Title

Further Development of a Shared Cataloging Resource for the Visual Resources Community: UCAI Phase Two: Final Report

Authors
Barnhart, Linda
Schottlaender, Brian E.C.
Westbrook, Brad
et al.

Publication Date
2006-02-01

Further Development of a Shared Cataloging Resource for the Visual Resources Community: UCAI Phase Two

Final Report to The Andrew W. Mellon Foundation

20 February 2006

Brian E. C. Schottlaender, Principal Investigator
Linda Barnhart, Project Manager
and the UCAI Project Team

CONTENTS

1.0 INTRODUCTION

Problem statement

The visual resources community has voiced an urgent need for a central bibliographic utility through which they could share and re-use metadata records for visual materials. In the absence of such a tool, image catalogers have been forced to independently and redundantly research and create descriptive records, an expensive and potentially wasteful undertaking. The current highly localized approach has resulted in a large number of legacy records, formed by divergent (sometimes conflicting) and idiosyncratic practices. Despite recent progress, the slow implementation of community-based standards and the absence of a common technical platform have made sharing image metadata records extraordinarily difficult.

Phase One achievements

Phase One of the Union Catalog for Art Images (UCAI) project proved that the legacy metadata held by three very diverse and large image collections could be standardized and combined into one prototype database. The team worked on Phase One from April 1, 2002 to December 31, 2003, and a separate final report was prepared covering that period. Key achievements from Phase One included:

- Developing an innovative prototype database
- Developing customized data maps to VRA Core 3.0 in XML from three different dataset structures
- Converting 715,000 records and loading them into the prototype system
- Developing the work unit and composite record concepts
- Articulating data standardization needs
- Beginning initial development of an automated clustering algorithm

Phase Two

Phase Two was designed to advance and stabilize the infrastructure for a shared cataloging resource by developing a set of production-quality tools that operate on a large, standardized set of legacy metadata. We are pleased to report that our goals were met, on time and within budget, through the creation of a development system used daily by UCAI staff. Phase Two began on January 1, 2004, and the project was given a no-cost extension to continue development through October 31, 2005. Findings and recommendations for future work follow.

2.0 FINDINGS

2.1 The concepts of “works” and “images” and their interrelationships need to be discussed and defined further within the community so they can be consistently understood and applied by image catalogers.

Until recently, the visual resource community has not clearly defined the distinctions between work and image information. Although definitions of works and images exist in standards such as Cataloging Cultural Objects (CCO) and the soon–to-be released VRA Core 4.0, there is no certainty about the community’s understanding of or agreement on those definitions.

The nature of visual resource cataloging, traditionally a highly subjective endeavor, adds idiosyncrasies to the process. A photograph, for example, could be considered a work by one cataloger and an image by another depending on whether the cataloger views the photograph as the surrogate of a work or as a work of art in its own right. For a general collection, images of the furnishings of a room might be appropriately cataloged using the room as the work, whereas a specialty collection might need each piece of furniture to be separately identified and cataloged. This is a granularity problem, and further elucidation is needed to uncover the problems posed when multiple levels of granularity are represented in a single database.

Records for architectural works present additional problems for distinguishing between works and images. Architectural buildings and sites are often complex works that contain multiple whole/part relationships and sometimes require two or more levels of hierarchy within the work record. Due to a lack of standard practice for cataloging architectural works, two institutions may catalog the same building or complex in very different ways. For example, a chapel located in a cathedral could be considered a work, or as a part of a larger work. In another example, a plan for a building might be considered a view of the building by one institution (which then catalogs it at the image level) while another institution may consider the plan a work itself, particularly if there is considerable descriptive information that the institution wants to record, such as designer, date, material, etc. Further discussion is needed to determine the relationships between works, and between works and images, and how relationships should be explicitly coded in the metadata record.

The definitions of entities such as a work are at an abstract level, and therefore are open to interpretation. The image cataloging community may never have total agreement about the proper cataloging level because it will always depend upon the cataloger’s perspective of a resource and the needs of the local institution. However, as cataloging communities increasingly share their records in broader contexts, reconciling how records for the same resource can effectively interoperate together becomes a greater necessity.

