12-24-2018

Come Together: Interdepartmental Collaboration to Connect the IR and Library Catalog

Amanda Makula
University of San Diego, amakula@sandiego.edu

Come Together: Interdepartmental Collaboration to Connect the IR and Library Catalog

Description, Abstract, or Artist's Statement

While institutional repositories (IRs) often include a built-in search tool and/or are indexed by Web search engines, some patrons go directly to the online library catalog with their information need. Rather than hope that users will stumble upon the IR from the library website or assume that they will start their research with a Google search, librarians can enhance IR discoverability and usage by integrating its content into the library catalog. With strong teamwork, good communication, and a shared vision, this endeavor transforms the IR and library catalog from separate, siloed platforms into a more cohesive collections package. At the University of San Diego, librarians and administrators across three departments came together to share information and work in concert to explore the benefits of auto-harvesting IR content into the library catalog. Driven by a vision of enhancing discoverability and access, as well as promoting the IR and enriching the catalog, the team members worked cooperatively to identify specific IR collections appropriate for harvest, investigate technical logistics, consult outside vendors (including Innovative Interfaces, Inc./III and bepress), and experiment with implementation.

Disciplines

Library and Information Science

Notes

This is an Accepted Manuscript of an article published by Taylor & Francis in *Serials Librarian* on 24 Dec. 2018, available online: https://doi.org/10.1080/0361526X.2019.1550337

Come Together: Interdepartmental Collaboration to Connect the IR and Library Catalog

Amanda Y. Makula
Presenter

While institutional repositories (IRs) often include a built-in search tool and/or are indexed by Web search engines, some patrons go directly to the online library catalog with their information need. Rather than hope that users will stumble upon the IR from the library website or assume that they will start their research with a Google search, librarians can enhance IR discoverability and usage by integrating its content into the library catalog. With strong teamwork, good communication, and a shared vision, this endeavor transforms the IR and library catalog from separate, siloed platforms into a more cohesive collections package. At the University of San Diego, librarians and administrators across three departments came together to share information and work in concert to explore the benefits of auto-harvesting IR content into the library catalog. Driven by a vision of enhancing discoverability and access, as well as promoting the IR and enriching the catalog, the team members worked cooperatively to identify specific IR collections appropriate for harvest, investigate technical logistics, consult outside vendors (including Innovative Interfaces, Inc. / III and bepress), and experiment with implementation.

KEYWORDS collaboration, institutional repositories, library catalog, metadata harvesting, Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH)

Background

In 2009, the Joint Information Systems Committee (JISC) funded a large-scale investigation, the *Online Catalogue and Repository Interoperability Study (OCRIS)*, which

highlighted the need for institutional repositories (IR) and online library catalogs to communicate with one another.[1] Though there are many issues to consider to achieve full interaction between the two systems, one simple step that academic libraries can take toward this goal is to employ automated harvesting from the IR to the library catalog. With an enterprising, collaborative team approach, metadata harvests are easy to implement with even limited staffing and resources. This process results in significant benefits: greater searching productivity and satisfaction among users, more exposure for the IR, and a more complete, enriched library catalog.

In her 2015 article, "OAI-PMH Harvested Collections & User Engagement," DeeAnn Allison, Professor and Director for Computing Operations & Research Services at the University of Nebraska-Lincoln, makes a strong case for integrating content from the IR into the library's catalog / discovery tool:

> Scholars are often frustrated with isolated databases that require them to repeat their searches . . . . The discovery tool is an attempt to answer this criticism by integrating resources with traditional catalog entries to provide a single point for searching . . . [and] can include more than just articles with the addition of locally developed resources that are frequently hidden in repositories and disconnected from the catalog.[2]

