Faceted Search for Heterogeneous Digital Collections

Citation	Zhang, H., Durbin, M., Dunn, J., Cowan, W., & Wheeler, B. (2012). Faceted search for heterogeneous digital collections. Proceeding JCDL '12. Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, 425-426. doi:10.1145/2232817.2232924
DOI	10.1145/2232817.2232924
Publisher	Association for Computing Machinery (ACM)
Version	Accepted Manuscript
Terms of Use	http://cdss.library.oregonstate.edu/sa-termsofuse

Faceted Search for Heterogeneous Digital Collections

Hui Zhang
Digital Library Program
Indiana University
Bloomington, Indiana USA
hz3@indiana.edu

Jon Dunn, Will Cowan
Digital Library Program
Indiana University
Bloomington, Indiana USA
{jwd, wgcowan}@indiana.edu

Mike Durbin, Brian Wheeler
Digital Library Program
Indiana University
Bloomington, Indiana USA
{midurbin, bdwheeler}@indiana.edu

ABSTRACT
The idea of faceted search has received growing attentions in the digital library field for its potential of improving user satisfaction by combing the query and browse strategies interactively. Furthermore, with the trend of using digital repositories as the central infrastructure for curation and preservation, there is a demand for a single search interface providing public access to all the diversified content stored in the repositories. In this demo, we present *Digital Collections Search*, a system that is designed to assist users who are unfamiliar with the subject of their information needs locating relevant items as well exploring related but unknown collections in the repository.

Categories and Subject Descriptors
H.3.7 [**Information Search and Retrieval**]: Digital Libraries

General Terms
Management, Design, Human Factors.

Keywords
Faceted search, discovery interface, fedora common, solr, blacklight, disseminator.

1. INTRODUCTION
A common challenge to discovery interface design is that the average users of digital library systems cannot specify their information needs appropriately due to the lack of knowledge on the subject. The problem is only intensified when libraries keep adding digitized contents of various types and from different domains into their repositories. For instance, the repository maintained by Indiana University's Digital Library Program (DLP) holds more than 100 collections with a variety of content types such as images, manuscripts, and sheet music. Although some of the collections have their own public websites, it is impractical to exploit this rich resource without a unified discovery interface because otherwise a potential user will have to search his topics on every single collection individually.

In this paper, we demonstrate *Digital Collections Search* (DCS), a system developed by DLP that supports exploratory search across collections preserved in its Fedora Commons repository.

JCDL'12, June 10–14, 2012, Washington, DC, USA.

Although a faceted search interface to a union catalog is not a new idea, for instance the North Carolina State University Library's OPAC (Online Public Access Catalog), the DCS system is innovative in its supporting of contents generated from different metadata standards (e.g., EAD and MODS) and complicated workflows (e.g., flexible collection branding) by: creating an index framework serving as the middleware to synchronize the repository and discovery interface, developing Fedora services to "normalize" the digitized items and provide standardized output through web service.

2. SYSTEM OVERVIEW
2.1 Architecture
The general procedure of creating Solr index for the collections in Fedora will be briefly described in this section. Figure 1 illustrates the system architecture, which includes three layers: the back-end layer of Fedora repository, the middle layer of index framework, and the front-end layer of faceted discovery interface. The repository layer is responsible for providing MODS record for each digitized item, using a locally developed Fedora disseminator called getMODS, as input for the index layer. The middle layer of index service, called Fedora-index-service framework [1], is in charge of creating the Solr and Lucene indices for the MODS records. The framework is implemented as a RESTful web service application written specially to generate index for digitized items with complex structures (e.g., journals and books) while in sync with the Fedora repository. However, this framework does not support searching and browsing the index, therefore a discovery interface is developed in separate to allow public access to the indexed contents. The next two sections will discuss in details on how the Solr index is created and the design of the discovery interface.

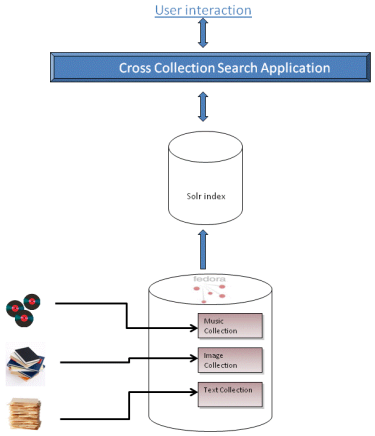

Figure 1: System Architecture

2.2 Building the Solr Index

Solr, an open source index server, is chosen as the framework for building the index because of its robustness and support for faceted search. Items in Fedora that will be included in the Solr index must associate themselves with the getMODS disseminator so the indexing routine only needs to handle one format no matter what metadata standards were used during the cataloging process. The generated MODS records will then be converted into a XML-style format that is native to Solr and indexed. All the input and output files for indexing are delivered as streams of raw bytes to save processing time and storage space.

