Institutional Repository Digital Object Metadata Enhancement and Re-architecting

A. Saundry

University of British Columbia Library
Vancouver, British Columbia, Canada
amber.saundry@ubc.ca

ABSTRACT

We present work undertaken at our institutional repository to enhance metadata and re-organize digital objects according to new information architecture, in an effort to minimize administrative object management and processing, and improve object discovery and use. This work was partly motivated by the launch of a new discovery platform at our institution, which aggregates metadata and full text from our four open access repositories into a cohesive, consistent, and enhanced searching and browsing experience. The platform provides digital object identifier (DOI) assignment, metadata access via various formats, and an open metadata and full text application program interface (API) for researchers, amongst other features. Functionality of these platform features relies heavily on accurate object representation and metadata. This work facilitates and improves the discovery and engagement of the diverse digital objects available from our institution, so they can be used and analyzed in new, flexible, and innovative ways by a myriad of communities and disciplines.

Keywords

Institutional repository; information architecture; metadata; digital objects; digital libraries

1. INTRODUCTION

1.1 Background

The University of British Columbia (UBC) Library manages four open access repository platforms: DSpace (the institutional repository known as cIRcle¹), CONTENTdm, Access to Memory (AtoM), and Dataverse. Previously, this decentralization and diversity of vendor-provided interfaces resulted in uncoordinated delivery and an inconsistent searching and browsing experience, with a limited chance of object discovery and minimal ability for object use and analysis.

1.2 cIRcle/DSpace

cIRcle provides access to published and unpublished material created by the UBC community and its partners, with an aim to showcase and preserve UBC's unique intellectual output. It includes a diversity of items, including but not limited to documents, images, sound recordings, videos, and datasets. It currently holds approximately 57,000 objects or records; each object can hold multiple files and content types within.

cIRcle currently runs on DSpace 5.1, an open source platform based upon the Apache Cocoon framework (XMLUI), which "provides a modular, extendable, tiered interface" [1] and structures cIRcle around communities containing collections. Previously, these communities and collections were used to provide contextual and structural information to digital objects and aggregate objects based on various criteria. Communities were broadly organized according to the administrative structure of the university, and collections were created for various groups, subjects, projects, etc. Objects were also mapped to multiple collections when applicable, a largely individual and manual process.

This structure and practice resulted in the proliferation of many collections; it was administratively challenging to maintain and keep current with changes in the university administrative structure. Further, information about diverse objects was stored within the collection descriptions, not within the object metadata, reducing the flexibility of object use and retrieval. Once the objects were pulled out of the DSpace environment, objects were decontextualized and it proved challenging to pull together specific objects according to various criteria.

1.3 Open Collections

Launched October 2015, Open Collections² (OC) is UBC Library's discovery platform for all the Library's locally produced and managed digital objects. It aggregates objects from each of the four repositories described earlier to provide a consistent and enhanced searching and browsing experience, following best practices of information architecture [2]. At the core of OC is the open software framework that inserts metadata and full text (when available) from the repositories, and indexes it in a locally-controlled, dynamic index utilizing the open source application Elasticsearch.

OC provides features that increase the discovery, analysis, and promotion of UBC research, including faceted searching, embedded metadata in the HTML for improved search engine harvesting (including Google Scholar), responsive design, object usage statistics, and assignment of a digital object identifier (DOI) for each cIRcle object. Along with the ability to download object media and full text, metadata can be downloaded in various formats (including JSON-LD). UBC Library has also provided a metadata and full text application program interface (API) as an open source tool for researchers to mine data and build their own custom views, applications, and widgets.

These features rely heavily on the accuracy and reliability of object metadata. Thus, a review and evaluation of cIRcle's object metadata guidelines and community/collection information architecture was necessary. This work has involved object metadata enhancement, and community/collection consolidation.

