public libraries, archives, and other less well-represented sectors in future work.

Response Analysis

29 of the 100 identified repository contacts responded to the survey, with 21 of these completing the survey. 136 persons responded to the survey sent out to mailing lists and blogs, with 81 of these completing the survey. In total, the survey had 165 responses, 102 of which answered every question.

The group examined the survey results for anomalous patterns, such as to see if a certain question or questions drove away respondents. Response rate was analyzed on a per-question basis to see if those who started but did not complete the survey disproportionately answered questions near the beginning of the survey versus those near the middle or the end.

It was determined that the first three questions were answered by the most respondents (155, 152, and 140, respectively) but beyond the third question, there is no indication that a significant percentage of respondents started the survey and abandoned it due to any particular question. The non-technical questions in the middle of the survey were answered by approximately 111-119 respondents and the same kind of questions near the end were answered by 106 respondents.

From the analysis, the group concludes:

- Non-technical questions were answered more than technical questions;
- Boolean questions (mean: 121 responses) were answered more than multiple-choice questions (mean: 92 responses);
- Multiple-choice questions (mean: 92 responses) were answered more than free-text questions (mean: 43 responses);
- Free-text questions (mean: 43 responses) were answered less than both Boolean (121 responses) and multiple-choice questions (mean: 92 responses);
- Within each type of question -- Boolean, multiple-choice, and free-text -- the general trend is that there are insignificantly more responses to lower-numbered questions than higher-numbered ones.

Findings

[N.B. Percentages may not add up to 100% as responses such as "N/A" and "Don't know" were not tabulated. Identifiers were not defined, and respondents entered their types of identifiers in free text.]

Institutional identifier usage

- 58.1% of repositories include identifiers for themselves, 49.7% of which are public. 41.9% do not include identifiers for themselves.
- 46.1% of repositories include identifiers for their organizations, 35.6% of which are public. 62.9% do not include identifiers for themselves.
- 74.2% of repositories that include institution identifiers also include identifiers for institutional subdivisions. 26.9% are used only internally.

- 37.1% use Handles for institutional identifiers.
- 24.7% use URIs or URLs.
- 3.4% use MARC organization codes.
- 3.4% use PURLs.
- 2.2% use Fedora identifiers.
- 2.2% use OCLC identifiers.
- 2.2% use UUIDs.
- 1.1% use ARKs.
- 1.1% use DOIs.
- 1.1% use DSpace identifiers.
- 1.1% use URNs.
- 1.1% use ISIL identifiers.
- 1.1% use NUC identifiers.
- 1.1% use OAI-PMH identifiers.
- 1.1% use OAO identifiers.
- 1.1% use TinyURLs.

Assignment of institutional identifiers

- 37.5% use systems to assign institutional identifiers:
 - Handle.net
 - DSpace
 - DNS
 - OCLC
 - ISIL
 - ePrints
 - EDINA
 - California Digital Library
- 41.7% use manual processes to assign institutional identifiers:

- By the repository team
- By a single individual
- By an outside department
- 9.7% use a combination of manual processes and systems to assign institutional identifiers.

Need for standardization

- 87.7% consider it important to standardize institutional identifiers for repositories; 43.4% consider it very important.
- 83.8% see a need for a standardized repository identifier.
- 78.4% see a need for a standardized institutional identifier.
- An average of 23.9% see a need for a standardized identifier at the campus, division, department, and office levels.
- Other levels identified by respondents:
 - Workflow
 - Multi-institutional group
 - Partner/Individual
 - Collection
 - Project
 - Research units
 - Service/Software instance

Issues potentially solved by a standardized institutional identifier

- 31.9% have yet to encounter any issues they would consider potentially solvable by standardized institutional identifiers.
- 14.9% state a standardized institutional identifier would have helped track institutions across name changes, disambiguate similarly-named institutions, and tie collections to institutions.
- 10.6% state a standardized institutional identifier would have helped identify and enumerate organizational units, especially in multi-lingual environments.
- 8.5% state a standardized institutional identifier would have helped tie authors to institutions.
- Other issues:
 - Uniqueness
 - Interoperability
 - De-duplication
 - Persistence
 - Statistics
 - Indexing
 - Workflow

Identification of subdivisions

- 36.0% prefer to surface hierarchy, identifying institutional subdivisions via affixing an additional element or elements to a high-level institutional identifier.
- 29.8% prefer to have unique identifiers for institutional subdivisions.
- 12.3% prefer not to identify subdivisions.

Sharing of data

- 89.9% have exposed repository data to search and discovery facilities (e.g., Google Scholar).
- 73.1% have shared repository data with collaborative initiatives.