Quality control and live update are two major challenges of incorporating the Solr index service with DLP's production system. Quality control, the procedure of keeping the indexed content correct and up-to-date, is involved across the whole workflow of indexing. For instance, an item may have to be removed from the index because it is unrelated with the collection. Therefore, the index service must implement routines to restrict problematic item from public access. Several measures are added to ensure this goal: a new field called *recordStatus* is created in MODS to register whether the item has completed the cataloging process; a Fedora service is created to confirm whether an item is eligible for public view; and the indexing program will check certain MODS fields to make sure they meet the expected pattern.

Live update is a customer raised feature that argues a digitized item should become available for the public view as soon as it has been cataloged. It requires that the Solr indexing speed must match the Fedora ingestion speed; it also demands a communication mechanism in between so that Fedora is able to notify the index service once an update to the repository is made. Through experiments, it is clear that the indexing process is slowed down mostly because of the Solr commit operation, and the solution is to specify the *maxTime* property of Solr's *autoCommit* configuration to the value (30000 ms) that will trigger the commit operation only when there were documents updated in the last five minutes. This change in Solr configuration is important to keep the repository and the index in synchronization especially during massive Fedora ingestion when thousands of digitized items are imported into the repository in batch mode.

The communication mechanism between Fedora and Solr indexer is built with the Java Message Service (JMS), where the Fedora server is specified as publisher and the index service is setup as the listener. Fedora will broadcast notice when an item is added or updated, and the indexing routine subscribed to the messages will in response start the process of (re) indexing that item.

2.3 The Discovery Interface

The discovery interface at the front-end is designed with the goal of assisting non-expert users in locating relevant items without the knowledge of the subject. Blacklight, an open source online public access catalog project, is chosen as the platform to develop the search interface because it is developed primarily for library usage and supports search with Solr index out-of-box.

Identifying the facet fields is a vital task for discovery interface design. All the facets used in DCS are data driven, which means their values are stored in Fedora and exposed using disseminator. After all, seven index fields are chosen as facets, where three of them, namely *Format*, *Collection*, and *Source*, are extracted from Fedora and passed directly to the indexer. In contrast, the other four facets are translated from the corresponding MODS fields based on the following mapping patterns:

- Topic facet: mods:subject|mods:topic
- Creator facet: mods:namepart
- Genre facet: mods:genre
- Year facet: mods:originalInfo[@keyDate and @w3cdtf]

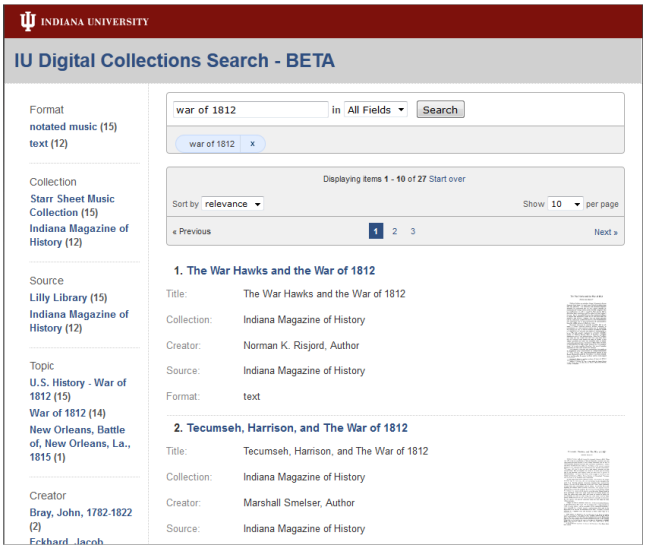

Figure 2. Snapshot of IU Digital Collections Search

A snapshot of the faceted search interface is depicted in Figure 2. In this example, it provides results to a keyword search of *war of 1812*, where the user can choose either narrowing down his request or exploring new items by picking the appropriate facets (e.g., picking a *topic* facet to specify requirement or a different *collection* to explore the repository) shown in the left pane. Selecting a different facet will initiate a new Solr search and the results will be rendered in the primary window on the right, and a click on the item title or its thumbnail will redirect the user to a webpage that will show the item in its full details.

3. FUTURE PLAN

DCS is still at Beta stage; we are adding new collections and collecting user feedbacks at this time. There is a plan for improvement and new features on top of the list are: supporting full-text search, recognizing and indexing new Fedora collections automatically. We also want to explore the possibility of integrating DCS with traditional library catalog and using it for video search.

4. REFERENCES

[1] Durbin, M. and Dunn, J. (2008) Accommodating Diverse Search Requirements over a Fedora Repository. In: Third international Conference on Open Repositories 2008. 1-4 April 2008, Southampton, United Kingdom.