2. METHODOLOGY

2.1 Metadata Enhancement

cIRcle object metadata fields were reviewed to provide various ways to search and aggregate objects based on numerous goals, needs, and criteria. cIRcle primarily uses the Dublin Core metadata schema [3]; while some fields new to cIRcle were added, some already existing fields required local standardization and refinement. Decisions were made so content would reflect the unique needs of UBC, stakeholders, and users of cIRcle [4].

With the refined metadata fields and guidelines in mind, identified communities and collections were evaluated to ensure that any contextual or specified criteria were present in the metadata of each enclosed object. Metadata was adjusted to new guidelines, moved, or populated accordingly, in order to pull together objects across cIRcle based on the various criteria that could be of interest to users.

The large majority of this work was devoted to the objects already held within cIRcle, while the new guidelines were applied to incoming objects. In this way, all objects in cIRcle adhere to consistent metadata standards.

2.2 Community/Collection Re-organization and Object Movement

The decision was made to shift away from the DSpace community structure built around the administrative structure of the university, to an information architecture based around the scholarly level or role of the content providers. Communities broadly include UBC faculty, graduate, undergraduate, or community and partners. Additional communities for UBC affiliated or hosted events and other special groups exist for administrative or historical reasons. Objects are placed in the highest scholarly level community; for example, an item authored by a UBC faculty member and graduate student is directed to the faculty community.

Once the metadata had been evaluated and adjusted accordingly, objects were moved to these new communities and included collections structured around the role of the content

provider, ensuring appropriate metadata allowed relevant items to be aggregated (when possible and appropriate).

3. RESULTS

This work is ongoing, but object metadata continues to be reviewed, and metadata fields standardized and enhanced. The large majority of the 57,000 objects have had their metadata adjusted according to new guidelines, with some work remaining for unique, sensitive, or complex cases.

Thus far, there has been a 63% reduction in the number of DSpace communities/collections, from approximately 475 to 175.

4. DISCUSSION AND FUTURE WORK

The metadata enhancements and changes to the information architecture of cIRcle improve object findability and usability, providing users with context and multiple ways to discover and access the objects [2,4]. The work has been intensive, and is ongoing. However, this creates more robust metadata with the potential for greater discoverability, use, and analysis. This work facilitates and improves the engagement of the diverse digital objects available from our institution, so they can be used and analyzed in new, flexible, and innovative ways by various communities and disciplines.

Documentation is being drafted to guide decision making in the assessment, direction, and management of new cIRcle objects. This documentation needs to be flexible enough to support the diversity and range of content that comes to cIRcle, and will need re-evaluation or modification at a later date as feedback or new information arises. Further work will include investigation of metadata challenges with the current lack of an authority system, including the balance of metadata consistency and accuracy [5]. cIRcle objects have seen an overall increase in usage since the launch of OC, and we hope to determine if usage continues to increase after improvement in the object metadata. Additional data and information should be reviewed to better understand how these changes are affecting object discovery and use over time. Further work includes evaluation of OC user feedback received since launch, with the goal to prioritize new features, abilities, and enhancements in the next development cycle (to be scheduled).

ACKNOWLEDGEMENTS

We are thankful for the ongoing collaboration and feedback between various team members within UBC Library (Digital Initiatives and Technical Services) and UBC Information Technology, and the expertise each member provides.

REFERENCES

- [1] DuraSpace wiki. 2015. XMLUI Configuration and Customization. *DSpace 5.x Documentation*.
- [2] J.C. Parandjuk. 2010. Using Information Architecture to Evaluate Digital Libraries. *The Reference Librarian*, 51, 124-134.
- [3] Dublin Core Metadata Initiative (DCMI). 2012. DCMI Metadata Terms.
- [4] H.L. Moulaison, F. Dykas and K. Gallant. 2015. OpenDOAR Repositories and Metadata Practices. *D-Lib Magazine*, 21(3/4).
- [5] J. Park and Y. Tosaka. 2010. Metadata Quality Control in Digital Repositories and Collections: Criteria, Semantics, and Mechanisms. *Cataloging & Classification Quarterly*, 48(8), 696-715.