Single Search:
The Quest for the Holy Grail

Leah Prescott
Digital Projects Coordinator
Getty Research Institute

Ricky Erway
Senior Program Officer
OCLC Research

A publication of OCLC Research

Single Search: The Quest for the Holy Grail

Leah Prescott and Ricky Erway, for OCLC Research

July 2011

OCLC Research
Dublin, Ohio 43017 USA
www.oclc.org

ISBN: 1-55653-426-4 (978-1-55653-426-3)
OCLC (WorldCat): 741331876

Please direct correspondence to:
Ricky Erway
Senior Program Officer
erwayr@oclc.org

Suggested citation:
Prescott, Leah and Ricky Erway. 2011. *Single Search: The Quest for the Holy Grail.* Dublin, Ohio: OCLC Research. http://www.oclc.org/research/publications/library/2011/2011-17.pdf.

Table of Contents

Acknowledgements

The authors and OCLC Research gratefully acknowledge the effort of the OCLC Research Library Partnership Single Search Working Group. Günter Waibel conceived and launched this activity before leaving OCLC Research for a new position at the Smithsonian Institution. Without his inspiration and the Working Group's significant contributions, this report would not have been possible.

- Heather Caven, Head of Documentation & Collections Management Services, Victoria & Albert Museum
- Emmanuelle Delmas-Glass, Collections Catalogue Specialist, Yale Center for British Art
- Dennis Meissner, Head of Collections Management, Minnesota Historical Society
- Youn Noh, Digital Information Research Specialist, Yale Office of Digital Assets & Infrastructure
- Paul R. Pival, Public Services Systems Librarian, University of Calgary
- Leah Prescott, Digital Projects Coordinator, Getty Research Institute
- Margaret Savage-Jones, Library Systems Manager, Wellcome Trust
- Francesco Spagnolo, Curator of Collections, The Magnes Collection of Jewish Art and Life, UC Berkeley
- Ching-Hsien Wang, Manager of Library and Archives System Support Branch, Smithsonian Institution

Introduction

As part of an OCLC Research investigation (2009) into convergence issues for libraries, archives and museums (LAMs), a series of workshops at five institutions raised many examples of successful collaborations, as well as obstacles to collaboration. One topic that came up at each of the institutions was the desire to offer users a way to easily discover the institution's resources regardless of where and how the resources might be managed. Following the workshops, the discussions about single search were continued with other institutions, many of whom had implemented single search, but were unhappy with the results and wanted to re-implement it. Others were just beginning to consider the options available to them. OCLC Research thought it would be useful to get some successful implementers together to share their experiences and help those who were just getting started.

Users can search Internet resources through a single search engine query, yet often the resources of a single cultural institution or university campus are segregated into silos, each with its own dedicated search system. The prominence of multidisciplinary research, the increase in the use of primary materials, and the desire to make new connections across disparate materials all would be advanced by the offering of single search to open up all the collections to the researcher.

There was some disagreement on this topic among OCLC Research staff. Some thought that single search at the institutional level was a red herring, pointing out that the goal should be to make the content accessible where the researchers usually work. But others felt that there were many justifications for single search at the institutional level. Once the user arrives (physically or virtually) at an institution, it's ridiculous to expect them to navigate the organizational structure of the institution and learn a new search approach at each stop along the way. A significant motivation for single search is to be able to show funders and content donors the riches of the institution and to enlist their contributions to enhance the collections. And integrating collections at the institutional level is the first step to exposing them to aggregators and search engine spiders for network-level discovery.

The report from the LAM workshops (Zorich 2008) highlighted nine catalysts for successful collaboration: vision, mandate, incentives, change agents, mooring, resources, flexibility, external catalysts (such as audience, peer institutions, funding organizations, and professional organizations), and trust. Many of these come into play in the single search environment.

OCLC Research facilitated the working group of nine single search implementers through discussions about the opportunities for, and obstacles to, integrated access across an institution. They told their stories, categorized a list of issues, and created and answered a questionnaire looking for similarities and differences in their approaches.

This brief report summarizes those discussions and highlights emerging practices in providing access to LAM collections, with a particular emphasis on successful strategies in the quest for single search.

The goal of the report is to foster successful single search implementations by sharing the experience of the working group with those who want to create single search but don't know where to start. In doing so, the report strives to be relevant to a wide range of professionals in different roles in the implementation. The broad considerations and concrete examples in the report can serve as a catalyst for discussion; they have been chosen to inspire action, rather than to inhibit it.

There have been sufficient single search implementations, and enough experience has been accumulated to allow us to take a step back and synthesize today's lessons for the next phase of implementers. If those new to single search can start where others left off, they will be far more likely to succeed.

Ricky Erway
OCLC Research

Institutional Considerations

An institution's decision to implement single search, and to devote or seek new funding for the endeavor, relates to issues of identity and change. Single search projects generally aim to better serve the needs of users as well as collections, and institutional identity is closely tied to both.