2.2 Local systems need to cleanly separate work and image information.

The VRA Core 3.0 element set and the CCO guidelines prescribe a hierarchical record structure: work records and (affiliated) image records. Some communities call this a parent/child relationship, with some elements from the parent inherited by the child. This is a relatively new concept in the visual resources community, and has not yet had much uptake there. In order for record sharing and for a central utility to be efficient and effective, record structures must be consistent. This in turn may call for the development or redesign of image cataloging databases to meet the hierarchical standard.

Hierarchical structures make a clear separation between work and image information, particularly in the area of titles (work title=*Red Barn*, image title=*detail of doorway*), by using different elements to record the data. Flat records usually combine the work and image title information into a single element, sometimes separated by punctuation such as a slash or colon (*Red Barn: detail of doorway*). Typically, the punctuation is not used consistently, making interpretation (both human and machine) quite difficult. A significant amount of legacy data contains titles constructed in a flat fashion due to slide label needs. This is an old practice in which catalogers attempted to save space on labels by concatenating as much information together as possible.

When these two types of record structures are combined in a single system such as UCAI, problems are created for both searching and clustering. In searching, listed titles (from flat records) may be separated by long, complex (and indistinguishable) image title strings. Reliable image title searches may be impossible. In clustering, records do not come together appropriately because the algorithm identifies matches by similarity of text. The algorithm does not see “Notre Dame” and “Paris: Notre Dame: interior view of apse” as similar enough to fall into the same cluster.

To follow the standards, work and image information should occur in separate records. If this is not possible in a local system, work and image information should appear in separate elements. It is essential that work and image data within a flat record be adequately differentiated in order for consistent mapping, clustering, and retrieval to occur.

2.3 Inconsistent cataloging practices, both within and across institutions, are one of the biggest barriers to efficient automated processing and retrieval.

The UCAI team asked its data contributors to send documentation of cataloging practice with their data: for example, element definitions, thesauri used and for which element, and stylistic consistencies. For most contributors, this documentation was sparse or non-existent. It is not hard to imagine two (or more) catalogers in the same institution finding their own individual interpretations of element definitions, and it is reasonable that practices change as well over time and as staff change. When mapping from native records to the UCAI standard record, UCAI metadata analysts studied institutional documentation of prescribed practice (in the rare instances that it existed) in addition to combing through thousands of records looking for cataloging and data entry patterns in order to determine practice.

When looking at data in the record clusters, the UCAI team could see where like works should cluster, but the data were not consistently presented in elements in a way that an algorithm could make this determination. For example, in one institution, building names were recorded in a variety of ways. One cataloger consistently noted all of the descriptive information in the Title element as follows: Building name(.) Image title(.). [e.g., Temple of Jupiter. Front façade.] At the same institution, a different cataloger used the BuildingName element to record 'Temple of Jupiter" and the Title element to record "Front façade." The problem of where and how to record data is exacerbated when looking across multiple institutions.

Another example of inconsistent practice are the various methods of documenting personal names associated with a work. Some institutions use an Agent element to hold all personal names, adding a distinguishing type (personal or corporate) and a role (artist, owner, subject, etc.). Other institutions require personal names as subjects to appear in a Subject element. While each practice is appropriate when applied consistently within its own database, it becomes problematic when the records are shared.

Such problems are compounded in a central utility. Routines can be written to attempt to standardize data values within an element, but it is significantly more difficult to standardize data values across different elements.   The UCAI mapping and ingest processes manipulate native data to standardize them within elements of the UCAI standard record. The UCAI clustering algorithms then compare textual similarities within an element. The bifurcation of data, split across two elements (for example, using both the BuildingName and Title elements to identify an architectural feature), is difficult to reconcile and could prohibit records for the same work from being brought together in the same cluster.

It is increasingly important for image catalogers to recognize that the records they create have multiple uses in multiple places and have users outside of their institutions. Implementing and committing to agreed-upon community standards will help, and data cleanup projects to adjust legacy records to these new standards would be beneficial to all if those records are to be shared.

2.4 Standardization tools and technologies, including controlled vocabularies, will improve record and database quality.

When the UCAI project began in April 2002, few common data standards were used by this community. Data element sets such as the VRA Core 3.0 and Categories for the Description of Works of Art (CDWA), which have become part of the image catalogers' lexicon, were new and not very widely implemented. Data content standards, such as Cataloging Cultural Objects (CCO), which cater to the image cataloging community, have only existed for the past year. The application of controlled vocabularies across institutions has been both limited and varied.

