Cross-collection Searching: A Pandora's Box or the Holy Grail?

Susan Schreibman, Jennifer O'Brien Roper and Gretchen Gueguen
University of Maryland Libraries, University of Maryland, College Park, MD, USA

Abstract

As digital libraries have expanded to absorb existing collections as well as to create new ones, it has become clear that cross collection discovery is not simply desirable, but is increasingly a necessity demanded by users. Similarly, in the digital humanities community, thematic research collections once distinct from one another now would seem to benefit from interoperability. However, efforts to aggregate disparate resources are often stymied by differing metadata schema and controlled vocabulary. Using the lessons learned from the Thomas MacGreevy Archive, The University of Maryland Libraries designed its digital repository to provide for discovery across object types and collections using Fedora as the underlying architecture. To facilitate access to multiple collections within one repository, University of Maryland developed a flexible metadata standard. This metadata schema is used to describe varying types of materials at varying levels of granularity, while allowing for controlled vocabularies appropriate to specific collections.

Correspondence:
Susan Schreibman, Assistant Dean, University of Maryland Libraries, University of Maryland, College Park, MD, USA.
E-mail:
sschreib@umd.edu

1 Background

During the 1990s, the majority of digital collections developed in digital humanities centers were conceived as stand-alone entities with their own look and feel, search and browse interfaces, controlled vocabularies, and technical underpinnings. This approach created digital silos: autonomous repositories that John Unsworth termed in 2000 'Thematic Research Collections'. These collections were typically developed by scholars to support open-ended research based on aggregating collections of digital primary resources. Throughout the 1990s, there were relatively few of these collections, with technical and theoretical objectives that made them seem (at least at the time) incompatible with other collections that shared the same discipline, time period, or culture.

Digital Library initiatives, on the other hand, often made some attempt to aggregate collections, particularly when these collections focused on a single broad theme, such as *Documenting the American South* (University of North Carolina)[1] or *Historic* Pittsburgh (University of Pittsburgh).[2] But just as frequently, digital libraries also followed a thematic collection model. Although these collections typically resided within the same architecture, interoperability was not necessarily a feature.

Sometimes cross collection search was possible within a single repository, but only within a single media type. This was typically due to a design feature of the repository framework—for

doi:10.1093/llc/fqm039 Advance Access Published on 10 December 2007

example, XPAT, University of Michigan's Digital Library Repository, was designed to search within object types, but not across them. Sometimes a repository's choice of encoding standard, HTML particularly, became a barrier to cross-collection searchability, except at the level of the text itself retrieved via a Google-like search. For collections created in SGML, and later XML, controlled vocabularies were often developed to fulfill research goals and were specifically tailored to the domain.

Other projects took a different tack by utilizing existing controlled vocabularies, such as *Documenting the American South* that utilized Library of Congress Subject Headings (LCSH). Since this project was developed within the library, it had the advantage of local expertise in applying the complex structure. Moreover, it is a vocabulary that is familiar to many academic library users. In addition to this environmental familiarity, LCSH intentionally covers a broad range of subjects and is therefore applicable to a large variety of projects. This is not to say, however, that this approach is without limitations.

LCSH is a precoordinated classification scheme, which means that, where possible, multiple concepts are strung together to create one longer subject heading that encapsulates the topic of an object. For example, a work on a specific campaign during the Gallic Wars would have a heading 'Gaul – History – Gallic Wars, 58-51 B.C. – Campaigns'. The result is a single heading in which the various concepts have been brought together to summarize the work. This approach to subject analysis has long been in use in the library community as a way to direct users towards materials in a precise fashion. The drawback to precoordinated subject headings is that it may be difficult for users who are not searching on a very precise subject, or for users who are unfamiliar with the entire vocabulary that has been coordinated: in such cases, postcoordinated subject headings may be more useful.

Postcoordinated headings place each concept in a separate subject heading. In the example above, a postcoordinated version would have separate headings each for *Gaul*, *History*, *Gallic Wars, 58-51 B.C.*, and *Campaigns*. Marcia Bates has discussed the usefulness of postcoordinated headings in Boolean search environments, noting that concepts can be indexed separately and then brought together by the user looking for combinations, rather than determining combinations ahead of time. This approach, however, also has limitations, as it is more difficult for users to discern the relationships between the subject terms. In the precoordinated version, *Gaul* and *History* were linked in a way that identified the subject as being a history of Gaul, whereas in the postcoordinated version it is possible for the terms to appear as subjects without representing that this is a work about Gallic history.

By the early 2000s, several fundamental shifts were happening in the digital library community. As digital library programs became larger and more established, the need for repository architectures that could accommodate multiple thematic research collections, or a larger more general collection of materials from multiple time periods across multiple formats became more of a necessity than a desire. Digital collections became less an 'exhibition' or 'highlight' on the library's web site and more an integral component of digital library services. Collaborations between digital humanities projects and libraries resulted in an increase in permanent curatorial responsibilities by libraries for thematic research collections. As digital libraries were absorbing existing collections as well as creating new ones, it was becoming clear that facilitating cross-collection discovery was not simply desirable, but being demanded by users. A similar realization was happening in the digital humanities community as thematic research collections which seemed in the 1990s distinct from one another, now would seem to benefit from interoperability.

One of the most prominent initiatives in the library community to address federated searching, the Open Archives Initiative (OAI) Protocol for Metadata Harvesting[3], dealt with this issue in designing a methodical way to search metadata composed in various schema through a single search tool. The OAI protocol grew out of the e-print server community to normalize both the type of metadata used to describe collections and the method by which the data was gathered and stored. This initiative particularly rejected the method advanced in the earlier Z39.50 protocol, a complex protocol to allow the searching of remote servers, and created a system of harvesting or gathering normalized metadata periodically from disparate repositories into a federated search service. Participating repositories make their metadata available in unqualified Dublin Core so that all metadata in the resulting federated repository uses the same sixteen fields. When a user initiates a search and a matching record is found in the federated repository, the user is redirected to the original repository where richer metadata may be found. This switch was made to reassign the workload away from distributed repositories to a single repository that would be consistently organized resulting in less-complicated searching methods and stronger centralized functioning. It also allows individual institutions to have a rich local metadata scheme while providing a normalized version for federated searching.