The phrase *hidden in repositories* emphasizes the discoverability disadvantage of content that is siloed. Harvesting content helps break down barriers between platforms, allowing users to save time and energy: fewer searches, less time sorting through results. Allison's research bears out her assertion that users appreciate integration. Usage statistics obtained from Google Analytics revealed that when users find harvested content via the library's discovery system (i.e., content that exists in the discovery system as a result of a metadata harvest), they spend more time with it – and they are more likely to return to it – than if they had discovered that same

content in some other way, such as a Web search.[3] As Wesolek, Comfort, and Bodenheimer point out, this preference for the library catalog may be due in part to researchers' comfort with and trust in it:

> The library catalog enhances access by virtue of being one of any library's most authoritative and widely available resources. In addition, it is more familiar to many researchers than the institutional repository. The library catalog is also used by researchers worldwide, either directly or through WorldCat, and when the content it has cataloged is made openly available, such as through an institutional repository, those researchers may access it. Finally, harvesting and crosswalking institutional repository metadata . . . also extends the function of the catalog to include non-traditional library materials.[4]

The revered status of the library catalog in the research habits of scholars makes it an excellent vehicle to maximize exposure to and engagement with content living in the IR. In turn, the catalog is enriched with a greater variety of materials, making it a more valuable tool for locating different kinds of content.

There are other, less tangible but equally valuable, advantages to implementing IR-catalog metadata harvests. Particularly at small or medium-sized institutions, the project necessitates collaboration among stakeholders across different departments to provide a variety of skills, perspectives, and relationships with vendors. This kind of collaboration promotes communication and cooperation and helps cultivate strong working relationships and respect for one another's roles and expertise. Working together in this capacity can also help build a foundation for future cooperative projects.

Project members depend on the institutional culture and organizational structure but are likely to be representatives from areas such as Technical Services, Systems, Digital Initiatives, and Reference (see Figure 1). The IR manager, the catalog / discovery service administrator, and the vendors of the platforms are critical participants, as they know the intricacies of the systems and can readily identify (and troubleshoot) potential obstacles.

Figure 1. Potential team members for a harvesting project

Institutional Context

The University of San Diego is composed of eight academic divisions and serves a combined student population of approximately 8,900 undergraduates, graduate students, and law students. The institution identifies itself as "a Roman Catholic institution committed to advancing academic excellence, expanding liberal and professional knowledge, creating a diverse and inclusive community and preparing leaders who are dedicated to ethical conduct and

compassionate service."[5] Marks of distinction include its designation as a Changemaker campus, robust undergraduate study abroad participation, campus sustainability efforts, and recognition by Best Choice Schools as the country's "most beautiful campus."[6]

The University Library is organized into three departments: Collections, Access, and Discovery; Reference; and Archives, Special Collections, and Digital Initiatives. The latter includes the Digital Initiatives Librarian, who oversees the IR ("Digital USD"), which is run on bepress' Digital Commons platform. Outside the library in Information Technology Services (ITS), the Head of Library and Web Services and the Library System Administrator administer the university's online catalog and discovery layer (Encore and Encore Duet) and integrated library system (Sierra ILS), all of which are Innovative / III products shared by the University Library and the law library.

Prior to the adoption of the IR, the library used Content Pro, a digital asset management system (DAMS), and harvested content from it into Encore for greater discoverability. In 2014, the library launched the IR on the bepress platform and migrated materials from Content Pro into the repository. Use of Content Pro discontinued, and with it the harvest. During the interview process for a new Digital Initiatives Librarian in 2016, the search committee asked then-candidate Amanda Makula about the potential to populate the library catalog with content housed in the IR. Both parties were enthusiastic about exploring the possibility. Upon her hire, Makula called a meeting with Laura Turner, Head of Technical Services; Diane Maher, University Archivist; Michael O'Brien, Senior Director, Library & Web Services; and Bee Bornheimer, Library System Administrator. The group identified reasons to pursue automated harvesting of IR content into the library catalog (namely greater discoverability and awareness of

IR content), discussed collections most appropriate for a harvest, and decided to attempt a pilot harvest.