Data standards are immature and their adoption is slow in the image cataloging community. The UCAI team heard many reasons for this, ranging from "this community is underfunded and is coming to automated cataloging after the bibliographic community" to "image cataloging involves more interpretation than book cataloging" to "each image is currently cataloged separately as an original item." Nonetheless, standardization would bring a wealth of new opportunities to this community, including the ability to share records and re-use each others' work.

This community of catalogers (which includes museums, libraries, archives, and image collections) needs to gain a common understanding of standards (content, structure, communication) and of the benefits of sharing. A domino effect will take place as common knowledge grows. An understanding of data structure aids in defining data elements, including their purpose and use. In order to understand the purpose and use of data elements, catalogers document their practice. Documenting practice fosters consistency of application. With consistent cataloging practice comes ease and efficiency of sharing data. The common good is reinforced through the use of a shared central utility.

Accepting standard definitions of data elements and strictly applying those definitions will take this community a long way toward sharing records and thus increasing cataloging efficiencies. An image cataloger recognizes that the titles "The Bride Stripped Bare" and "The Bride Stripped Bare by her Bachelors, Even" in all likelihood refer to the same work. An algorithm that is comparing text may not draw the same conclusion. A cataloger might recognize an object described as a "painting" as the same object as a "watercolor", but a computer likely will not. Following agreed-upon standards for what constitutes a title or the level of granularity for the type of material will improve consistency in indexing and retrieval.

The UCAI team often found that the clustering algorithm brought together records from the same institution, thus showing that when data standards, controlled vocabulary, and cataloging practice were consistent within an institution, like records clustered. Imagine if this were true across institutions. Broad and ongoing educational opportunities for image catalogers in the importance of data standards, tools, and techniques should be developed and promoted.

2.5 Standardizing repository names and WorkType data would yield improved processing and retrieval.

The UCAI clustering algorithm was based primarily on Agent and Title data. To a lesser extent, Date was also incorporated into the algorithm, because of the need for wide flexibility between broad date matches (e.g., for antiquities) and very precise matches (e.g., for contemporary art). As UCAI analysts worked through the daunting task of analyzing clusters to determine the appropriateness of inclusion and exclusion, it became clear that utilizing other data elements could potentially increase the precision of the clusters. The repository name (in the element Location.Repository) and the WorkType element became a focus for further investigation.

The three different paintings by the same artist, each titled “Watson and the Shark,” for example, could be differentiated and separated into distinct clusters because they were located at three different museums. Using the repository name seemed promising, but more analysis was needed to ascertain how reliably the element had been coded and how consistent the data values were. In the UCAI database of more than one million records, the Location.Repository element appeared in 58% of the records, however the data values showed a staggering range of forms and abbreviations. The team attempted to find an authoritative electronic database of international museums, but was unsuccessful.

The UCAI team believes that if a standardized, authoritative list of repository names existed and were it used by image catalogers in coding this element, the Location.Repository element could significantly improve the precision of clustering and subsequent retrieval.

A similar exercise was carried out with the WorkType element. The WorkType could potentially distinguish, for example, architectural drawings of Chartres from paintings of Chartres from photographs of Chartres. However, the data in the UCAI database revealed both an absence of the element and a serious inconsistency in understanding of the meaning of the element. The granularity problem also resurfaced here. UCAI staff tried to find a hierarchical list of WorkTypes that could be used, for example, with categories in the Art and Architecture Thesaurus, but did not find anything satisfactory.

The UCAI team believes that an agreed-upon hierarchical list of WorkTypes would be beneficial to the community as a way to group like works (or images) and to improve clustering, searching, and retrieval.

Controlling the terms both for repository names and WorkTypes would help image catalogers create standardized and shareable metadata records. Spelling errors, typographical or otherwise, and unassociated abbreviations could be avoided. Formatting can also be structured. Controlling the data values for these two elements would improve precision and efficiency.

2.6 Unique object identifiers would significantly improve processing and retrieval.

Internationally recognized numeric identifiers have been enormously useful in the bibliographic world as metadata is shared and re-purposed. ISBNs and ISSNs provide instant recognition and a quick match point for monographic and serial textual works. For artworks and architectural structures, there is no equivalent to an ISBN or ISSN. Museum accession numbers can sometimes serve this purpose, but they are rarely recorded in bibliographic records for images. Museum accession numbers are hard to find, and they are only available for artworks that are located in a museum or collection. The numbers are specific only to that institution, so there is no central registry for that information. Architectural structures and sites have no such identifying system.