An extremely successful implementation of the OAI protocol is *The Sheet Music Consortium*[4], a joint project of UCLA, Indiana University, Johns Hopkins University, and Duke University, which provides a single portal to search across collections while providing users with the ability to easily locate an item belonging to a particular institution. The consortium harvests metadata from each institution's catalog using the OAI protocol while providing a framework to integrate, search, and deliver the harvested metadata and the associated sheet music images.

The OAI protocol offers some valuable lessons on the macro-level to the institution working towards interlinking their own micro-universe. Normalization of content structures is an obvious benefit in terms of faster, easier searching; centralizing functions to redirect traffic out to different areas is another promising model. While this method allows federated services to search through records faster, it also allows for the creation of different levels of metadata. The OAI protocol, however, is not without its drawbacks. One large problem is the lack of content standards. If no single, controlled vocabulary is used consistently, there is no guarantee that users will find all relevant items without a considerable effort to expand queries to include synonyms. While separate controlled vocabularies may work very well at the local level, when combining repositories that level of local specificity can cause problems. Although each smaller collection may have used a standard controlled vocabulary, these vocabularies may not be interoperable.

For example, two authoritative sources to describe art, The Getty Art and Architecture Thesaurus (AAT) and the Library of Congress Thesaurus of Graphic Materials (TGM), are used to describe both technical details of art works as well as the subject matter within them. But, while TGM indexes the plant 'Cat tails', AAT recognizes 'cattail'. Were two collections dealing with art, one using TGM and one using AAT, to combine their data, any search on this or any other dissimilar term would yield incomplete results to an unwitting user. There is no easy solution to this problem and it is one that has plagued cross-collection initiatives. Certainly, this problem exists in any catalog or metadata repository to some degree, but with a distributed group of record creators who do not interact, vocabularies abound.

Legacy projects with metadata created at various levels of granularity and with different data types (XML, HTML, relational databases) in addition to various content standards show problems similar to OAI protocol implementation. They often suffer from the proliferation of metadata schema and controlled vocabulary that make them individually so unique. Attempts to make these projects interoperable, such as the NINES project[5] (A Networked Interface for Nineteenth-Century Scholarship), often struggle. One of the goals of NINES is to provide an interface in which a federated search can be carried out across a number of collections. Unlike *The Sheet Music Consortium* that harvests metadata from MARC catalogue records that have a long history of interoperability within the library community, NINES found itself harvesting from a variety of metadata, frequently implemented in non-uniform ways. Moreover, at least one of the NINES projects is still in HTML with little metadata that is amiable to automated harvesting. So, in this instance, both the irregularity of metadata schema and the absence of a single, controlled vocabulary have made this task extremely challenging.

2 Development of University of Maryland Digital Collections (UMDC)[6]

With these protocols, projects, and consortia in mind, The University of Maryland Libraries began the design of a digital repository that provides for the discovery across object types both within a single thematic research collection as well as across them. After investigating several different digital library systems, it was decided to adopt Fedora[7] as the underlying architecture. Fedora was attractive for a number of reasons. It is an extremely flexible open source software that provides a service-oriented architecture for managing and delivering digital content. Its digital object model supports multiple views of digital objects, as well as the relationships among objects. It is also extremely scalable. At the time of this writing, the National Science Digital Library (http://nsdl.org/) has ingested and made available over 4.7 million objects. Moreover, dynamic views of content are possible by associating web services with objects. It is XML based, and supports multiple metadata schemes and searches across object types.

Additionally, Fedora was attractive because it is possible to configure the repository so that unique collections could retain their own look and feel and be searched individually while allowing searching across all collections. This was important so that users could discover objects outside a collection context. This necessitated forethought in not only designing the digital repository architecture, but the metadata, allowing for different levels of subject descriptors to be developed and stored: one that is collection specific and one that provides for cross-collection discovery.