Motivation

Looking at the single search implementations at the nine working group members' institutions gives us a glimpse of the motivational factors for embarking on a project that can impact institutional priorities. The main factors appear to be:

- The desire to respond to the needs of users
- The need to harmonize collection management practices
- The willingness to redefine priorities
- The sensitivity to respond to factors that motivate individuals

User Needs

Within the working group, the most prevalent motivational factor is the desire to serve users by helping them find all content relevant to their informational need; as one member expressed, "to see topically unified resources, regardless of institutional home."

With the significant increase in digitization activities in all types of cultural institutions, there is increased focus on providing single search access that includes this new wealth of digital resources. And digital resource management is a significant factor in planning an integrated access project. Whether single search involves highlighting aspects of well-known collections, finding collections that would otherwise be overlooked, and reconciling catalog records and digital objects, the attempt to simplify the user experience is a goal that impacts the whole institution.

Collection Management Practices

Another important motivational factor expressed by the working group is the ability for collection staff to manage objects and information created through widely different professional practices. Single search allows librarians, archivists, curators and registrars to obtain a bird's-eye view on holdings and descriptive records. The resulting perspective enhances collection understanding, expands the awareness of varied professional expertise and enables streamlined collection management workflows.

Institutional Priorities

Integrating the intellectual output of different areas of an institution can in some cases fundamentally change how an institution identifies itself. This is more often the case when the affected areas comprise the entire institution, and not as likely when the LAM component is part of a much larger organization such as a university.

Cultural institutions evolve in different ways, and the impetus to develop a single search mechanism can come from different directions, such as Boards of Trustees, managers, collection professionals, user constituencies, or all of the above. Impetus for a project can be "top down" or "bottom up," although it helps to have a mandate and tangible support from the institutional leadership. Some of the participants in the working group clearly have a mandate to put integrated access in place, but even with a mandate they may be lacking the institutional muscle to compel or cajole weakly aligned program units to work towards a single search goal. Institutional leaders may finance and direct such an objective, but if it is not prioritized and enforced at operational levels the outcome may not be fully realized. There is value in such circumstances of utilizing "managing up" as a strategy toward achieving single search goals.

Individual Motivation

In addition to the factors that motivate an organization, there are factors that motivate individuals involved in a single search project that are equally important. In one instance a working group member reported that the staff were "equally motivated but with varying goals." Another noted that the "public services staff [didn't] seem to be as motivated" in their institution. Other personal motivational factors reported included the "ability to explore innovative cultural and professional horizons," and "intellectual curiosity, the challenge of creating something new [and] learning new skills."

Often there are varying levels of motivation in a wide-reaching project, and dealing with these issues among collaborators will likely involve defining a vision, demonstrating a practical path to achievement, and developing champions in key areas of the institution.

Single search projects can be complex and may require activities such as data-mapping and data migration as well as learning a new system. Frustration and weak support are understandable reactions to such big changes, and one or two programs or units may need to carry the ball at first and work to overcome the challenges to collaboration posed by other units. One working group member explained that their "working group [is] motivated and [has] been selected for their interest, skills and capacity to work on the program." The early leaders in the effort will develop a real sense of mission about the value that is added to access by pursuing a single search solution, and their efforts may serve as a proof of concept that, if successful, can be used to begin pushing the challenge back to more recalcitrant units. Illustrative is this comment from one member of the group: "We worked with a couple of early implementers to showcase the positive results. Later we approached more natural units to move forward. For units that push back, we simply wait until they are ready. Since a majority of the major units are on board now, the hold-out units are changing through peer pressure."

Even if the objectives of disparate programs within an institution align well, it is not uncommon for single search projects to lose traction because of motivational disparities among the partners, requiring support from enterprise leaders in clarifying priorities. A successful test involving a few formats or custodial areas might be a powerfully persuasive tool that can not only convince peers within the institution, but also achieve greater tangible support from enterprise leaders.

Organizational Structure and Institutional Culture

A single search project must be considered within the context of a given institution's organizational structure. Single search is not just for large institutions, but as one participant noted, "A small, nimble institution can innovate quickly and afford the risks involved in innovation. Staff with different professional backgrounds can be enthusiastic about working together." However, larger institutions are likely to have a much broader range of expertise to draw upon.

When a project cuts deeply across an organizational structure, it is not uncommon for the structure itself to change. Some organizations will define whole new departments and fundamentally change the formal structure of the institution, while other institutions will form cross-departmental teams that don't change the formal structure at all. Because they work within the established organizational structure, these teams require clearly defined authority definitions, and strong, fluid communication avenues. Organizations may have the skills needed in-house, but they may be spread across several different departments from the Web team to the Information Technology (IT) department, or from the library to the collections management team. These teams may report to different senior managers and have

different agendas and objectives. There is a need for integration across the LAM sectors, and to put the user experience at the center.

Even if the organizational structure doesn't change, there will likely be significant policy changes of many types in areas such as rights, processing practices, metadata creation, and others.

Developing an integrated access solution offers many challenges that may be political, professional or institutional in nature, and important aspects include how public access enters into an institution's values and priorities; how archivists, museum curators and librarians view their respective responsibilities and methodologies; how different programs compete for resources within a parent institution; and how disparities in the audiences are served.

