Semi-automated collection evaluation for large-scale aggregations

Katrina Fenlon
kfenlon2@illinois.edu

Peter Organisciak
organis2@illinois.edu

Jacob Jett
jjett2@illinois.edu

Miles Efron
mefron@illinois.edu

Graduate School of Library and Information Science

University of Illinois, Urbana-Champaign

501 E. Daniel St.

Champaign, IL 61820

ABSTRACT

Library and museum digital collections are increasingly aggregated at various levels. Large-scale aggregations, often characterized by heterogeneous or messy metadata, pose unique and growing challenges to aggregation administrators – not only in facilitating end-user discovery and access, but in performing basic administrative and curatorial tasks in a scalable way, such as finding messy data and determining the overall topical landscape of the aggregation. This poster describes early findings on using statistical text analysis techniques to improve the scalability of an aggregation development workflow for a large-scale aggregation. These techniques hold great promise for automating historically labor-intensive evaluative aspects of aggregation development and form the basis for the development of an aggregator's dashboard. The aggregator's dashboard is planned as a statistical text-analysis-driven tool for supporting large-scale aggregation development and maintenance, through multifaceted, automatic visualization of an aggregation's metadata quality and topical coverage. The administrator's dashboard will support principled yet scalable aggregation development.

KEYWORDS

Digital collections, digital aggregations, collection evaluation, subject analysis, subject access, digital libraries, latent topic models, document representation.

INTRODUCTION

Library and museum digital collections are increasingly aggregated at various levels, from large-scale national and international digital libraries to local, institution-specific collections of collections. Diversity of the content in aggregations presents many challenges: not only user-centric concerns, such as facilitating discovery across massive data, but also administrative concerns, such as managing and integrating heterogeneous content. This poster describes work in progress on applying statistical text analysis to support an administrator's dashboard for building and managing digital aggregations.

CONTEXT

Since 2002, the Institute of Museum and Library Services (IMLS) Digital Collections and Content (DCC)[1] has grown into the largest cultural heritage aggregation in the U.S., with more than 1,200 collections and 1 million items from a diverse range of institutions in 44 states. DCC maintains a collection registry and item-level metadata repository, which together serve as a research test bed. This poster draws on the work of the Digital Collections Evaluation and the Subject Analysis working groups of the DCC project.

The IMLS DCC has developed a workflow for aggregation development, illustrated in Figure 1:

1. Identify collections for inclusion according to collection development policy.
2. Describe collection in a collection metadata record for ingest into the collection registry.
3. Harvest item-level data, usually via the Open Archives Initiative-Protocol for Metadata Harvesting (OAI-PMH), for the item repository.

1 http://imlsdcc.grainger.uiuc.edu/

ASIST 2011, October 9-13, 2011, New Orleans, LA, USA.

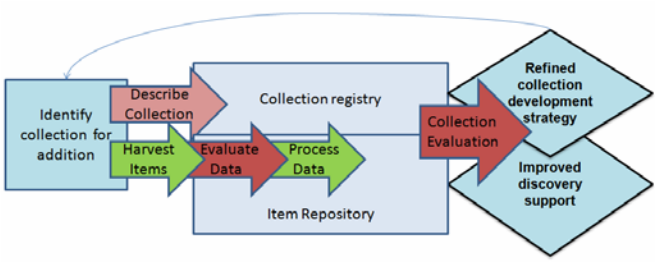

Figure 1. IMLS DCC aggregation development workflow

4. Conduct post-harvest evaluation for data idiosyncrasies that will affect processing. Process data to normalize and augment for search and display.
5. Perform an evaluation to determine the topical scope and breadth of the added collection and how it fits into the aggregation as a whole. Results are used to refine the collection development strategy and inform interface and functionality improvements, as described in Palmer et al. (2010).

The impetus for this research is DCC's growing need to improve scalability of the aggregation development workflow. While data harvesting and processing are largely automatic steps, the tasks of collection description, initial post-harvest data evaluation, and overall collection evaluation are labor-intensive processes that impede aggregation development. This has significant consequences for the visibility and accessibility of digital cultural heritage resources. This poster reports early findings on improving the scalability of data and collection evaluations. The techniques discussed here represent reapplications of earlier work that focused on supporting end-user behavior. Specifically, we explore ways to leverage emerging probabilistic subject analysis techniques to improve evaluation of large-scale aggregations, with the goals of (1) making metadata cleanup more efficient, and (2) improving collection development and access features.