3 Creation of a Local Standard

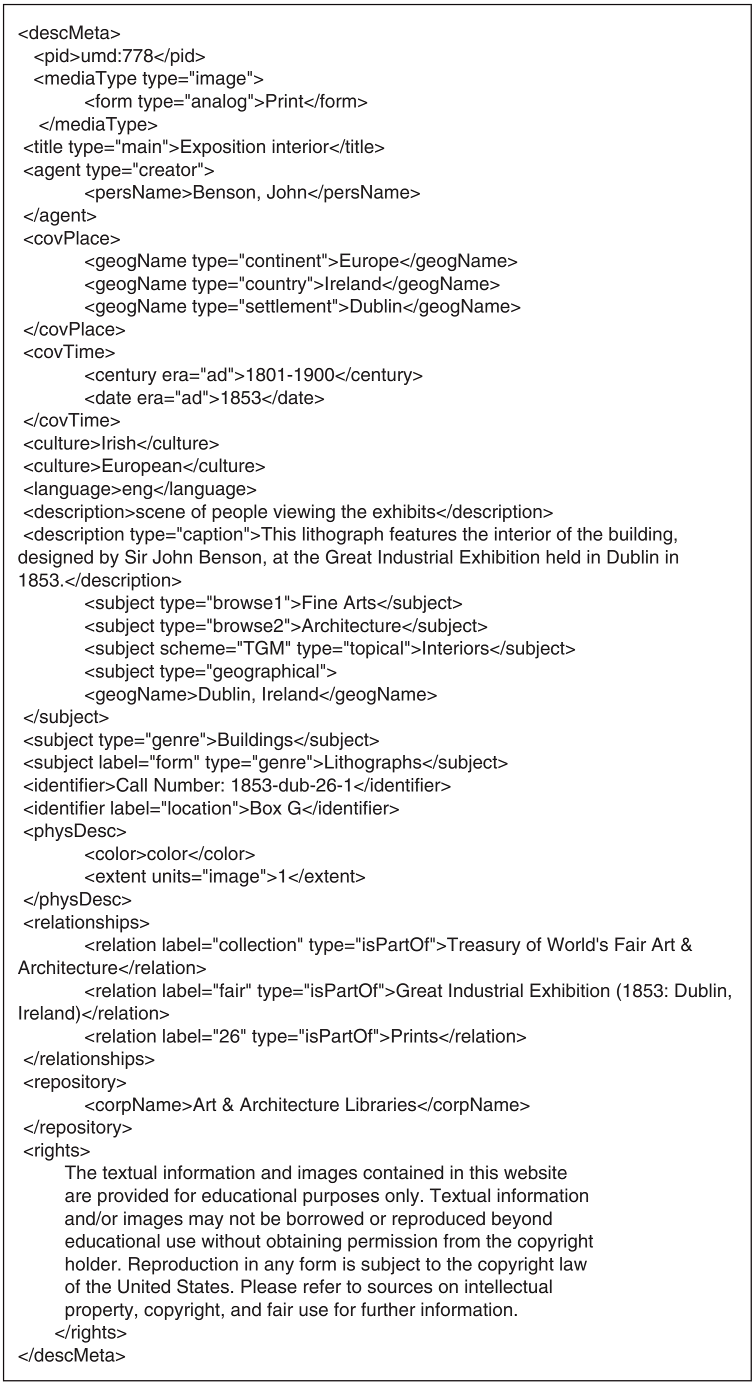
It was understood that the selection of metadata standard(s) would be a key element in facilitating both collection specific and cross-collection searching. Dublin Core and Visual Resources Association (VRA) Core were closely analyzed for suitability. As noted earlier, using the most basic of descriptors, Dublin Core does not provide the rich description of an object usually desired in a thematic collection. VRA Core, on the other hand, excelled at providing a truly rich description of digital image objects, but it is intended to solely describe still images.

In discussions of the strengths and weaknesses of these major descriptive standards, the idea of blending the two emerged. The University of Virginia had already created such a hybrid schema,[8] and this localized standard was tested for usefulness in the University of Maryland setting. The UVA standard contained most of the elements from Dublin Core and VRA. The testing proved positive.

The MODS (Metadata Object Description Schema) was also investigated. While MODS allows for the capture of the same types of information that UMDM captures, it also limits some options for metadata design built into the UMDM. For instance, the UMDM elements <culture> and <style> are easily placed within the MODS <genre> element, with type qualifiers to separately identify their purposes. However, making these elements qualifiers on a larger, more general element, limits the opportunity to further qualify the information they express. For instance, with <style> represented as a qualifier on a more general <genre> element, it is difficult, if not impossible, to further subdivide the 'style' data into logical sub-genres such as 'architecture', 'literary', and 'music' while retaining the larger 'style' category. This subdivision of style could prove useful to creating a tool to help users looking for a particular type of 'style'. The ability to provide this further level of granularity argues for the value of having a local standard, as local desires will invariably exceed what a generalized standard can provide.

Thus the UVA standard was customized into a descriptive standard meeting local needs and specifications, the University of Maryland Descriptive Metadata (UMDM; Fig. 1). At the same time, the UMDM standard can also be seen as a highly customized local application profile of MODS, as any UMDM record can easily be mapped into a MODS record, and thus conform to a national standard for collaboration outside of the institution.

In customizing the DTD, decisions regarding minimum descriptive standards were also made.



```
<descMeta>
  <pid>umd:778</pid>
  <mediaType type="image">
          <form type="analog">Print</form>
   </mediaType>
 <title type="main">Exposition interior</title>
 <agent type="creator">
          <persName>Benson, John</persName>
 </agent>
 <covPlace>
          <geogName type="continent">Europe</geogName>
          <geogName type="country">Ireland</geogName>
          <geogName type="settlement">Dublin</geogName>
 </covPlace>
 <covTime>
          <century era="ad">1801-1900</century>
          <date era="ad">1853</date>
 </covTime>
 <culture>Irish</culture>
 <culture>European</culture>
 <language>eng</language>
 <description>scene of people viewing the exhibits</description>
 <description type="caption">This lithograph features the interior of the building,
designed by Sir John Benson, at the Great Industrial Exhibition held in Dublin in
1853.</description>
          <subject type="browse1">Fine Arts</subject>
          <subject type="browse2">Architecture</subject>
          <subject scheme="TGM" type="topical">Interiors</subject>
          <subject type="geographical">
          <geogName>Dublin, Ireland</geogName>
 </subject>
 <subject type="genre">Buildings</subject>
 <subject label="form" type="genre">Lithographs</subject>
 <identifier>Call Number: 1853-dub-26-1</identifier>
 <identifier label="location">Box G</identifier>
 <physDesc>
          <color>color</color>
          <extent units="image">1</extent>
 </physDesc>
 <relationships>
          <relation label="collection" type="isPartOf">Treasury of World's Fair Art &
Architecture</relation>
          <relation label="fair" type="isPartOf">Great Industrial Exhibition (1853: Dublin,
Ireland)</relation>
          <relation label="26" type="isPartOf">Prints</relation>
 </relationships>
 <repository>
          <corpName>Art & Architecture Libraries</corpName>
 </repository>
 <rights>
       The textual information and images contained in this website
       are provided for educational purposes only. Textual information
       and/or images may not be borrowed or reproduced beyond
       educational use without obtaining permission from the copyright
       holder. Reproduction in any form is subject to the copyright law
       of the United States. Please refer to sources on intellectual
       property, copyright, and fair use for further information.
      </rights>
</descMeta>
```

Fig. 1 A sample record encoded in UMDM

Productive cross-collection searching could only be achieved with a rich set of data, and so the minimum descriptive standards were set very high. In addition to the standard media type, title, creator/provider/contributor, physical description, repository, and subject elements, the UMDM for every object must contain information at least about the century and continent of creation, general subjects to facilitate the browse, and any useful relationships. With twelve required base elements, some of which further require subelements, the system would be populated with full descriptions on a variety of levels for each object.