It is important for the success of the project to have representation not only from the primary LAM stakeholders, but also from IT units, as weak motivation within the IT area of an organization has the power to paralyze such a project. It can be difficult to communicate complex information concepts across a technological divide and develop a common understanding of the ultimate goal. One group member expressed being "unhappy that so much of the design, objectives and search mechanics are defined by IT staff rather than program staff." It is crucial that there be adequate interaction between these groups so that a common vision can be developed and common goals established.

It is also a challenge to get the disparate program partners to align their respective goals and objectives. As one group member stated, single search "has become a political issue at this point." Misalignment might be the result of professional differences—the "great LAM divide." For example, curators, archivists and librarians may assign different importance to broadening public access. Within the working group examples of misalignment include observations such as "librarians don't like not knowing specifically where a record came from." Another participant was "surprised to learn how little curators seem to want to expose their collections, compared to librarians," and another observed that in their institution "the archivists are interested in retaining access through the hierarchical display in their current archives system and not replicating that in any single search interface."

The reality of achieving an integrated access vision could mean overturning years or decades of institutional thinking, which has segmented collections management practice among the three different sectors of LAMs. Existing methodologies in many cases represent the embodiment of distinct professional values and perspectives. Program units within an institution may be locked into longstanding competition for resources and status. Partners will need to acquire and master new standards, tools and workflows and finding a new middle way can be challenging. Descriptive metadata and systems from one program may preference

public access, while another program may emphasize preservation information, and still another, ownership and rights management. Curators and program managers may feel their own professional values encroached upon, or they may fear that important metadata are being cast aside in the single search paradigm. These are significant impediments to achieving the buy-in necessary to enable single search.

Some potential partners may fear that single search will effectively eclipse—or even replace—their own discovery systems. If the individual systems are being retained, it will be necessary to make (and enforce) reassurances in order to build strong partnerships. There is logic to existing discovery systems, and single search objectives can support rather than replace them. Respect for proficiencies should be fostered, and equal participation in the conversation should be ensured. The project may be perceived primarily as a technology project, an information management project, or public access project, and it is important to respect and balance other perspectives.

Funding

Funding models include combinations of regular operational funds (which highlight the pursuit of single search as a core institutional need), specially allocated funds (which highlight the research aspect of single search pursuits), and grant sources (which often connotes special project status). If grant funding is a strategy, it may be advisable to include staff from the development area in the core project group. Their buy-in could be fundamental to presenting the project to a granting agency, or to knowing which granting agency might best align with the project.

In the best possible scenario, an alignment of governance, collection management, and fundamental understanding of user needs are the forces behind the kind of innovation that has made single search projects successful, and funding that is balanced between operational and special budget resources reflects that.

Technological Considerations

There are several technological strategies that can be adopted in the implementation of single search functionality, as evidenced by the institutions represented in the working group. Diversity of collections provides a significant challenge, and a mix of approaches which take account of particular local circumstances may be the best solution. Increasingly vendors are recognizing the need for a modular approach, but existing LAM systems are often well established and have very different indexing, search and display capabilities.

Metadata flows can be push (e.g., batch export) or pull (e.g., API calls). The former gives the metadata producer control over how and when metadata are disseminated, avoiding the unanticipated spikes in traffic that may occur with the latter. Pull methods allow the metadata consumer to make calls in response to requests, typically initiated by the user, generating metadata that are likely to be more up-to-date and better configured for the consuming application. Push methods may be favored when hardware cannot easily be scaled, when demand is difficult to anticipate or estimate, or when collections are largely static. Pull methods may be appropriate for highly dynamic collections and may reach a broader audience.

Single System

The most tightly integrated method for enabling single search is to have all data entered into a single system using the same data entry tools. The database itself is the only limitation on what data are collected and how they are recorded. A single system may have a common data format where regardless of the type of LAM material, the descriptions must all fit into a common structure. A single system can also be modular, with separate data entry tools for different areas, but on the back-end the data are all synchronized and stored in the same data structure.

While this method may have less technological overhead, it is also least likely to afford an institution the ability to manage each of its LAM collections according to specialized professional practices. When a collection with its own context is juxtaposed with other collections in other professional contexts, there can be an uncomfortable contrast. In addition, there may be unique functionality required for various types of materials that can complicate the structure of the whole. For instance, there may be library material that must

circulate, archival material that must be put into a finding aid structure, or museum data that must include the ability to finely describe the physical aspects of a collection as well as a public level display description (exhibit labels, for instance). Even the concept of "collection" is often different between the separate professions. Richard Rinehart describes this in a report from 2003 that discusses integrating materials for the Online Archives of California:

> "Archival practice defines a collection as the unself-conscious by-product of the activities of a person or organization. The collection comes to the archive where it is kept in the original order of the previous collector and described in terms of that provenance and its contents. A certain amount of objectivity is desired in the description, which is limited to the scope and history of the collection. Archival collections are maintained with the integrity of the whole, also defined by provenance, and are described from the top collection level, and only sometimes at further levels of detail. . . . Museum collections are often acquired at, and oriented toward, the item level where most description tends to focus less on the archive-like history of the object, or the library-like subject of the object, and more on the 'thingness of the thing' including the physical properties such as material, dimension, and object or genre classification. If expanded cataloging can be afforded, the catalog record often includes details about the creator (not necessarily collector) of the artifact, as well as interpretive text. Those are the respective descriptive practices in theory anyway—practices that at first glance seem almost irreconcilable in creating an integrated access system."