An international and coordinated object identifier registry, perhaps modeled after the ISBN/ISSN, could provide an efficient method of identifying objects (and, presumably, works). Searching, record matching, clustering, and retrieval would be expedited and improved with a unique identifier system.

2.7 Minimal-level record standards should be determined so that catalogers will know what is required to uniquely identify (or distinguish) a work or an image.

The bibliographic community has long recognized the need for minimal-level records, since meeting the ideal (or fullest) record standard is not always necessary or possible. Identifying information for art works relies on textual information such as Title, Agent or Culture, Date, Site or Repository, and WorkType. The UCAI Project Team did some analysis of its legacy data to determine its own minimal-level record standard.

- **Title.** 98% of the records in the UCAI database contained a title. Many of those are descriptive titles, assigned by the cataloger, making them difficult for a computer algorithm to match related records.
- **Agent or Culture.** 65% of the records in the UCAI database had agent names and 21% named cultures. Cultural affiliation, rather than personal name, is often the only known information for pre-modern artworks.
- **Date.** 82% of the records in the UCAI database showed dates. A significant amount of normalization was necessary to make dates useful in clustering because of widely varying formatting practices.
- **Site or Repository** (in the fields Location.Site and Location.Repository). For architectural works and archeological sites this is a geographic location. For art objects this element is the individual or institution that owns or houses the work. Repository name proved problematic due to the lack of consistent vocabularies available to image catalogers. Therefore the format of this information across contributors varies widely
- **WorkType.** WorkType identifies whether the object being cataloged is a painting, photograph, or sculpture. Like repository name, this element is problematic because there is no specific vocabulary from which to choose these terms. Catalogers approach this element with different levels of granularity and subjectivity. Content standards for WorkType would be useful for both clustering and for providing ways of grouping and browsing records.

For statistical purposes, the UCAI team established minimal record requirements. In order to be considered a “good” record, the following elements must have been present: Title, Agent or Culture, Date, Site or Repository and WorkType. The number of records that met those criteria was counted. Based on those criteria, only 45% of the records in the UCAI database were “good.” Reducing the criteria for “good” records to the presence of the Title, Agent or Culture and Date elements increased the percentage to a still disappointing 69%. This finding can help the image cataloging community focus on areas for improving the quality of its records. Rather than putting time, energy and money into agreeing on data work-by-work, the community can work on normalizing or standardizing particular elements. The elements used for UCAI's minimal record are useful to the end user and would improve the effectiveness of the clustering algorithm. Prioritizing these elements for standardization seems worthwhile.

2.8 Thumbnail images are important for identification but should not be required.

After considerable debate, the UCAI team has found that thumbnail images should not be required for every record.

Thumbnail images are extremely useful and valuable data elements for the visual matching of images and for distinguishing between similarly named works. If a cataloger has an image at hand, a quick comparison with thumbnails from a central utility could usually determine whether the images are the same or different. Occasionally there might be need for a larger image to confirm details; links to larger images were not pursued for this project but should be considered for a central utility.

The UCAI team posed the following question to itself, the UCAI Partners, its consultants, and to others in the image cataloging community: is a thumbnail image a required metadata element for a central utility? The first answer from most people was "yes." Upon further consideration, most people backed off from that unqualified "yes." If the thumbnail were a required metadata element, half of the UCAI database (500,000 records) would not meet the requirement. Upon hearing these numbers, most of those polled felt that a record without the thumbnail metadata element was still a useful record.

Requiring thumbnails as a metadata element should be revisited when a central utility is realized and the universe of digitized images grows.

2.9 Automated processes help with standardization, but record and database quality will be unacceptable without significant investment in manual cleanup.

The re-use of legacy data is highly desirable because the data represent an intellectual investment too valuable to be lost.

Despite the difficulty in working with legacy metadata, the UCAI records revealed an incredible richness, particularly in subjective areas such as subject and culture. Starting over and losing many collective years of research and discovery would be unnecessary and wasteful. It is worth the effort to rescue legacy data, but the process must be a mix of automated and manual techniques. Automated handling alone is insufficient.