4 Use of METS

The Fedora repository used by the University of Maryland relies on the UMDM metadata standard for all digital objects, but in fact, can and does support other metadata schemas. The UMAM (University of Maryland Administrative Metadata Scheme) contains administrative metadata about each digital object. The repository also uses as its roadmap XML files created according to the Metadata Encoding Transmission Standard[9] (METS). A standard created by the Library of Congress and the Digital Library Federation, METS records act like a framework to hold and structure different types of metadata about digital objects. Records are made up of seven components: the METS header, Descriptive Metadata, Administrative Metadata, File Section, Structural Map, Structural Links, and Behavior.

While the UMDM and UMAM provide detailed information regarding digital objects, the METS record goes a step further and integrates specific file information and the composition of those files within a digital object. For this reason alone it is a valuable tool when building a digital repository. But what has been possibly more valuable to the creation of a cross-searchable repository is the use of external descriptive and administrative metadata and the behavioral control provided by the Behavior section.

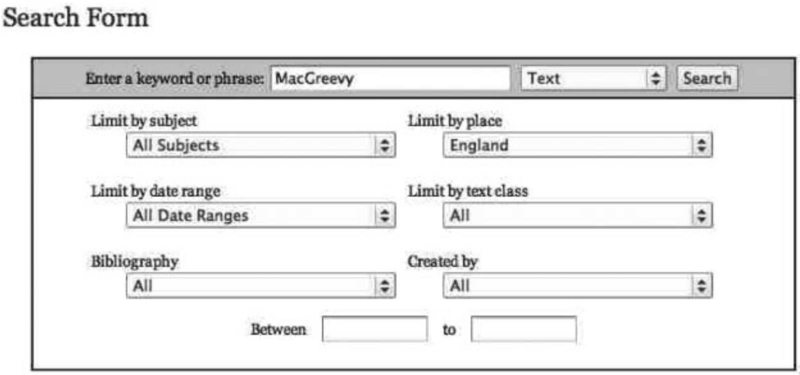
With these tools it is possible to ingest a single TEI file into the repository and use the behavior to carry out a mapping into the UMDM record based on information in the file. This workflow allows authors of TEI files to use appropriate software to write and validate their TEI document against a DTD or schema outside the repository. Following the conventions of the mapping that was created for the project, a UMDM record is generated on the fly when the object is called as a result of a user query. The same case is true for EAD or other full-text encoding schemes.

In the case of other legacy metadata record schemes, like VRA Core or Dublin Core, as long as a mapping onto the UMDM record is created and stored, the METS record can handle the translation from other schemes. MARC or VRA records could be ingested for use only in a specialized research-specific interface, while a mapping of those fields into UMDM is used for cross-collection searching. Just as earlier when the MHP was discussed, a basic metadata scheme is dictated and as long as that minimal requirement is met, extra information can be handled in the specific schema.

5 Mapping to Multiple Standards

When any new project is integrated into the University of Maryland's digital repository, existing metadata is evaluated and mapped onto the UMDM standard. A large portion of the work in metadata design for any new collection is the creation of mapping documents that are used by the programmer to script existing metadata into UMDM records. Such mapping documents usually take the form of tables with specific instructions about the location of information in both the old and new records, any transformation notes such as alterations in punctuation or the combination or separation of files, and the creation of any static fields or phrases.

Once data has been mapped records can be ingested in multiple ways: records can be ingested in one format like TEI or EAD and UMDM records can be dynamically generated when needed. Data can also be scripted from a database into XML records and parsed against the UMDM DTD. A third option is to add the record through a web form in the repository's administrative interface. If a full-text record is being added the process is



Search Form

Enter a keyword or phrase: MacGreevy
Text

Limit by subject: All Subjects
Limit by place: England
Limit by date range: All Date Ranges
Limit by text class: All
Bibliography: All
Created by: All
Between ___ to ___

Fig. 2 The first page from the UMDM object upload form

as simple as choosing the XML file to attach and submitting the record. When separate objects and metadata records are being added (for images, audio, and moving images) with the online form, the user is prompted to input specific metadata values that match the fields of the UMDM record and is provided controlled vocabulary choices through drop-down menus and other options (Fig. 2). As records for individual collections may have specific attribute values or input standards, optional fields are available when items are added to particular collections. Once the metadata form is filled up, objects are attached and the entire package is submitted.

6 Controlled Vocabularies

Even the most robust standard, however, is only as useful as the input data. Except in fields such as <title> or <description>, efforts are made to regularize vocabulary and ensure that like objects can be consistently found, both within a thematic collection and across them. Subject terms receive particular attention, as they are important in describing objects richly at the collection level, as well as critical to guiding users to objects held in a variety of collections. Thus, objects can be described using a number of subject terms with these two different purposes in mind. For example, a lithograph depicting the interior of a Worlds Fair exhibit hall and a closing awards ceremony occurring within can receive specific headings such as ‘Awards ceremonies’ and ‘Closing ceremonies’ in addition to broader terms such as Fine Arts and Architecture.

At the outset of each project, an appropriate controlled vocabulary is selected that will describe the objects in the manner best suiting the materials as well as the purpose of the project. In addition to these collection-determined vocabularies is a second set of subject terms, determined independent of any project and designed to promote cross-collection discovery, although the exact nature of this collection-independent vocabulary has yet to be determined. Initial experiments with a vocabulary based on the Library of Congress Classification System (LCCS) were only moderately successful, as this vocabulary worked well for some objects and collections, but it was difficult to apply in specialized cases. The current direction is to alter or supplement the LCCS-based terminology so that it remains broad in focus while allowing for users to easily navigate towards more specific content.