Single systems may be most appropriate when one LAM area greatly outweighs the others. For instance, an organization may find it acceptable to catalog a small number of 3-D objects in a much larger library catalog, or describe a small collection of archival materials in a museum collection management system. It may also be an acceptable solution when an overall collection, including all LAM materials, is so specific to a single topic, that the practices of the individual professions are a much smaller concern. Overall, it can be the most cost-effective method, as one member of the research group noted, "being able to use one system brings efficiencies and cost savings."

Multiple Systems

Harvesting to Central Repository

Another type of system or method that may be used to integrate LAM collections is to keep each type of data in separate domain-specific repositories and subsequently "harvest" the data into a separate central repository. In this instance each specific area can have a system that caters to the needs of that domain, with all appropriate functionality. This is often the case with large "digital encyclopedia" projects, and the challenge is to coordinate what could be widely divergent data structures. The data must be coordinated across the individual repository levels so that harvesting is successful.

Much care is needed to decide which levels of data to harvest and how to present the data to the user. One of the institutions represented in the working group is investigating use of an API in the archives system which will return the hierarchy as XML which can then be rendered to reconstitute the same tree structure within the overarching single search delivery system.

This method is most often considered when there are a limited number of data "streams," and when the aggregated whole is not so large as to make a central repository unwieldy. One of the challenges with this method is in coordinating changes between LAM repositories and the central repository. For instance, if data get edited in a pre-existing record in a remote system, how does that change get incorporated into the central repository? And does that methodology go both ways, so that data enhancements flow back to the originating repository? Some harvesting paradigms cannot facilitate a two-way data exchange, so policies must be developed and workflows built that account for this reality.

Federated Search

A third methodology is a federated search strategy. It, too, involves coordinating data between individual LAM systems, but instead of gathering and replicating the data, the central system translates a user's query into syntax that is specific to each remote system, and brings back result sets that are then coordinated and presented to the user.

The upside to this method is that LAM collections can be managed in domain-specific systems and the data do not need to be reconfigured. Various types of gateways are built to interpret the incoming query for the individual system, as well as interpret the query result so that it can be combined with result sets from the other systems involved.

This method is also attractive because theoretically it is more extensible than a centralized repository, but the major downside to this method has proven to be speed. The more repositories it must query, the slower it may become. Often the presentation of the integrated result set can only be as fast as the slowest gateway, although some systems now start presenting results as soon as they come back and integrate with other incoming data sets on the fly. Depending on the uniqueness of the remote systems, the gateways can be quite challenging to build.

Another downside is the difficulty of offering relevance ranking and faceted manipulation of search results. Often users are presented with separate sets of results from each system. If there are a lot of systems and the content isn't very logically dispersed amongst them, this can be confusing to the user.

Central Index

A method that is increasingly being used is to maintain separate domain-specific systems and create a separate centralized index. At the time of a query, the index is searched rather than the stored values. This is the strategy that search engines such as Google use, so that they can return results quickly. When a user performs a search, it is the index that is being searched rather than the individual systems. Indexes are created through harvesting, and as with record harvesting, there are complexities, including decisions on harvest frequency and level. Harvesting may be daily or less frequent, and scripted manually or automatically. It may harvest only newly added, updated or deleted data, but some systems may not be able to identify these and instead make the full dataset available on a regular basis. Harvesting and indexing can be labor intensive, but in this model, there is more flexibility in the presentation of the aggregated results to the user.

One of the features often offered to users is faceted navigation. Faceting takes advantage of categories of data, which allow a user to "drill down" to increasingly refine their search results (many commercial Web sites use faceting to help a buyer find exactly what they want). Difficulties in faceting often result from the different data sources; this was a common thread for the working group. Most sites reported using facets, developed commercially or in-house, as a way to enhance user navigation and assist with determining relevance.

Digital Asset Management Systems

Accessing digital objects is an important part of the user experience when searching library, archive and museum collections online, and social media sites such as Flickr, YouTube and Facebook, have increased user expectations. Users not only wish to view and listen to digital assets, but to visually browse, tag and reuse them. Effectively managing digital files is a key responsibility of the curatorial units. Access to, and preservation of, the files often requires a digital asset management system (DAM). [The DAM is not an essential part of a single search system (in fact digital content isn't necessarily an essential part either), but since DAMs often come up in discussions of single search, we include them here.]

The value of accessing digital objects alongside catalog information is backed up by research. When asked what improvements or changes they would like to see, a majority of users wish to see more images made available while fewer wish to see more detailed catalog records. LAM institutions, large and small, are also collecting more and more "born digital" material, from e-books to digital art work, films and audio files to digital archives. The value of digital assets is even more significant when metadata are skeletal or nonexistent, as they can provide visual browsing to enable users to discover and access collections, despite sparse descriptions.

