Stanford University’s SearchWorks is a library discovery portal built with Blacklight, an open source tool. In this chapter from Nicole C. Engard’s *More Library Mashups*, Stanford’s Bess Sadler explains how the mashup approach to the library’s collections offers opportunities to present diverse data formats in a unified interface design with a consistent look and feel.

Stanford’s SearchWorks: Mashup Discovery for Library Collections

*Bess Sadler*

SearchWorks (searchworks.stanford.edu) is Stanford University’s library discovery portal, built with the open source Blacklight (project blacklight.org) software. SearchWorks provides access not only to books and MARC records, but increasingly also provides access to Stanford’s digitized archival materials, photographs, sound recordings, and geospatial data. This mashup approach to the library’s collections has given us opportunities to think about the best ways to present diverse data formats in a unified interface design with a consistent look and feel. We are also increasingly publishing APIs for our digital library collections and infrastructure, and we hope to become a source of data for mashups that others will create.

A Shift in Perspective

The history of mashup applications is one of shifting perspectives and user expectations. According to the Wikipedia article on the history of mashup applications (en.wikipedia.org/wiki/Mashup_%28web_application_hybrid%29), in the Web 1.0 business model there was an expectation that a website would manage and control access to all of its own content. With the shift to Web 2.0 thinking, there is no longer an expectation that all data will be stored in a single system or an expectation that a single delivery mechanism is sufficient for many collections. Discovery and delivery interfaces increasingly want to include content from other systems via APIs, and they increasingly
want to expose their own content via APIs for inclusion elsewhere. SearchWorks is an example of how that shift in perspective has played out in library discovery with tremendous success. By concentrating on establishing application-building patterns, instead of putting effort into monolithic systems, Stanford’s digital library has been able to accommodate a wide range of content and make it available in a variety of contexts, something that would have been impossible in the Web 1.0 mindset.

**Blacklight as an OPAC**

Blacklight is an open source discovery application for collections of digital objects. Originally conceived as a next-generation online public access catalog (OPAC), it has been widely adopted by institutions wishing to improve discovery for content traditionally managed by an integrated library system (ILS). The ILS is primarily a tool for managing collection inventory and library business workflows, and although ILS vendors have usually packaged a user-facing search component (the OPAC) with ILS software, these have often failed to measure up to rapidly evolving user expectations for what web-based search applications should be like.

Online catalog systems have long been a source of frustration for librarians and patrons alike. This frustration is due in part to librarians’ desire—and patrons’ expectations—for their online catalog to keep pace with the rapid development seen in other web applications, in contrast to the slower progress of new library catalog features. The relatively slow development of new features for library catalogs is in part caused by most catalogs’ close ties to ILSs that run all of a library’s transactions. Upgrading an ILS is a difficult and expensive process that has the potential to disrupt workflows throughout the library and is undertaken only when absolutely necessary. This is a suitable pace of change for business and workflow operations but not for front-end systems used by patrons accustomed to the pace of development of services like Netflix (netflix.com), Facebook (facebook.com), Google (google.com), and Flickr (flickr.com), each of which rolls out new features every few months.

By indexing the contents of the ILS into web-scale open source search engines like Solr (lucene.apache.org/solr), libraries can define their own measures of relevancy and customize indexing procedures
to match the needs of local collections and audiences. By providing a flexible front end for a Solr index, one that allows different behaviors for different kinds of objects, Blacklight allows for flexibility and new innovation possibilities. This flexibility will become ever more important as libraries grapple with increasingly diverse collections and ever higher user expectations for delivery and discovery of content.

**Discovery for Digital Collections**

Digital library content has often suffered from a lack of user-facing visibility. At Stanford University Library, which has invested heavily in the creation and curation of digital library content over the years, Blacklight has provided a generalizable approach to discovery. Just as de-coupling the discovery interface from the ILS provided new opportunities for innovation, de-coupling discovery from object management systems like Fedora Commons (fedora-commons.org) has breathed new life into digital library interfaces without requiring the re-engineering of already established back-end management and preservation systems. Blacklight provides an opportunity for unified discovery of content across collections as well as user interaction and development patterns that can be used for building tailored discovery applications for specific collections or content types.

**Unified Discovery for Heterogeneous Content**

SearchWorks is Stanford University Library’s primary collection discovery interface. The Solr index that SearchWorks runs on contains not only the millions of MARC records indexed from our SirsiDynix ILS, but also content from our digital library, including digital collections we host for other institutions, and content that has been submitted via our self-deposit interface for preservation in our digital library.

In Figure 17.1, two digitized collections from Stanford University Library’s special collections are visible. Because Blacklight provides the flexibility to define different behaviors for different kinds of objects, we were able to define behavior tailored to the needs of a digitized image collection. This included creating a collection-level view, which merged metadata from the MARC record with metadata from the digital library, and a “filmstrip” view of a collection, which allows a user to quickly preview the images visually.
In Figure 17.2, the view of an individual image is shown. Note that a member of a collection maintains knowledge of its collection membership (RDF triples modeling each object’s relationship to other objects is part of the metadata we store and can draw on for discovery), as seen in the left hand column where contextual information about the collection this image is part of is displayed.