METHODS

We are developing automated techniques for resource enrichment that can be adapted to support aggregation administration: a topic modeling approach derived from latent Dirichlet allocation (LDA, originally described by Blei et al. 2003), collection topic visualization, and machine-driven metadata augmentation using external content. We plan to apply these methods for improvement in the following aspects of aggregation administration:

1) Augmenting initial post-harvest data evaluation to discover weakly topical metadata records, by visualizing the information content of a collection's records at a glance. Metadata quality across large aggregations is notoriously heterogeneous (Dunsire 2008, Hillmann et al. 2004, Jackson 2006), creating difficulties in finding coherent topics across the collection. Individual record assessment is impractical across providers that serve up many thousands of item records. We approach this problem through two methods. First, we have been exploring metadata augmentation through a variety of external sources, for example matching record terms with controlled vocabularies or deriving topical information from general online resources. Second, we have developed techniques to improve the quality of LDA-based topic modeling, by

Method visualization	
Adaptation for use in administrative workflow	
1	
Discover and manually assess documents that appear to have high lexical overlap:	
Is the metadata suitable for detailed, item-level display?	
Will indexing all data from a collection impede full-text search?	
Can the data be better organized, e.g. into hierarchically organized records, to reduce redundancy and support discovery?	
2	
Discover and evaluate primary topical strengths and weaknesses of collections/aggregations:	
What are the contours of subject coverage in the aggregation; how are subjects distributed across collections and items?	
How should the collection development policy change to increase topical strengths and decrease weaknesses?	
Where are opportunities for development of thematic research collections on strong topics to support in-depth inquiry?	
Is collection-level subject description (manually assigned) accurate, in view of topical coverage of all items? Can it be augmented?	

Table 1. Adapting methods to support aggregation administration.

identifying and excluding unreliable metadata (Efron et al. 2011). Table 1.1 demonstrates how we can adapt this method to support post-harvest data evaluation, along with metadata normalization and enrichment.

(2) Visualizing subjects of collections and aggregation to improve overall collection evaluation. Previous collection evaluations were manual, relying on collection-level metadata that is often incompletely representative of all items' topicality. Our method transcends the limits of top-down subject category assignment by helping administrators discover primary topical strengths of a collection and of the aggregation as a whole, in a replicable way, relying on both item- and collection-level subject information. Table 1.2 shows an example of how an end-user browsing feature developed by this method can be adapted to support evaluation.

FINDINGS

These techniques are useful along both fronts: finding messy topical data and determining the overall topical landscape of the aggregation. In both cases, they provide useful and vastly more efficient evaluative elements if they can be systematized. Currently, they are used ad-hoc to assist in traditionally manual collection evaluation work. Moving forward, they will be integrated into an aggregation administrators' dashboard, a graphical interface to the results of a collection evaluation reliant on statistical text analysis.

CONCLUSION

Semi-automated assistive techniques for topical mapping of large collections are foundational to the administrator's dashboard, which we expect to serve as an administrator-centric portal view to the aggregation's metadata. This tool holds great promise for improving the modularity and efficiency of large aggregation administration, a burgeoning challenge for cultural heritage institutions. These methods will also inform future work on reconciliation and faceting of subject metadata across the aggregation for better control and improved discovery; on enrichment of metadata records using third-party / authority data; and for informing new end-user discovery features, including dynamic virtual collection generation in response to user queries.

ACKNOWLEDGMENTS

THIS WORK WAS SUPPORTED BY INSTITUTE OF MUSEUM AND LIBRARY SERVICES LG-06-07-0020, A PROJECT HOSTED BY THE CENTER FOR INFORMATICS RESEARCH IN SCIENCE AND SCHOLARSHIP (CIRSS). OTHER CONTRIBUTING PROJECT MEMBERS INCLUDE: CAROLE L. PALMER, OKSANA ZAVALINA, TIMOTHY W. COLE, THOMAS HABING, LARRY JACKSON, SARAH L. SHREEVES.

REFERENCES

Blei, D. M., & Jordan, M. I. (2003). Modeling Annotated Data. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. New York, NY, USA: ACM.

Dunsire, G. (2008). Collecting metadata from institutional repositories. *OCLC Systems & Services: International digital library perspectives*, 24(1), 51-58.

Efron, M., Organisciak, P., Efron, M. (2011). Building Topic Models in a Federated Digital Library Through Selective Document Exclusion. In *Proceedings of the ASIS&T Annual Meeting*. (New Orleans, LA, Oct. 9-13).

Hillmann, D., Dushay, N., Phipps, J. (2004), "Improving metadata quality: Augmentation and recombination." In *Proceedings of the 2004 international conference on Dublin Core and metadata applications: metadata across languages and cultures*, 11-14 October 2004, Shanghai, China.

Jackson, A. (2006). *Preliminary Analysis of Item-level Metadata Harvested* (White paper). University of Illinois at Urbana Champaign. Retrieved from . <http://www.ideals.illinois.edu/bitstream/handle/2142/720/Itemlevelmetadata.pdf?sequence=2>.

Palmer, C. L., Zavalina, O., Fenlon, K. (2010). Beyond size and search: Building contextual mass in digital aggregations for scholarly use. In *Proceedings of the ASIS&T Annual Meeting*. (Pittsburgh, PA, Oct. 22-27).