These digital collection assets may be ingested into a single digital asset management system, but could be catalogued across multiple collections management systems in multiple domains. Digital collection assets can raise significant technical, metadata and rights management issues when it comes to delivering single search.

Larger organizations are most likely to have invested in one or more DAMs, as the cost may put it beyond the reach of many small to medium sized organizations (open source systems for digital asset management were not used by any of the working group participants). In addition to the initial cost of implementing a new system, there are ongoing staff and maintenance costs. However, these costs could be offset by system efficiencies elsewhere, for example an existing system can be adapted to cater to more than one type of collection, or existing paper-based workflows can be replaced by automated processes. Implementing effective workflows requires interoperability between the DAM and other "back of the house" collections systems, but not all DAMs have this functionality and this can inadvertently reinforce the silo nature of internal information sources.

Whether or not a DAM has been implemented, there needs to be a clear understanding at the management level of why digital assets must be systematically named, managed and assigned a unique asset identifier. The danger of not doing so is to risk asset silos being created at a local level, or assets saved onto storage devices without appropriate digital management processes such as backup and preservation.

Implementation of a DAM, (or a centralized repository and policy framework for digital assets), provides an opportunity to build organizational understanding of the aggregate value of the collection. Adoption by internal stakeholders of a shared metadata standard for digital assets, and digital workflow for library, archival and museum items, can have a positive effect when implementing single search.

Open Source vs. Commercial

The decision to use open source software or to adopt a commercial package may depend on skills in-house and what the local infrastructure and policy allows. Open source tools can offer more functionality and economy even when in-house programming and support costs are factored in. Change is being driven by awareness of open source alternatives, making them a serious alternative or addition to commercial products—and they are increasingly used in commercial offerings. The development work undertaken by one site was looked at by three vendors during the process of developing their own products.

Within the working group, the most common indexing software being used is Lucene, with Solr as the search appliance, although at least one institution is programmatically creating index

records as "Solr docs" and populating the index with them rather than indexing with Lucene. While Solr and Lucene can be implemented by an institution, they are tools that are increasingly being used by LAM vendors.

Lessons Learned

There is no single answer to the technology underlying single search. Different approaches work in different situations. Lessons learned about systems architecture are that there are significant efficiencies in bringing repositories together, but such interoperability requires an investment and is not without risk. Thorough technology planning requires feasibility studies, validation of assumptions and clear communication with suppliers about specifications. There is a need to choose extensible systems so that new partners with totally different needs can be included.

Metadata Considerations

As we have seen, LAMs often have what amounts to different information world views, and while they have fundamental similarities, metadata practices have not developed in concert with one another. In addition, the three LAM domains contain collections that are different in subject, format, and context; each of them may have fundamentally different metadata requirements. Juxtaposed against this reality is the ever-increasing demand by users to be able to access the entire universe of data in ways that are simple, seamless and metadata format agnostic.

The quality and quantity of metadata affect the quality of the single search experience. Because metadata can be expensive and labor-intensive to create by hand, it makes sense to gather any metadata that are created—including automatically generated metadata, that are potentially useful to end users. A variety of platforms are used to generate metadata, such as library information systems, content management systems, digital asset management systems and media production tools. Many types of software generate header metadata that can be enriched and edited. For applications such as single search, the ability to access metadata from source systems is crucial. Even in fully integrated systems, metadata need to be ingested from production systems and exported for conversions, upgrades and migrations.

Standards

Adherence to standards is essential, as integrating metadata is one of the most challenging aspects of a single search implementation. Libraries and archives have a long tradition of descriptive metadata and museums are increasingly developing metadata standards to suit their needs. But data content standards and data value standards used by these institutions are different. For example, the metadata format for describing bibliographic materials (MARC) is different from the formats used for cultural objects, such as CDWA (Categories for the Description of Works of Art). Additionally data content standards and data value standards in these communities are in different stages of stability and maturity. For example, the data content standard, Cataloging Cultural Objects (CCO), is just a few years old, while Anglo-American Cataloging Rules have been in use for decades.

Single search has the challenge of providing users with seamless cross-collection discovery while at the same time respecting the standards used in each domain in order to represent the collections appropriately. A good example of the challenges that must be faced is attempting to integrate metadata from a museum collection, where there tends to be a one-to-one relationship between the object and record, and from an archive, where there can be a single record that represents thousands of objects. EAD (Encoded Archival Description) records have tree-like hierarchies with potentially many levels, and each level can contain rich or scant metadata. These multiple levels provide the user with the collection context; at the item level there may be little or no information. This becomes confusing when those records are harvested and displayed with flatter records such as museum object records, which may have a lot of description, but only at the item level.

Another issue that can arise is relative representation of the different data sets in the search result. For instance, item-level records can overwhelm collection-level records to the point of making that content virtually invisible. One participant in the research group confirmed the "need to be sensitive to [the] possibility that one source will overwhelm others due to numbers—but not necessarily importance."