When aggregating large numbers of metadata records, significant duplication exists between institutions, as every institution catalogs the (same) most popular works. There is also significant duplication within institutions using flat databases, as complex or heavily-studied works have many records corresponding to different views, details, etc., as well as separate records for multiple copies. Clustering involves tradeoffs between data redundancy and data invisibility. An unclustered database would contain many duplicate records. For popular and complex works, hundreds of records may exist for a single work, making displaying, sorting, and evaluating the records a difficult task. In contrast, a clustered database (particularly an imperfectly clustered database) would render some records invisible within heterogeneous clusters (which are confusing).

Unique identifiers would be the most efficient method of identifying duplicates, but there is no universal system of identifiers, and partial systems of identifiers (such as museum accession numbers) are not widely used. Inconsistent descriptive practice also complicates the identification of duplicates, as there is substantial variation in how the key identifying information (artist names, work titles, dates, and location names) is recorded. In addition, many records lack some of these data, sometimes because they are inapplicable to the record (such as works by unknown creators), but more often because of incomplete information being available to the original cataloger, information not being required locally, or because the data were lost in the mapping process.

There is a wide body of work on automatically grouping a set of records based on their properties, which provides many techniques and potentially useful algorithms for overcoming these problems. However, most clustering algorithms are of limited use, as their assumptions are very different. Most clustering algorithms assume that there are no duplicates. Related to this is the fact that the algorithms generally feature similarity measures that produce a scale of values. When identifying duplicates, however, a binary yes/no answer is typically desired.

In particular, there are several clustering algorithms that use statistical approaches to clustering text, such as Latent Semantic Indexing. Unfortunately, these algorithms all require much more text than is available in typical metadata records for cultural works. Additionally, these algorithms typically work by identifying words or word-stems that appear in two records and calculating the significance of the common words. But because of inconsistent wording and abbreviations, there are sometimes no common words (or only a very small number) in two records that describe the same work.

While automated methods of clustering records showed some promise, the UCAI team is not optimistic that they will be effective on their own. Methods of manually correcting the automatic processes should be integrated: for example, by allowing manual decisions about grouping or splitting records to be fed into the clustering process to override the automatic record comparisons.

When comparing two values, effective normalization is essential to acceptable results. All text strings should be normalized to remove differences in case, punctuation, spacing, etc. Additionally, removing common word endings (using PorterStemmer or similar software), expanding abbreviations, and employing synonym rings can also help reduce variability. Any numeric values (such as dates) should be parsed and reformatted in a consistent manner. However, this is not a trivial operation; we found more than sixty different date formats in our source data.

Controlled vocabularies can also help reduce variability. We matched artist names against ULAN, and then compared ULAN identifiers when possible. Where there is no existing vocabulary (or existing vocabularies aren't widely used), the list of unique values can be manually grouped to create an ad-hoc vocabulary. The list of unique values can sometimes be surprisingly small. In our database of nearly 1,000,000 records, there were only 25,000 unique Location.Repository values, and only 38,000 unique Location.Site values. While these numbers are still significant, they are much more manageable.

Many records are missing data, even for key elements such as Agent, Date, Location.Repository, and Location.Site. The data may be missing because it is inapplicable to the work being described (the repository of a work that's been destroyed), because the information is not generally known (the Agent of a prehistoric artifact), or because the person who cataloged the item did not have access to the information or was not required to enter it. The first step in handling missing data is to determine, based on the nature of each element and the cataloging processes of the contributing institutions, whether missing data in each element should be considered significant. We decided that Agent was the only element where missing data was significant, and grouped records without Agent elements with records with "unknown" or "anonymous" Agent values. For other elements, comparisons can only be made when both records being compared have values for a given element. Our clustering process compared Agent and Title independently; then compared the Date, Location.Site, and Location.Repository values of two records; and required that there be no mismatches in those elements and that at least one of the elements was present in both records and matched. The last requirement was added to prevent large groups of sparse records from being grouped together based only on Agent and Title.

Works with unknown creators are typically much less rich than records with known creators. Differing techniques for these two categories of records may be helpful. Greater consistency with names and titles of records with known creators may allow them to be clustered relatively well using those elements alone. Works without known creators may need Date, Location.Site, Location.Repository, and WorkType to effectively differentiate them.

Bringing metadata up to current standards will be time consuming and expensive. The ability to use automated techniques will depend on how consistently rules were applied. To achieve an acceptable level of quality, a strategy for appropriate human intervention is necessary.