Implementation

Copley Library's Special Collections owns thousands of twentieth-century postcards, the majority of which have been digitized, curated, and ingested into the IR.[7] Makula and Maher suggested using these collections for the pilot harvest, as they would provide valuable images to scholars conducting historical and cultural research in the library catalog. Bornheimer contacted support representatives at Innovative and explained the harvest project; in response, they requested the information listed in Table 1.

Table 1. Questionnaire from Innovative

- What is the repository base URL?
- Is this an image repository?
- What is the name of the repository?
- What is the estimated number of records that will be harvested?
- How frequently is the repository updated? (daily/weekly/monthly)
- Please provide at least one staff email address to receive automatic notifications upon collection harvesting into Encore.

Meanwhile, Makula communicated with bepress, who supplied documentation and support for harvesting repository records via OAI-PMH.[8] Through correspondence with both vendors, the team learned that bepress supplies an avenue for harvest by way of a Uniform Resource Locator (URL) that transforms IR data into Extensible Markup Language (XML), which Innovative receives and populates into Encore. Using Encore, librarians can schedule harvest jobs to run automatically on a daily or weekly basis, or they can opt instead to run a manual harvest whenever there is new content to pull from the IR. On their end, bepress can map the fields in collections targeted for harvest to Dublin Core elements in order to expose desired

metadata. While most default fields in Digital Commons come already mapped to Dublin Core elements, custom fields created at the institution's request may need to be mapped by a bepress consultant.

When the initial harvest was complete, team members were pleased that catalog records pointed users back to the IR in two ways: via a hyperlinked *Digital File* button near the title and a hyperlinked text URL in the *Identifier* field. Other aspects of the harvest, however, were unsatisfactory. First, the large amount of metadata assigned to the postcards in the IR proved unnecessary and distracting in the corresponding library catalog records. Extraneous fields such as the latitude and longitude, object size, date digitized, and image file size cluttered the catalog record and obscured more important information such as the direct link to the item in the IR. Makula and Bornheimer contacted bepress and Innovative respectively and asked if it were possible to harvest only desired fields and exclude others. From bepress, they learned that they would have greater flexibility in customizing the output of XML data from the IR by utilizing the metadata prefixes *dcs* or *dcq* rather than *oai_dc* in the harvesting formula. Innovative, however, stated that the sole metadata format they could ingest into Encore was *oai_dc*. At this point, team members feared a dead end. But O'Brien and Bornheimer petitioned Innovative, agitating for a work-around. Innovative agreed to experiment with ingesting Simplified Dublin Core (*dcs*); to everyone's surprise and delight, it worked. Therefore, the team requested that bepress customize metadata output in *dcs*, which resulted in more control over the catalog record display.