The appropriateness of a metadata standard for a particular application depends on a number of factors, such as the users of the application, the features the metadata are intended to support and the content of the collections. Given that collections have many uses, some unknown or unanticipated, it is often useful to apply the most domain-specific metadata standard available to describe a collection. This provides an opportunity to record as much information as possible about a resource in a manner that is true to how its current stakeholder community views it. This information can subsequently be mapped and transformed for use in different applications.

Many metadata standards are now implemented as XML schemas, and this makes it possible to validate metadata records against many standard schemas. Validation may also be applied to values to check for data types, regular expressions (e.g., date formats) and strings (e.g., controlled vocabulary terms). Concepts drawn from an ontology may be checked for logical consistency or used for inference.

Vocabularies

Adopting published and well-documented controlled vocabularies and authorities will make the processing of the data by machines much more reliable. At the local/institutional level, controlled vocabularies and authorities have features that can enhance searching and improve access to end users. LAMs have traditionally used different vocabularies because they have developed data content standards and data value standards independently of each other.

Within the same professional communities there can be multiple vocabularies implemented. For instance, natural history museums and art museums might utilize the same collection management systems; although they are both museums, their vocabularies would be completely different.

Cataloging environments can unwittingly dictate vocabulary choices when connectivity with terminology Web services is not available. For example, an art museum's collection management system might not offer library vocabularies natively. Another problem of integrating vocabularies has to do with their difference in granularity, that is, one vocabulary will represent a concept on one level of specificity, while another at a more specific or less specific level, and often there is no easy way to automatically connect the two.

As a result of these challenges, LAMs often create silos of data that can be difficult to integrate. In single search, one of the ways this is manifested is by presenting discordant data in facets. For example, the name of a creator might be repeated twice in the same facet: once under the format preferred by the Library of Congress Name Authority and another time under the format preferred by the Oxford Dictionary of National Biography. These various problems can make for an unsatisfying user experience.

Dealing with heterogeneous metadata can be eased if a few steps are taken. The basic goal should be for data providers to create and contribute shareable metadata, e.g., metadata that are intelligible outside of the local context and coherent, consistent and standards-compliant. Ways to "harmonize" metadata may include the standardization of terms, such as date qualifiers ('ca.' versus 'c.' and 'about'), and implementation of shared lists of terms. Some vocabularies are more difficult to harmonize than others and partners might agree to focus only on values that will be integrated. Of course, harmonization is contingent on the willingness of data providers to collaborate in a way that will change their cataloging practices and their ability to carry out these changes. This route is likely to be time consuming and labor intensive, depending on the number of partners and their resources.

Another way to develop greater data interoperability would be to apply data content standards and vocabularies by material type and not by collection type/domain. A vocabulary such as CCO could be used for this as it is a data content standard used to describe cultural objects. For example, all globes would be catalogued with the same standards and vocabularies whether in a library, archives or museum collection.

Changing cataloging practice is bound to encounter many obstacles (cultural, technological, financial), and it may be desirable to also allocate resources to post-processing the data whenever possible. The process of programmatically normalizing spaces, punctuation, diacritics and case sensitivity is a useful step that will facilitate data retrieval. Automatic

stemming in the search process is also beneficial in that it will allow end users to get pertinent results whether they query on singulars or plurals.

The biggest challenge is to fully implement vocabulary interoperability at the aggregation level. Going forward, the goal of exposing and interconnecting data on the Web may benefit from "linked data" efforts. Linked data will offer data as a service rather than an end in itself and allow linking on the fly to other pieces of data. This is contingent on the data providers being able to express their metadata in RDF (Resource Description Framework), an ability greatly facilitated by compliance with metadata standards, including vocabularies.

Mapping and Crosswalks

It is a fundamental principal of building good collections to conform to known community standards for metadata, rather than developing idiosyncratic schemes. Possible target formats for data include MODS, MARC, Dublin Core, EAD, VRA, CDWALite, and Darwin Core. An institution that creates metadata in accordance with an accepted standard will be able to provide different views of its data with the creation and manipulation of crosswalks, rather than manipulation of record data itself.

A crosswalk is a tool that theoretically "translates" between different metadata standards. Some refer to the application of a crosswalk in a specific instance as a mapping. A mapping will include the idiosyncrasies specific to that given instance, and will often include explanations about how data must be processed before or after they are transformed.

Regardless of what technological method is chosen to implement single search, crosswalking is a likely a part of the data integration process. Whether deciding about what type of data to enter, bringing together data from different systems, building gateways, or defining search strategies, crosswalks are helpful tools and help to ensure satisfactory results for the end user.

In some ways, creating a mapping is like negotiating a compromise, as it will be relating two (or more) detailed and functionally-specific data sets. In order to do this, the data structures will need to be looked at a broader, more generic level. Mappings and transformations for single search applications will typically route from a domain-specific standard to a general standard, and subject specialists and domain-specific metadata experts who are accustomed to sophistication in the ways in which they describe their collections can sometimes have difficulty with the concept of losing any of that original data or functionality. When this is the case, an important distinction to make is that the mapping and subsequent data transformation is for the purpose of resource discovery, not to replace original cataloging.