Metadata element ranking

- 61.0% consider Parent Institution required; 15.2% preferred; 13.3% unnecessary.
- 31.1% consider Related Institution preferred; 26.2% unnecessary; 24.3% required.
- 78.1% consider Institution Name required; 18.1% preferred; 1.0% unnecessary.
- 33.3% consider Variant Name unnecessary; 27.3% preferred; 22.2% required.
- 35.0% consider Address unnecessary; 23.0% required; 18.0% preferred.
- 30.7% consider City unnecessary; 25.7% preferred; 21.8% required.
- 38.1% consider State/Province unnecessary; 22.7% required; 18.6% preferred.
- 50.0% consider Region unnecessary; 14.9% required; 11.7% preferred.
- 32.4% consider Country preferred; 28.4% required; 24.5% unnecessary.
- 41.6% consider URL required; 41.6% preferred; 5.0% unnecessary.

- 68.9% already collect some or all of the metadata marked as required or preferred.
- 25.5% collect all of the metadata marked as required or preferred.
- 29.2% collect none of the metadata marked as required or preferred.

- In addition to the metadata elements enumerated in the survey, respondents suggested that the following be added:
 - Contact E-mail
 - Institution Type (Gov., Edu., Com., Org.)
 - Campus
 - Relationships to Other Institutions
 - Geographical Coordinates (Lat., Long.)
 - URI

Other identifiers in metadata

- 54.4% use HTTP Uniform Resource Identifiers in their metadata. 47.5% plan to implement them in the next 2-3 years.
- 53.3% use CNRI Handles. 40.0% plan to implement.
- 51.1% use Digital Object Identifiers. 45.0% plan to implement.
- 12.2% use MARC organization codes. 7.5% plan to implement.
- 10.0% use OCLC symbols. 5.0% plan to implement.
- 4.4% use OCLC WorldCat Registry identifiers. 17.5% plan to implement.
- 1.1% use ISIL (ISO 15511) identifiers. 2.5% plan to implement.
- 0.0% use ISDIAH. 5.0% plan to implement.
- 0.0% use ISNI (ISO 27729). 5.0% plan to implement.
- 0.0% use Standard Address Numbers. 0.0% plan to implement.
- 0.0% use Global Location Numbers. 5.0% plan to implement.
- 0.0% use Data Universal Numbering. 0.0% plan to implement.
- 0.0% use NISO EDItEUR ONIX Serials Online Holdings. 5.0% plan to implement.
- Other identifiers used in metadata:
 - ISSN
 - ISBN
 - URN
 - ISMN (International Standard Music Number)
 - ARK
 - person ID
 - DAI (Digital Author Identifier)
 - ISO codes (e.g., country codes)
 - PubMed ID
 - Geo-locations
 - Internal building/department IDs
- Other identifiers planned for implementation:
 - CERIF
 - ARK
 - VIAF-ID
 - Academic Institution Internal Structure Ontology

Identifiers and contexts

- 56.6% report that institutional identifiers used in the repository are not used for other library activities (e.g., electronic resource sharing, ILL, etc.)
- 22.6% report that these identifiers are used in other contexts.
- 60.3% consider it important to have a single identifier that serves all organizational purposes. 25.4% do not consider it important.

Legacy identifier standards

The following are percentages of respondents who believe the listed identifier standard to be an important one to consider in developing a new standard for institutional identifiers:

- 16.7% : MARC
- 16.7% : CNRI Handle
- 11.1% : OCLC
- 5.6% : GKD
- 5.6% : Names
- 5.6% : ARK
- 5.6% : ISSN
- 5.6% : DOI
- 5.6% : Library & Archives Canada
- 5.6% : OAI
- 5.6% : LCCN
- 5.6% : EAN
- 5.6% : UK Access Management Federation

Resolvability

- 92.4% consider it important that institutional identifiers be resolvable, half of whom consider it very important.
- 5.7% consider it not important.

Identifier registry participation

- 56.3% would like to participate in a registry of managed institutional identifiers (assuming it's voluntary and cost-free).
- 35.9% are somewhat likely to participate.
- 1.0% are unlikely to participate.
- Registries that are already being participated in:
 - ROAR
 - DOAR
 - Handle
 - OCLC
 - RIAN (Ireland)
 - LC/NAF
 - DSpace
 - SPASE
 - GKD-ID
 - MARC
 - ISIL

Identifier registry management

- 16.2% recommend OCLC as a stakeholder that is best-suited to manage a registry.
- 10.8% recommend the registry be distributed among national libraries.
- 8.1% recommend the U.S. Library of Congress.
- 8.1% recommend the registry be completely decentralized.
- 5.4% recommend the National Library of Australia.
- 5.4% recommend Universities Australia.
- 5.4% recommend large professional organizations, e.g., SPARC or AMA.
- 5.4% recommend "anyone but a vendor."
- 5.4% recommend state or national governments.
- 2.7% recommend any North American stakeholder for North American institutions.
- 2.7% recommend SURF (Netherlands).
- 2.7% recommend KNAW (Netherlands).
- 2.7% recommend CAIRSS (Australia).
- 2.7% recommend UNESCO.
- 2.7% recommend the British Library.
- 2.7% recommend JISC (UK).
- 2.7% recommend IFLA.
- 2.7% recommend NISO.
- 2.7% recommend MIMAS (UK) for UK-based institutions.
- 2.7% recommend the UK Access Management Federation.