2.10 A central utility must have authoritative records, not just merged records.

While the UCAI merged record is lengthy and rich, merged records are not useable in their current form within local catalogs. Multiple measurement elements, for example, without the unit of measurement or without a clear description of what is being measured (sculpture with or without its base, individual parts vs. the collective triptych) are not useful. Variations in dates are interesting, but retaining multiple conflicting dates in a catalog record would be confusing. If senior staff were doing the cataloging, a series of checkboxes in the record would allow a user to create a highly customized record informed by their expertise. The UCAI team realized, though, that it is much more likely that junior staff (or students) will be doing copy cataloging, and thus a central utility would need to provide an authoritative record that would not require high-level (or element-by-element) decision making.

Not having an existing corpus of authoritative records is a substantial barrier. The bibliographic world did not face this problem, as the Library of Congress had long been viewed as an authoritative source whose cataloging could be trusted. Their collection of millions of MARC records provided a respected "seed set" for central shared cataloging utilities while the community refined common practices and interpretations.

Analysis of the components of UCAI merged records revealed several categories, which helped the UCAI team move toward the concept of an authoritative record as distinct from the current merged record.

- **Objective data** is physically verifiable, and includes measurements, current repository [and repository identifier], inscription, material, technique, and thumbnail.
- **Defining (but often ambiguous) data** is associated with a work but is not physically verifiable. This data is somewhat objective and might even be generally agreed upon, and includes title, creator, and date.
- **Subjective data** is more open to interpretation. It is information brought to a work by the cataloger (or collaterally by a scholar or an expert), and includes subject terms, topical descriptions, keywords, iconographic terms, and style or period descriptors.
- **Non-existent data** could potentially be blank or null characters in any element. Unique object identifiers are also unfortunately non-existent.

There are significant challenges in identifying authoritative data, and in reaching agreement about authoritativeness. One assumption that could be tested is that authoritativeness comes from the object owner. Museum data could be incorporated for the objects they hold; their objective data elements might be presumed to be definitive (material, technique, measurements, etc.) Ideally, these data should be tracked over time, because even objective data can change. Museum-supplied data elements should be displayed within records as such so the source is known. Incorporating museum data into a cataloging system raises many questions, not the least of which is the standards and formats used. Will records created by museums, presumably containing objective data, be compatible with cultural heritage records created by catalogers at other types of institutions?

How will objective data be obtained for objects outside of museum collections, or when a museum chooses not to share their data? Acknowledged experts (apart from the object owner) could also contribute authoritative records, however experts differ and change over time. Getting data from experts would be a more difficult and expensive path.

Richness comes from including the ambiguous and the subjective data from many viewpoints. The extraordinary wealth of subjective data from the combined legacy records of visual resources catalogers provides an extraordinary range of perspective and potential search terms for users. Adding these elements to the objective data from a museum record could be the beginnings of an authoritative record, and could provide a valuable incentive for use of a central utility for image cataloging.

2.11 Record synchronization between feeder systems and a central utility should be explored further.

The UCAI team began to explore models for connectivity between feeder systems (those systems whose records seed the initial database) and a centralized utility. Catalogers will not make corrections to records redundantly in two systems, and adding a central utility to their toolkit will force them to choose the system on which they will focus their quality improvements. Workflows and incentives must be carefully considered when planning a central utility. Should there be close synchronization (regular file sharing) so that the changes made in the local system can be leveraged by the central utility? If so, can centralized improvements made by others be protected?

Three models were suggested, and others could be defined and explored.

- The **separate model** posits a one-time download of records from a donor source. These records become part of the central utility, but would presumably be identified as having their origins at the donor site. The donor records and the copy at the central utility would grow apart over time as each changes; these changes are made independently ("separate") and are not linked or coordinated. Over time, the centralized records would become recognized as distinct from the donor institution, and would be valued for the collective improvement made to them.
- The **linked model** would also pull data from a donor source but would develop an ongoing mechanism (such as the OAI Protocol for Metadata Harvesting) to regularly retrieve new and changed records. Those records would overlay or be added to the central utility. The central utility would be seen as a central hub with many connecting spokes that bring together the work of many collections or institutions in one place. The records could (but wouldn't have to be) linked back to the donor systems so they could be viewed in a local context if desired.
- A **hybrid model** could implement the best features of both the separate and linked concepts. Datasets could be pulled from partner institutions using an ongoing mechanism to manage new and changed data. In addition, a separate internal database could be established, in effect making the central utility itself an equal partner in record creation. The records in this database could be newly entered or customized records from the larger database. These records would be shorter than merged records, and the hand-tooled, evaluative aspect of their creation would position them toward being more authoritative. Truly authoritative records would develop over time, and would gradually become the major focus of the system. At some point, harvested records might no longer be necessary, and the merged record concept would drop away.