Another issue was that the customized titles that a librarian had assigned to fields in the IR did not carry over to the catalog record. These names had been customized for the clarity of human users but the harvest machinery did not know what to do with them. Instead, the Dublin Core elements to which they had been mapped overrode the customized titles. For example, the
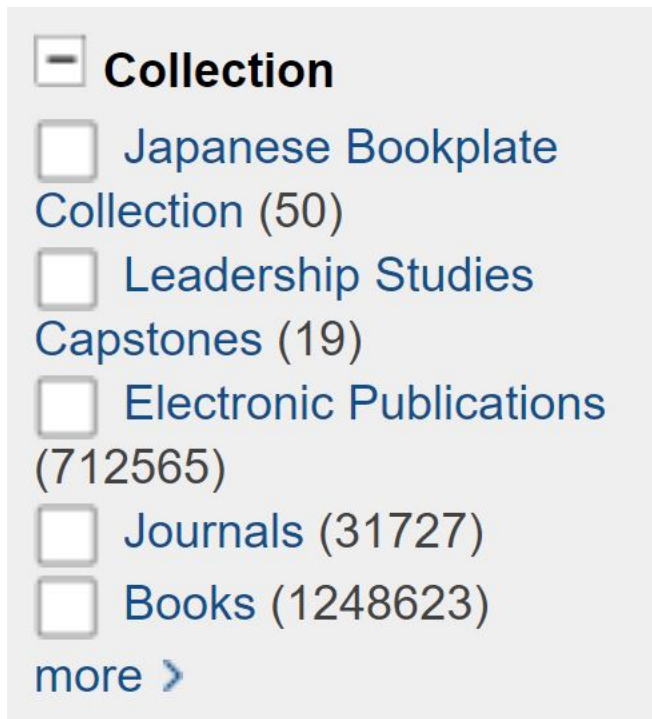
librarian had re-named the Dublin Core element *Coverage* to read instead as *Location* because the information in the field was a geographic location. The title of the field reverted to *Coverage* in the harvested catalog record. The harvest was also unable to separate the contents of multiple fields that had been mapped to the same Dublin Core element. Although subject headings and keywords resided in separate fields in the IR, because those fields had all been mapped to the same Dublin Core element (*Subject*), they appeared as a single string without delineation in the catalog record. Additionally, the harvest captured the time stamp (rendered as a nonsensical sequence such as *T07:00:00Z*) and attached it to the end of the *Date* field. bepress explained that this is because the *Publication Date* field exports the full *ISO 8601* format. Fixing this problem would be laborious – necessitating the addition of a text field with date information, mapped to Dublin Core's *Date* element – with the potential for introducing errors during the inputting process, so the team has not pursued it.

The issue of greatest concern to team members turned out to be the simplest to remedy. When the postcards were ingested into the IR, separate records were created for the fronts and backs. A button, titled *See Reverse Side*, linked the two records and enabled viewers to toggle between them, offering a complete view of the postcard. While this configuration functioned well in the IR, harvesting these records into the catalog eliminated this connection and thus the context. When looking at the catalog record for the back of a postcard, users often encountered handwriting or a blank card and had no easy way of matching it to its front side counterpart. Fortunately, bepress can add a field allowing repository managers to mark specific records within a single structure for exclusion from the harvest. However, due to the time-consuming nature of revising the approximately sixty postcard structures, the team decided to halt the postcard harvest and begin the process anew with other collections that needed no exclusions:

Japanese bookplates[9] and the capstone project papers of graduate students in the Master's degree program for Leadership Studies.[10]

When Innovative attempted to initiate these harvests, a mysterious error appeared, blocking its completion. Bornheimer, Makula, Innovative, and bepress worked together to determine the cause. The timing of the error coincided with the IR's transition from *HTTP* to *HTTPS*; when Bornheimer contacted Innovative, they confirmed that the current version of Encore did not support harvesting from *HTTPS* sites. Makula asked bepress to roll back the IR to *HTTP*; the error disappeared. The harvest proceeded, with a plan to move the IR to *HTTPS* as soon as Innovative releases a new version of Encore that supports *HTTPS* harvests.

Currently, as a result of automated harvesting, two collections – the bookplates and the capstone papers – populate the library catalog, each with its own facet under the *Collection* menu (see Figure 2). These records read in a more user-friendly format than did records from the postcard pilot harvest, thanks to the lessons of trial-and-error and close communication with the vendors. Despite the improvements, however, issues remain. Catalog records of Japanese bookplates omit a space between the *Description* and *Type* fields and catalog records of Leadership Studies capstone papers dump the document type, degree name, and department into a single thesis field. Presumably, this is because all these fields in the IR are mapped to facets of the Dublin Core value *thesis*, but it is worth investigating more fully. Catalog records for both collections also duplicate the hyperlinked name of the collection in the *Collection* field and, in the case of the capstones, lead to a dead end. Figure 3 shows an example of a catalog record of content harvested from the IR, with its connections to the IR highlighted.

Figure 2. Harvested collections appear in the facet menu of the library catalog.

Figure 3. Catalog record of a capstone paper harvested from the IR.