Each object in our digital library also has a permanent URL (purl), as seen in Figure 17.3, and that purl has a machine-readable version, accessible via adding .xml to the end of the URL, for a full xml rendering of the public parts of a digital object, or .mods for an XML serialization of the MODS record for that object. Providing permanent, predictable URIs means that users of the collection can provide citations for publication, and that not only is our catalog a mashup
itself, but it is also a source of re-mixable data that allows others to include digital library content in their own mashup efforts.

User-Submitted Data

One rapidly growing source of data in our digital library is user-submitted research data. The Stanford Digital Repository (SDR) provides a permanent home for data publication by Stanford affiliated researchers (Figure 17.4). Depositing in SDR provides researchers with a permanent citation URL they can use in publications as well as a preservation strategy they can reference in their data management plan—a formal document that outlines how researchers will handle
the data they collect during the research process and after the project is completed. Data management plans are increasingly required by research funders, including many federal government funding agencies. The presence of research data in our digital library has meant expanding SearchWorks to accommodate its presentation and discovery. Unlike digitized surrogates of physical library holdings, these objects never had a MARC record in the ILS. Instead, SDR uses the MODS standard, which has meant defining new MODS-based presentation logic in SearchWorks.

We further leveraged the modular nature of our digital library architecture by using this as an opportunity to encapsulate MODS logic in ruby gems, including one for general MODS parsing (github.com/sul-dlss/mods) that should be broadly useful to the library software community, one that contains local Stanford customizations
(github.com/sul-dlss/stanford-mods), and one that contains display logic (github.com/sul-dlss/mods_display) so that presentation of MODS records is consistent and easy to add to any of our applications that needs to include that functionality.

**New Views for Archival Collections**

Part of the shift in perspective of the mashup mindset is letting go of the notion that a single interface will be able to accommodate the needs of every user and every collection. Instead, the Stanford Digital Library strategy is to adopt and build on workable patterns, applying them in new contexts as the need arises. While SearchWorks represents an opportunity to bring many diverse object types into a unified discovery solution, the application patterns established for SearchWorks have also allowed us to easily build sustainable discovery interfaces tailored for specific collections.
One recent example of the success of this strategy has been the interface for the Bassi-Veratti Collection (bv.stanford.edu), visible in Figure 17.5. The online collection is a multi-year collaboration of the Stanford University Libraries, the Biblioteca Comunale dell’Archiginnasio, Bologna, Italy, and the Istituto per i Beni Artistici, Culturali e Naturali della Regione Emilia-Romagna, to produce a digital version of the archive of the influential 18th-century scientist Laura Bassi. Bassi was the first woman to earn a professorship in physics at a European university and also to be offered an official teaching position at a European university. Particularly given the significance of the collection and the tremendous effort that went into its organization, description, and digitization, we wanted to create a discovery interface that did justice to the content.

We were greatly aided in this aim by the fact that we had well-established patterns already in place for the behaviors of digital
objects. Just like the images and data sets already referenced, each
digital object has a purl and a MODS record. Because we had already
invested in building MODS-based indexing and presentation behav-
iors for Blacklight, setting up a Blacklight instance that harvested
only the Bassi-Veratti objects from the larger digital library and sorted
them into collection-appropriate facets for discovery was much easi-
er than it otherwise might have been.

We also benefited from participation in a larger community of open
source software development. Blacklight has a variety of pluggable
UI components developed by community members and contributed
back to the project. One of these is the date slider visible in the left
column in Figure 17.5, originally developed by the Johns Hopkins
University Library. It allows users to narrow their search to a range
of dates, while also providing visual feedback about how much of the
collection is represented within a given slice of time. One can see at
a glance that this collection spans the years 1591–1836 and that the
bulk of the material is from the late 1700s and early 1800s.

The Bassi-Veratti archive also benefited from the increasingly
rich world of geospatial mashup components. Using Google’s geo-
coding API, we were able to transform place names into points on
a map, allowing for geographical exploration of the content and
allowing users to get a better sense of where the story told by this
archive took place.

Finally, because sometimes the best tool is the one you’re most
familiar with, we also provided access to the complete archival find-
ing aid, as seen in Figure 17.6. In the spirit of mashup, however, we
provided access not only to an online display and a downloadable
and printable PDF, but also to the actual EAD XML, in case research-
ers want to index the content themselves or run it through some
other EAD presentation system.

**Discovery for Space and Time**

Near term development for our digital library will include expanding
our infrastructure and software components to include more support
for geospatial data. The geographic information system (GIS) is play-
ing an increasingly important role in many academic fields, and GIS
data are already well represented in our faculty-submitted content
in SDR. The management of GIS data in libraries is also experiencing
some exciting but challenging shifts in perspective, as GIS assets such as licensed data sets are exploding in popularity and requiring more sophisticated and systematized management strategies, such as those offered by a robust digital library infrastructure.