Around the same time these decisions were being made for the University of Maryland Libraries digital repository, the opportunity arose to see if these same principles could be applied to a legacy thematic research collection that was growing in ways that made it difficult to apply the original metadata scheme to new collections.

7 The Thomas MacGreevy Archive: A Thematic Research Collection with a Repository Attitude

In 2004, *The Thomas MacGreevy Archive,*[10] a TEI-based thematic research collection was undergoing significant expansion. When the project was first conceived in 1996, the only document type envisioned for the collection was previously published books and articles. Nevertheless, over the years, the project expanded to include a name database (*Who's Who in the MacGreevy Archive)* initially developed for internal use to provide authority control, and *Thomas MacGreevy Writing a Poem*, a selection of poems available in multiple versions through *The Versioning Machine.*[11] Although these collections are available through the browse page, content in them is not discoverable from the project search page. For objects to be searchable, they must be TEI documents which conform to *The MacGreevy Archive* DTD and utilize a controlled vocabulary as described subsequently. In 2004, two unrelated events revealed how some of our design decisions precluded the integration of other data types: a microcosm of the XPAT architecture issues mentioned previously.

The MacGreevy Archive was conceived as a text-centric repository. If images were available through the site, they were available only though HTML pages as part of the site wrapper, or discoverable via a link from within a TEI-encoded text. Since MacGreevy was an art critic, it did not take long for there to be almost 100 images available through the archive, many of them unavailable anywhere else on the web. It was thus decided that users might want to view them independently of the articles, or indeed, might want to view an image and then see what texts it was associated with.

The second event that put a strain on the site architecture was the decision to include two collections of unpublished letters. These collections were conceived as scholarly editions with several levels of apparatus. Unlike the design decision taken with articles (which are represented by a TEI-encoded text transformed to HTML for display), it was decided that including images of the manuscripts was crucial so that a reader could view a diplomatic version of the letter next to a facsimile image. These new letter collections, intimate, gossipy, and informal, also put a strain on the controlled vocabulary that was developed for a more formal, public writing style. And lastly, many of the individuals and organizations mentioned in the letters are relatively obscure, and are mentioned repeatedly: thus having a seamless link between the name database and the letters was crucial.

A University of Maryland Libraries Faculty Research Grant enabled the authors of this article to retrofit *The MacGreevy Archive*, making it less like a thematic research collection, and more like a multi-faceted repository.[12] Armed with the knowledge gained from developing a University of Maryland Libraries Digital Library Architecture, we investigated how, without significantly altering the architecture currently in place (which was beyond the scope and resources of the current project), interoperability between the various components of the collection could be enhanced.

8 Development of an Enhanced Ontology

The first step in making the entire archive searchable through a single search page was a review of the keywords currently being used to describe articles, and a determination of how to apply them to the new collections of letters and images. The original keyword scheme for the *Archive* was based on the ten major classes of the Dewey Decimal System. It described the collection in simple broad strokes, as subjects in this relatively small collection were relatively wide-ranging, with no large concentrations of materials in any one specific subject area. Dewey class terms that did not correspond to anything in the collection (technology, for example) were dropped, while others that had some significant depth, 'the arts', for example, were expanded to 'ballet', 'opera', 'music', 'fine art', etc.

The keywords had been reviewed in the summer of 2004 when the idea of creating a letter-based collection was originally proposed. At the time when cross-searching was not considered a priority and an ontology was suggested that was uniquely tailored to each letter collection. Over the course of time, however, it became clear that a cross-searchable repository would be more useful to users than individually searchable subcollections. These new keywords became less useful as they hindered cross-searching and the ontology itself was extremely specific—i.e. terms sometimes applied to only a few letters and could not be combined effectively.

Next an approach was considered based on the subject control currently being implemented for the library-wide Fedora repository, wherein all items would receive keywords from a broad control group as well as more specific terms chosen from specialized vocabularies chosen on a project-by-project basis. The original MacGreevy keywords were to be considered as the broader group, and the unique ontology created in 2004 was to be modified and used as the project-specific list. This approach also did not work as many of the letters and images did not correspond to any original keywords.

Finally, a single set of keywords based on the original MacGreevy list, with some modifications based on terms developed specifically for letters, was created. For the most part, the new terms – 'Biography', 'Career and Finances', 'Domestic Life', 'Irish Culture', 'Politics and Government', and 'Social Life'—would apply only to letters and images. The appropriate terms from this group, plus any applicable existing ones would be assigned to new documents. The terms 'Irish Culture' and 'Politics and Government' proved slightly problematic as there was a high probability that articles already existing in the archive would fall within the definitions of these terms. In the case of these latter two terms, it was necessary to revisit the archive and apply these terms where applicable. While this is not a desired way of proceeding, the small number of terms to consider (only two) and the moderate size of the existing archive (around 450 documents) made the task manageable. Indeed, many of the articles could be ruled out without close examination due to the General Editor's knowledge of the subject matter.

9 Redevelopment of the Search Page

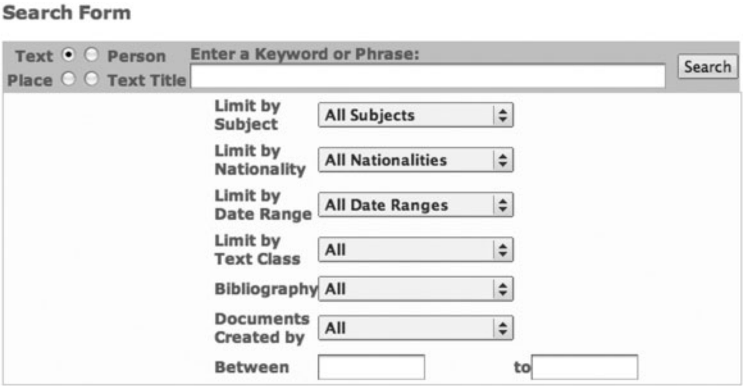
To accommodate the new cross-searching capabilities, the search page needed modification. The previous design featured an optional free-text search within the body of the TEI document, the <title> element, and persons or places encoded within <persName> and <placeName> elements. In addition, searches could be limited by a number of other qualifiers, all related to the subject of the article.