Some common challenges and considerations that may come up in the course of creating a mapping include:

- There may be a many-to-one situation in which metadata in the original scheme can be mapped to more than one place in the target record.
- There may be no clear translation between a field in the original scheme and the target record. In this case there may have to be a less-than-perfect mapping.
- There is a fairly frequent tendency (when describing digital objects) to use the same record for both the metadata about the original physical object and the metadata about the digital surrogate. An example is with a creation date: if the record has data about both the original (physical) object and the surrogate (digital) object, it can be very difficult or impossible to know what the creation date is referring to. The further the metadata are being taken out if its original context (through harvesting for example), the less clear this can become.
- Crosswalking metadata is more problematic and error-prone when the source data are from a less rigid metadata scheme and when the target scheme is highly specific and tightly controlled.
- The original data may in themselves represent a mix of standards in that the source may have changed the standard over time but the legacy records were not changed accordingly.

One of the first steps in metadata mapping is to get a complete specification of all applicable standards. It is also good practice to investigate what crosswalks may already exist and are generally accepted. There should be a fundamental understanding of source and target data, as well as the overall priorities and purposes for the mapping.

In addition to clearly showing relationships between fields in different metadata schemes, a mapping should include:

- The direction of the mapping. There will often be imprecise matches, and in most cases the mapping will not work equally well in both directions due to different practices within LAM communities. If data must travel in both directions, it will likely be necessary to create two different mappings, A-to-B as well as B-to-A.
- An indication of which fields in the target scheme are required, or should be required, and determine if the source data will support that requirement. If not, consider programmatic strategies to enter real data or placeholder data during transformation.
- An indication of other data to be entered programmatically. For instance, there could be data that were not originally entered into the source system, because from the perspective of the source it was self-evident (from within a painting collection it

might seem needless to enter "painting" as a term because it would apply to everything). From the context of a broader integrated search, that information could be vital. This is also a place where more specific vocabulary used to catalog the type and genre of painting might need to be converted to a broader and more generic term such as simply "painting."

- An indication of what to exclude. It's important to consider the objectives to be achieved by the transformation and perhaps exclude any highly domain-specific data that may hamper the usefulness of the resulting records. If the single search strategy is for the purpose of resource discovery, then there may not be a need to include the complete cataloging record.

- Instructions for transforming data that live in multiple fields in the source records and must be concatenated into a single field in the target record.

Plan on documenting and maintaining the crosswalks. As automation opportunities arise in the future, it will be crucial to refer to the crosswalks and associated documentation for maximum effect. There are possibilities for using structural standards such as METS to package and manage multiple crosswalks in a common way, and as a standard way to share, and deliver crosswalks as well as communicate mapping strategies with automated systems.

Well thought-out metadata mapping, and well-documented crosswalks will not only meet the immediate needs of present single search strategies, but will also act as part of the foundation for newer information environments such as the Semantic Web.

Access Considerations

Most of the efforts involved in cataloging, data standardization, digitization and indexing are invested for the purpose of discovery, so extra attention should be devoted to making sure the users can find and use it.

Web presence is the number one most important element when it comes to public access to LAM collections. Most users prefer searching for information from their computers via the Internet, and all major organizations have extensive Web sites. The questions are:

- How integral are the library catalog, digital library database and electronic resources to the main Web site of an organization?
- Is there a one-stop-searching center for all resources?
- What is the level of access to materials—how much immediate online access to images, articles, videos, and sound recordings, etc. are available to end users?

Web design is a full-fledged discipline of its own and professional designers should be hired or consulted when possible; at the very least, design principles ought to be studied. That being said, there are additional single search factors and considerations that will impact access.

The position of a single search center within a Web site often makes a difference. Some Web sites choose to put their single search mechanism front and center in the site, while others position single search at deeper levels. Certainly, this has much to with the priorities of the organization; however, making the single search center more prominent than just a simple link from a page is generally a better solution. When a Web site includes an integrated search box on multiple pages, placement of the search box should be prominent on every page including the search result page. This allows users to enter new searches at any time.

This new discovery tool can be further enhanced by incorporating browsing and filtering tools. Data from existing authority controlled fields such as names, places, subjects, and material types are perfect candidates to help users jump into the system without being intimidated by the empty search box. With proper system planning and implementation, these standardized terms can become context-sensitive guides that help users browse and navigate through their search results. These terms can also be used as filtering tools to include or exclude records

for more focused search results. Browsing and filtering is often implemented in a faceted searching environment where a Google-like keyword search creates the basic search result and the faceted term categories guide users to a more refined result. In this type of combined searching and filtering environment, the "advanced searching" can be de-emphasized or eliminated.

Digital Objects

The search result should clearly indicate whether digital objects are available immediately. With digital library databases being created at many organizations, immediate access to online material is frequently possible. While complete citations, inventory lists and catalog records are useful, users want access to the "real stuff" such as online images of photographs, text of full articles, sound file of interviews, digital surrogates of objects. How these digital objects are displayed is likely to be the topic of extended conversations within an institution.