Current development efforts are focusing on data modeling for GIS content, attempting to encapsulate logic about ISO-19139, the GIS metadata standard we’ve chosen, the way we have MODS and MARC. A major part of our strategy for enabling management and discovery for GIS content is establishing predictable behavior patterns for each object and building discovery interfaces that can be assembled from re-usable components, as we have for other kinds of data. This will also provide us with the opportunity to enable more spatial discovery of non-GIS content, such as the map interface on the Bassi-Veratti website. Geolocating digitized historic maps and semi-automated place name entity extraction for full-text and archival
collections are some of the opportunities we see on the horizon for improved spatial exploration of library collections.

Data Integration Challenges

The Stanford digital library has benefited greatly from embracing the mashup ethos, but combining data sources also comes with some inherent challenges. These are well recognized and widespread enough to have been detailed in the previously mentioned Wikipedia article on web application mashups, and include text data mismatch, abstraction level mismatch, and general data quality issues.

Text data mismatch can be understood as a problem with ambiguous naming. Text representations of personal names and place names are inherently ambiguous, since names in the real world are seldom unique, and names can shift over time (e.g., Istanbul was Constantinople, Jane Stanford was once Jane Lathrop). Data-driven applications, however, often assume the ability to gather sets of records together based on unambiguously shared characteristics. Libraries are well-versed in maintaining name authority records, but creating unambiguous name authorities is a challenging enough task for published authors of books; current name authority methods are unlikely to scale to the level needed to accommodate self-deposited data, journal articles, and persons of historical interest mentioned in archives. Instead, a shift in practice to recording both text representations and linked data URI representations in metadata, where possible, greatly increases the ease with which a given collection can be incorporated into mashup-style applications.

For example, as we establish cataloging practice for GIS data in library collections, Stanford University Library has started assigning entries from the GeoNames ontology (geonames.org), giving us much more value (e.g., the ability to co-locate records about the same place but that use a different place name, greatly improved mapping functionality) for a comparable level of cataloging effort. Similarly, by leveraging efforts like ORCID (orcid.org) and encouraging researchers to link their SDR deposits to their ORCID number, we hope to gain the ability to disambiguate authors with the same name as well as to make the data in our repository compatible with research aggregation services outside the library context.

Abstraction-level mismatch refers to the challenges inherent in combining full-text data records in the same discovery interface with
metadata-only records. If naïve relevancy algorithms that simply look at word count matches are used, full-text records will overwhelm briefer records. One of the benefits of using a system like Solr is that we can tune our own definitions of relevancy. Combined with a robust suite of automated repeatable tests gleaned from staff and community feedback about how relevancy should behave, we are able to add more and more content to our discovery index without overwhelming and breaking search strategies that our users have come to count on.

Abstraction-level mismatch is also present in the challenge of how to represent collections of objects that should be discoverable both individually and within a collection context. A digitized letter from an archival collection, for example, might be discoverable and interesting on its own, but often much of the cataloging effort has occurred at a collection level, and archival content is also most valuable when it can be viewed in context. To represent objects at the collection level, we’ve incorporated behaviors such as the filmstrip view of image collections visible in Figure 17.1. To tie individual objects back to their collection context we’ve incorporated patterns like those seen in Figure 17.7; the breadcrumb links above the main image allow the user to quickly see what other letters were in the same folder as this one, and the button labeled “View this item in content inventory” will take the user to this object’s location within the full context of the archival finding aid.

Finally, data quality issues are an ever-present challenge in any information system, but re-contextualizing and re-using metadata in new ways exposes previously hidden inconsistency. To address this we are attempting not only to remediate erroneous data when we find it, but also to put in place more automated data consistency procedures. Encapsulating expected behavior for MODS records in a suite of software tools, for example, allows us to quickly determine whether the records for a given collection match our expectations and improves our ability to achieve some measure of consistency across diverse sources of content.

The Future

Stanford University Library has reaped tremendous benefits from embracing a mashup mindset in its applications. In addition to allowing for more rapid innovation, it has challenged us to re-think
our cataloging and data quality processes to keep pace with desired new features. We have also benefited greatly from participation in larger communities of software development effort, in the form of specific collaborations like Project Blacklight and Project Hydra (projecthydra.org) as well as informal communities of practice like Code4Lib (code4lib.org).

Future efforts will include further investment in leveraging linked data vocabularies in our metadata practices, and as our body of linked data grows, discovering new ways of leveraging connections between objects to enable new forms of discovery and exploration. We are also investing in the creation of training curriculums and improved software documentation, so that the data and infrastructure we create will more easily be incorporated into the fabric of mashup applications elsewhere.
About the Author

Bess Sadler is the Manager for Application Development in the Digital Library Systems and Services group at Stanford University Library. She is one of the co-founders of Project Blacklight (http://projectblacklight.org). She has spent over a decade working on digital libraries and related digital humanities projects, with a particular focus on enabling better discovery and navigation for digital collections. Bess also contributed to a chapter in the first edition of Library Mashups and is pleased with how far we’ve come since those early days of mashup practice in libraries.