These qualifiers provided a faceted approach to a specific and complex subject matter, but were often confusing to users. Thus users could combine keywords based on the subject of the text (Art, Literature, History, etc.), nationality (which refers to the focus of the article such as 'French Literature' or 'Italian Art'), date range (indicated by the century of the article's focus), text class (a reference to the type of article, such as an art or book review, an obituary, letter to the editor, review, etc.), and its place within a bibliography of works written by or about MacGreevy.

As it stood, this search interface presented problems from a usability standpoint. When asked informally, many users noted that the options available for use with the free-text search box were not seen or understood. In addition, the limiting categories and terms themselves were often confusing. Terms like 'bibliography' and 'text class' were often not understood at all, while topics like 'date range' and 'nationality' could be confused for the time and place of writing rather than the subject of the text. A redesign of the search page would not only facilitate access to multiple collections, but would be able to address these usability issues without losing any of the sophisticated subject access functionality that the *Archive* search page already employed.

The newly designed search interface (Fig. 3) features a more prominent placing of the free-text searching options as well as a clearer labeling of the different aspects of subject description:

For example, a search that was previously labeled 'Nationality' is now referred to as 'Find documents about a place'. Changes to the search interface also


Search Form

Text ● ○ Person
Place ○ ○ Text Title

Enter a Keyword or Phrase:

Limit by Subject: All Subjects
Limit by Nationality: All Nationalities
Limit by Date Range: All Date Ranges
Limit by Text Class: All
Bibliography: All
Documents Created by: All
Between to

Fig. 3 The redesigned MacGreevy search page

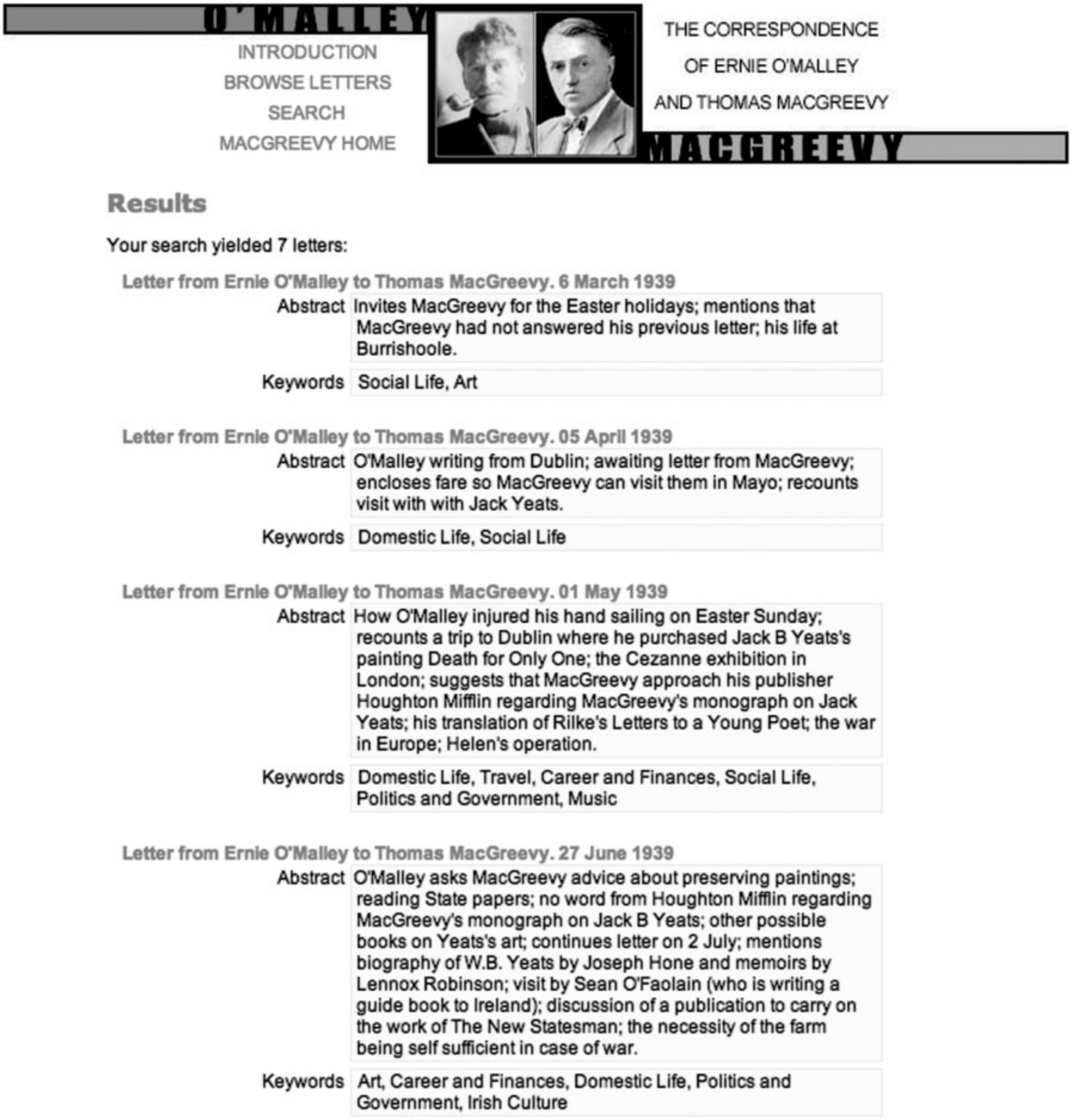
allowed greater access to subcollections and groupings of documents. Documents are classified as belonging to a large document class (text versus image), a document-type subclass (photograph versus painting versus illustration versus manuscript in the case of images), as well as thematic collection (*The Correspondence of Thomas MacGreeevy and Ernie O'Malley*, for example; Fig. 4).