Identifier registry usage

- 31.5% are not aware of any non-local registries that store their institutional identifiers.
- 11.1% in an OCLC registry.
- 9.3% in CNRI Handle registry.
- 3.7% in Google or Google Scholar.
- 3.7% in MARC organizational codes.
- 3.7% in OpenDOAR.
- 3.7% in ROAR.
- 1.9% in Australian Digital Theses Program.
- 1.9% in Employer Identification Number registry.
- 1.9% in Global Komunika Dewata (Indonesia).
- 1.9% in Higher Education Funding Council for England.
- 1.9% in Higher Education Statistics Agency (UK).
- 1.9% in ISIL.
- 1.9% in Library and Archives Canada.
- 1.9% in Library of Congress Name Authority File.
- 1.9% in National Center for Education Statistics (NCES) Integrated Post-secondary Education Data System (IPEDS-ID; US)

- 1.9% in National Center for Education Statistics (NCES) Integrated Post-secondary Education Data System Office of Post-Secondary Education (OPE ID; US)
- 1.9% in North American Industry Classification System (NAICS, replacing SIC)
- 1.9% in Open Archives Initiative (registered data providers).
- 1.9% in SPASE.
- 1.9% in state catalog.
- 1.9% in Texas Digital Library.

Appendix A: Repository-Related Listservs

- code4lib -- <http://dewey.library.nd.edu/mailling-lists/code4lib/>
- DC-IDENTIFIERS -- <http://dublincore.org/groups/identifiers/>
- digital-curation -- <http://groups.google.com/group/digital-curation>
- dspace-general -- <http://mailman.mit.edu/mailman/listinfo/dspace-general>
- fedora-commons-users -- <https://lists.sourceforge.net/lists/listinfo/fedora-commons-users>
- ir-net -- <http://mailman.anu.edu.au/mailman/listinfo/ir-net>
- JISC-REPOSITORIES -- <http://www.jiscmail.ac.uk/archives/jisc-repositories.html>
- metadataLibrarians -- <http://metadatalibrarians.monarchos.com/>
- PALINET-IR-L -- <http://larch.palinet.org/archives/palinet-ir-l.html>
- REPOMAN-L -- <http://www.lsoft.com/SCRIPTS/WL.EXE?SL1=REPOMAN-L&H=LISTSERV.INDIANA.EDU>
- SPARC-IR -- <https://arl.org/Lists/SPARC-IR/>

Appendix B: Survey Questions

1. Does your repository include an identifier for itself?
2. Does your repository include an identifier for the organization(s) it serves?
3. If your repository includes identifiers for organization(s), does it also include identifiers for divisions, departments, or other subordinate units of the parent organization it serves?
4. If the identifiers indicated in questions 1-3 are for internal use only, how do you identify your repository, organization, and departments to others for data discovery or sharing data?
5. If you use any of these identifiers (repository, organization or subordinate units), how and by whom are they assigned?
6. If you use any of these identifiers (repository, organization or subordinate units), what form(s) or standard(s) do you use (e.g., Handle, URI, etc.)?
7. At which organizational level(s) do you see a requirement for a standardized form of an identifier?
8. What is your preferred strategy for identifying subordinate units of your organization?
9. Have you exposed your data to search facilities (e.g., Google Scholar) for data discovery?
10. Have you shared data from your repository with collaborative initiatives?
11. What are the minimum metadata elements sufficient to identify your organization?
12. Do you already collect in your repository the metadata you marked as Required or Preferred in Question 11?
13. What identifiers do you currently use in your metadata other than item identifiers? For instance, identifiers for repositories, institutions, individuals, reference code values, etc.
14. Are institutional identifiers used in the repository also used for other library activities (e.g., electronic resource sharing, inter-library loan, etc.)?
15. Is it important to have a single identifier that serves all organizational purposes? If so, why?
16. What identifiers do you plan to implement in the next 2-3 years?
17. How important would it be for an institutional identifier to be resolvable, for instance, to return information about the institution?
18. What issues have you encountered with your repository that would have been helped by an institutional identifier?
19. Are there existing institutional identifier systems that you use which would be important to consider in developing a new standard? If so, please briefly describe them.
20. How important is a standardized institutional identifier?
21. Are identifiers for your institution stored in any non-local registries? If so, which one(s)?

22. Would you participate in a registry to manage globally unique institutional identifiers, assuming participation is voluntary and cost-free?
23. If you answered “Yes, we already do” to the previous question, please name the registries where you participate.
24. Can you recommend any stakeholders that you feel are best suited to manage an identifier registry (e.g., professional societies, library service providers, etc.)?
25. OPTIONAL: Please provide a brief description (types of content, number of items, etc.) of your repository.
26. OPTIONAL: Please provide your name, title and e-mail address, if you are willing to share it with us. Your contact information will not be distributed beyond the members of the NISO Institutional Identifier Institutional Repository Working Group. Whether or not you provide your contact information will not impact the use of your responses to the survey questions.