Particularly when utilizing an indexing methodology for single search, there is also the question of what the index points back to—will it bring up a thumbnail of the digital object right away, or will it point back to the native record, which will subsequently lead to the digital object? Individual repositories may continue to be viewed as "best of breed". When this is the case, a success factor may be the ability to link back from search results and generate traffic to the native site, which will allow users to continue their search from there. Other implementers may decide to turn off native interfaces or make them less easy to find. From the user's perspective this may be irrelevant as long as users are able to get what they want.

Rights Management

"Clean" rights information lets users understand what they can do with a digital asset. It also allows the institution to manage legal, financial and reputational risk that can arise from misuse of digital assets. Two survey respondents noted that the single search project at their organization had precipitated a review of their rights management policy and processes.

The value of investing in clean rights management data is manifold. Organizations are increasingly repurposing data, publishing it across different platforms and sharing metadata and digital assets with third parties. This requires institutions to develop clear policies for accessing collections information online; to define when content should be free to access, to determine if charging models should apply and to consider what resolution and proportion of assets should be made available. In Europe, European Commission funding requires that collections information and associated digital assets be made freely available through the Europeana portal. This portal is intended to provide single search access to the collections

housed in Europe's cultural organizations. Additional protection can be achieved by embedding licensing terms in the digital assets, using models such as Creative Commons, a global licensing framework.

User Feedback

Although formal user experience testing had not been undertaken by most of the sites represented by the working group, there were informal studies and casual feedback. In one instance where user testing had been done with an initial prototype, a great deal was learned about the difficulty of representing a faceted system, and they learned that the prototype was not generally well understood. It is difficult to get user feedback from realistic testing without having a system in place that is sufficiently developed and populated, so it is sometimes hard to determine when to begin testing. Since the whole point of all the effort that goes into single search is to increase and improve user access to LAM materials, it is of paramount importance to learn if the system is successful. Whereas one can conduct user testing to see if certain tasks can be completed certain features understood, there are other ways to measure success. Web analytics and other measures can help answer key questions. We should know whether single search attracts new users to the site. We should ascertain whether collections are getting more online use and whether in-person use is increased due to the increased visibility.

Parting Words

We have seen that while LAMs generally share compatible goals and are often hosted under the umbrellas of a single cultural heritage or academic organization, they are indeed very distinct entities, traditionally divided by their diverse histories, professional practices and user communities. Descriptive practices, in particular, constitute an important divide among these institutions. Through the description of their collections, LAMs not only promote the preservation of the cultural past, but they also inform it.

In the era of global search engines, however, users are often puzzled by the realization that they can search the Internet through a single interface, yet the resources of university campuses and other institutions are often compartmentalized in a plethora of informational silos, each with its own dedicated system, search categories and user interfaces. This is particularly evident in the segregation of records relating material and visual cultures in museum collections from the textual and pictorial holdings in libraries and archives. The result of this segregation is to place the burden of discovery on individual users, who may or may not be methodologically and technologically equipped to conduct searches in multiple information repositories. The multiple archival, library and museum collections in a single institution, such as a university campus, are also segmented. The challenges inherent in this informational divide ultimately expect researchers to compartmentalize their interests in a similar manner, rather than encouraging more multi-disciplinary approaches that focus on the research inquiry (rather than the nature and custody of the resources).

The apparent differences might be insurmountable if not for the fact that all these LAMs share one value in common: their clear desire to serve all audiences better by simplifying and empowering search, resource discovery and delivery. The desire to make collections more easily accessible to a wider audience is typified by a 2007 study of researchers in the United Kingdom, which concluded that "[w]hat researchers need above all is online access to the records in museum and collection databases to be provided as quickly as possible, whatever the perceived imperfections or gaps in the records. This is an essential first step towards improving discovery services that will benefit researchers as well as other users" (RIN 2008).

Many institutions large and small are starting to ask questions about how to optimally present their audiences with the breadth of resources locally available. Nine institutions participated

in distilling the essential aspects of single search projects. This report set out to lay the foundation for single search implementers so they can learn from our experiences. Hopefully it will inspire them to contribute to the community discussion by sharing their insights, strategies and approaches. We all found the sharing of experiences to be very helpful and hope that others do, too.

References

OCLC Research. 2009. "Library, Archive and Museum Collaboration." Last modified 11 August. http://www.oclc.org/research/activities/lamsurvey/.

RIN (Research Information Network) 2008. "Discovering Physical Objects: Meeting Researchers' Needs" London, UK: RIN. Last modified 30 August 2008. http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/discovering-physical-objects-meeting-researchers-.

Rinehart, Richard. 2003. "MOAC—A Report on Integrating Museum and Archive Access in the Online Archive of California." *D-Lib Magazine* 9(1). http://www.dlib.org/dlib/january03/rinehart/01rinehart.html

Zorich, Diane M., Günter Waibel, and Ricky Erway. 2008. *Beyond the Silos of the LAMs: Collaboration Among Libraries, Archives and Museums*. Dublin, OH: OCLC Research. http://www.oclc.org/research/publications/library/2008/2008-05.pdf.