Reflection

Academic libraries have an opportunity to enhance discoverability of IR content, enrich the online library catalog, and cater to users' needs and preferences by implementing metadata harvesting. In particular, institutions without the staff or resources to manually catalog and embed IR content in the library catalog can benefit from an automated process.

Teamwork across relevant departments such as Technical Services, Digital Initiatives, and Systems, as well as with vendors, is vital in all stages of the process. Team members from each area contribute unique skills and acquire new knowledge from one another and from the project itself. For example, Makula, who came from a public services background and had no metadata training, acquired a basic understanding of Dublin Core schema over the course of the process. Collaboration can take many forms and is not prescriptive, but derived from the unique institutional context. At the University of San Diego, team members shared information via e-mail, chat, and face-to-face meetings. They logged questions and issues on a shared Google doc. They held calls with vendors and conveyed updates to other members of the group. Interactions were informal, nonhierarchical, and collegial. Relationships born from the project can serve as the groundwork for other collaborations in the future. In regard to vendors, a harvest project offers an opportunity not only to learn more about what the platforms can do, but also to articulate your library's needs and to push for improvements to support those needs. Working with Innovative throughout the harvest revealed just how novel a process it is for them. The University of San Diego seems to be one of its first clients pursuing OAI-PMH automated harvests, helping pave the way for other libraries that use the Encore platform.

Going forward, the team plans to investigate the remaining problems in the catalog records of harvested bookplates and capstone papers and then to target additional collections

(such as undergraduate honors theses) for harvest. A newly hired cataloging librarian will be invited to join the discussions. The IR will move to *HTTPS* as soon as Innovative is able to accommodate it. New content ingested into the IR will be curated with special attention as to what the metadata mappings will yield in the online catalog record.

ADDENDUM

Following the NASIG presentation, team members learned that Innovative had delayed their ability to harvest metadata from *HTTPS* sites. Rather than wait indefinitely for Encore to be able to handle *HTTPS* harvests, they decided to migrate the IR to *HTTPS* and put the harvesting project on hold.

NOTES

1. Kathleen Menzies, Duncan Birrell, and Gordon Dunsire, "An Investigation of Information Systems Interoperability in UK Universities: Findings and Recommendations," *New Review of Information Networking* 16, no. 2 (2011): 92-140, doi:10.1080/13614576.2011.619917.

2. DeeAnn Allison, "OAI-PMH Harvested Collections & User Engagement," *Journal of Web Librarianship* 10, no. 1 (2016): 1. doi:10.1080/19322909.2015.1128867.

3. Ibid., 11.

4. Andrew Wesolek, Jan Comfort, and Lisa Bodenheimer, "Collaborate to Innovate: Expanding Access to Faculty Patents through the Institutional Repository and the Library Catalog," *Collection Management* 40, no. 4 (2015): 226. doi:10.1080/01462679.2015.1093986.

5. "Mission, Vision and Values," University of San Diego, accessed May 18, 2018, http://www.sandiego.edu/about/mission-vision-values.php.

6. "Rankings and Recognition," University of San Diego, accessed May 18, 2018, http://www.sandiego.edu/about/rankings-and-recognition.php.

7. "Postcard Collections," Digital USD, University of San Diego, accessed May 12, 2018, http://digital.sandiego.edu/postcards/.

8. "Digital Commons and OAI-PMH: Harvesting Repository Records," bepress, accessed May 18, 2018, https://www.bepress.com/reference_guide_dc/digital-commons-oai-harvesting/.

9. "Japanese Bookplate Collection," Digital USD, University of San Diego, accessed May 12, 2018, http://digital.sandiego.edu/japanese_bookplates/.

10. "M.A. in Leadership Studies: Capstone Project Papers," Digital USD, University of San Diego, accessed May 12, 2018, http://digital.sandiego.edu/solesmalscap/.

CONTRIBUTOR NOTES

Amanda Y. Makula is Digital Initiatives Librarian, University of San Diego, San Diego, California.