From the beginning of the project, it was known that the display of initial results for a collection of letters would be inadequate given the current paradigm of *The MacGreevy Archive*, wherein documents that match the search query are displayed by title and author. A similar display of letters would result in a many hits of the same title—'Letter from Thomas MacGreevy to George Yeats, [date]'—distinguished only by a difference in the date of the letter's composition. Unless the searcher had extensive knowledge of the author's and recipient's lives, as well as the social and political issues that might be of interest at a particular time, the short display results would prove cumbersome as it provides little differentiation from which to select a letter for further investigation. As a remedy, a decision was made early on to write brief abstracts for each letter which would be displayed along with the title, date, and author in the first result page. Although these abstracts do not aid in retrieval, they will assist users with relevance decisions. For collections of letters with less scholarly apparatus, displaying keywords would provide some measure of disambiguation between search results.

Since images were not discoverable as objects, they were never listed on the first result page. Now that they would be independently discoverable, similar issues of initial results and relevancy determination would have to be made. Metadata limited to the title or caption of an image and the subject keywords along with a thumbnail version of the image were deemed appropriate for first results (Fig. 5).

These thumbnails are considered analogous to the abstract and the blurb included with articles and letters to aid researchers in determining relevance. Although image creator may seem to be an obvious choice for this display, gaps in the earlier metadata records have made this field difficult to capture with any certainty. And in the case of the many family photos, it may not even be possible to determine the creator. When an image is chosen from these results, the more extensive metadata record is displayed alongside a larger version of the image.

Functionally, search was redesigned to include a full-text search searching, which would search through the entire TEI document, instead of a search of only the <text> element. This was particularly important for image metadata as the majority of metadata is contained in the TEI Header. Collections are separately identified via <keyword>, while different classes of texts are distinguished by the value of <classCode>. When returning results, documents are split into categories based on these <keyword> and <classCode> elements. Additionally, individual texts are searched for the presence of <figure> tags that are indicated in the search by the image icon.



Results

Your search yielded 7 letters:

Letter from Ernie O'Malley to Thomas MacGreevy. 6 March 1939

Abstract	Invites MacGreevy for the Easter holidays; mentions that MacGreevy had not answered his previous letter; his life at Burrishoole.
Keywords	Social Life, Art

Letter from Ernie O'Malley to Thomas MacGreevy. 05 April 1939

Abstract	O'Malley writing from Dublin; awaiting letter from MacGreevy; encloses fare so MacGreevy can visit them in Mayo; recounts visit with with Jack Yeats.
Keywords	Domestic Life, Social Life

Letter from Ernie O'Malley to Thomas MacGreevy. 01 May 1939

Abstract	How O'Malley injured his hand sailing on Easter Sunday; recounts a trip to Dublin where he purchased Jack B Yeats's painting Death for Only One; the Cezanne exhibition in London; suggests that MacGreevy approach his publisher Houghton Mifflin regarding MacGreevy's monograph on Jack Yeats; his translation of Rilke's Letters to a Young Poet; the war in Europe; Helen's operation.
Keywords	Domestic Life, Travel, Career and Finances, Social Life, Politics and Government, Music

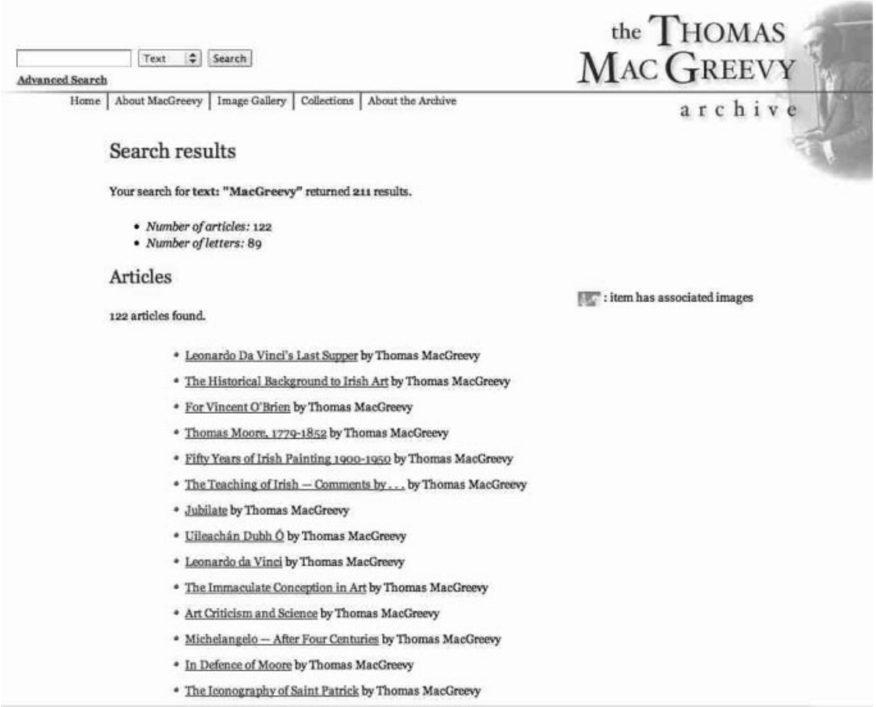
Letter from Ernie O'Malley to Thomas MacGreevy. 27 June 1939

Abstract	O'Malley asks MacGreevy advice about preserving paintings; reading State papers; no word from Houghton Mifflin regarding MacGreevy's monograph on Jack B Yeats; other possible books on Yeats's art; continues letter on 2 July; mentions biography of W.B. Yeats by Joseph Hone and memoirs by Lennox Robinson; visit by Sean O'Faolain (who is writing a guide book to Ireland); discussion of a publication to carry on the work of The New Statesman; the necessity of the farm being self sufficient in case of war.
Keywords	Art, Career and Finances, Domestic Life, Politics and Government, Irish Culture

Fig. 4 First results page for the correspondence between MacGreevy and O'Malley

10 Future Directions

The choice of utilizing TEI as the core metadata scheme for *The MacGreevy Archive* was a philosophic approach employed by many digital humanities projects of the 1990s. It was adopted at a time before the development of XML when integration with other standards and formats was extremely difficult, and there were few choices to display SGML-encoded text on the World Wide Web. Ten years after *The MacGreevy Archive* was founded, the environment is almost unrecognizable: advances


Text
Search
Advanced Search
the THOMAS MACGREEVY archive
Home | About MacGreevy | Image Gallery | Collections | About the Archive

Search results

Your search for text: "MacGreevy" returned 211 results.

- *Number of articles:* 122
- *Number of letters:* 89

Articles

: item has associated images

122 articles found.

- Leonardo Da Vinci's Last Supper by Thomas MacGreevy
- The Historical Background to Irish Art by Thomas MacGreevy
- For Vincent O'Brien by Thomas MacGreevy
- Thomas Moore, 1779-1852 by Thomas MacGreevy
- Fifty Years of Irish Painting 1900-1950 by Thomas MacGreevy
- The Teaching of Irish — Comments by . . . by Thomas MacGreevy
- Jubilate by Thomas MacGreevy
- Uileachán Dubh Ó by Thomas MacGreevy
- Leonardo da Vinci by Thomas MacGreevy
- The Immaculate Conception in Art by Thomas MacGreevy
- Art Criticism and Science by Thomas MacGreevy
- Michelangelo — After Four Centuries by Thomas MacGreevy
- In Defence of Moore by Thomas MacGreevy
- The Iconography of Saint Patrick by Thomas MacGreevy

Fig. 5 A screen shot from the newly designed first results page

in the technologies of the World Wide Web have made integration of metadata and media possible in a way one could only dream about in the early- to mid-1990s.

In 2002, *The MacGreevy Archive* migrated from SGML to XML. The next migration will be into an architecture more in keeping with that developed for University of Maryland Libraries. Although the UM Digital Repository is a work in progress, we have set our sights on the holy grail. Learning from the architecture of digital repositories from the 1990s, the UM Repository utilizes as its core schemes standards not developed to describe any one object type but which recognize and interact with standards like TEI, EAD, VRA Core developed to describe specific content in detail. The controlled vocabularies used adhere to recognized standards, like the LCSH and the Getty Art and Architecture Thesaurus, so as to provide a stable set of terms, but may not necessarily follow the application rules for the chosen vocabulary, such as perhaps providing LCSH terms as postcoordinated rather than precoordinated where appropriate. This approach, which both adheres to standards, but uses them in a flexible way, is the path down which we believe the treasure will be found.

References

Bates, M. J. (2002). The cascade of interactions in the digital library interface. *Information Processing and Management*, **38**(3): 381–400. Republished at http://www.gseis.ucla.edu/faculty/bates/articles/cascade.html (accessed 20 September 2006).

Besser, H. (2004). The Past, Present, and Future of Digital Libraries. In Schreibman, S., Siemens, R., and Unsworth, J. (eds), *A Companion to Digital Humanities*. Oxford: Blackwell Publishing, pp. 557–75.

Chan, L. M. and Lei Zeng, M. (2006a). Metadata Interoperability and Standardization – A Study of Methodology Part I. Achieving Interoperability at the Schema, Level. *D-Lib Magazine*, 12:6. http://www.dlib.org/dlib/june06/chan/06chan.html (accessed 20 September 2006).

Chan, L. M. and Lei Zeng, M. (2006b). Metadata Interoperability and Standardization – A Study of Methodology Part II. Achieving Interoperability at the Record and Repository Levels, *D-Lib Magazine*, 12:6. http://www.dlib.org/dlib/june06/zeng/06zeng.html (accessed 20 September 2006).

Dempsey, L. (2006). The (Digital) Library Environment: Ten Years After, *Ariadne* 46. http://www.ariadne.ac.uk/issue46/dempsey/ (accessed 20 September 2006).

Goldsmith, B. and Knudson F. (2006). Repository Librarian and the Next Crusade: The Search for a Common Standard for Digital Repository Metadata, *D-Lib Magazine*,12:9. http://www.dlib.org/dlib/september06/goldsmith/09goldsmith.html (accessed 20 September 2006).

Nines: A Federated Model for Building Digital Scholarship (2005) http://www.nines.org/about/9swhitepaper.pdf (accessed 20 September 2006).

Plamer, C. L. (2004). Thematic Research Collections. In Schreibman, S., Siemens, R., and Unsworth, J. (eds), *A Companion to Digital Humanities*. Oxford: Blackwell Publishing, pp. 348–65.

Unsworth, J. Thematic Research Collections. Paper presented at the Modern Language Association Annual Conference, December 28, Washington, DC. http://www.iath.virginia.edu/~jmu2m/MLA.00/ (accessed 20 September 2006).

Notes

[1] http://docsouth.unc.edu/

[2] http://digital.library.pitt.edu/pittsburgh/

[3] http://www.openarchives.org/OAI/openarchivesprotocol.html

[4] http://digital.library.ucla.edu/sheetmusic/

[5] http://www.nines.org/

[6] The authors wish to thank their collaborators at University of Maryland Libraries, David Cooper, Sean Daugherty, Paul Hammer, David Kennedy, Jean Phillips, and Ben Wallberg.

[7] http://www.fedora.info/

[8] http://www.lib.virginia.edu/digital/metadata/descriptive.html

[9] http://www.loc.gov/standards/mets/

[10] http://macgreevy.org

[11] http://v-machine.org

[12] The authors would like to thank their collaborators on this project, Sean Daugherty, Amit Kumar, Tony Ross, and Eric White.