The relationship between local data and shared data is a discussion that needs to continue within the visual resources community, and has a bearing on the roles and relationships between local and central processing systems. With a closely synchronized (harvesting) model, local data standards and practices are perpetuated, which may not serve the greater goal of reducing redundant work and re-purposing. There are risks (records maintained locally instead of centrally) and tradeoffs (redundant maintenance), however with a looser connection between systems. Ultimately, a central utility is a standardizing influence within a community, and it would be wise to leverage this advantage.

2.12 Planning for a central utility will need to include quality control and database growth processes.

The planning process for a central utility must address some serious questions about database and record quality issues, as well as database coverage and growth. If record quality is poor upon rollout of a central utility catalogers will disregard it, and persuading them to return could prove a serious challenge. A substantial effort must be defined and undertaken prior to rollout to assure useable, if not high quality or authoritative, records that will encourage routine uptake. A variety of ongoing quality control techniques should be considered, including community-led maintenance efforts, a central quality control team, incentives for record cleanup, an editorial board that maintains standards and guidelines, etc. It is essential to develop and make known a coherent plan (including a timetable) for establishing authoritative records.

In addition, a central utility should strive to include metadata for the broadest range of work and images, covering all cultures and time periods. This includes identifying gaps in coverage and actively identifying institutions that have metadata files that could fill those gaps. Some thought should be given to ingesting thumbnail-only files and matching them to the appropriate records. The service should grow through file loading as well as through the keying of individual records.

3.0 RECOMMENDATIONS

3.1 **Continue efforts to establish a central utility through which image metadata records could be re-purposed and shared.** The scholarly community would be better served by improved and consistent access to still images. Such a utility could form a "central hub" for the cultural heritage community, and would encourage and promote further standardization. Expert catalogers could provide coverage of a broader range of works because their redundant copy cataloging workload would be streamlined.

3.2 **Develop a business model for a central utility.** Some of the areas that should be addressed include:

- Mission, vision, values, and goals
- Branding
- Product or service description
- Needs assessment; market research
- Environment and competition
- Markets and services; customer development
- Pricing
- Communication, including the need for an Editorial and/or Advisory Board
- Organizational structure
- Operations, including facilities and equipment, management and staffing, and legal issues; workflow model, hardware requirements, quality control plan
- Metadata rights issues
- Financial planning, including revenue options, pricing structure, potential ROI
- Maintenance and enhancement processes, including sustainability
- Product evaluation and usability assessment
- Deployment/rollout and training plan

3.3 **Develop and support research projects** to explore:

- the specific challenges posed by architecture and three-dimensional object metadata
- using museum data as a basis for creating authoritative records
- the establishment of a unique object identifier registry
- developing a tool that could associate specific instances with broad categories, so that gap analysis could be better quantified

3.4 **Support the cultural heritage community in education and standardization efforts** to achieve:

- deeper understanding of the implications of sharing metadata
- clearer understanding and implementation of hierarchical record structures
- clarification of the concepts of "work" and "image"
- guidelines for minimal level records
- agreed-upon vocabularies/standards for Work Types and repository names

4.0 PROJECT PROCESSES

The goal for UCAI Phase Two was to further advance and stabilize the infrastructure for a shared cataloging resource by developing a set of production-quality tools that operate on a large, standardized set of legacy metadata. Specific objectives focused on three key areas: needs assessment for a production environment, database augmentation, and work on refining existing and developing new database processing tools.

4.1 Needs Assessment

The UCAI team did some preliminary analysis to determine the needs for a central utility's technical infrastructure. The primary audience for this shared cataloging utility is, unsurprisingly, image catalogers. An informal market survey revealed that the size of the membership of the Visual Resources Association (approximately 700 people) and the number of higher education institutions (approximately 3,200 institutions) would give us the order of magnitude for the potential number of real users. Our partners described the variety of user tasks undertaken by catalogers, including searching, browsing, selecting, and comparing. We assume additional tasks such as editing and exporting would also create system load, although that functionality was not developed in the Phase Two system. A wide range of query types should be expected, with keyword searches anticipated to be the most frequently used search. See Appendix A for more detailed projections.

In addition to needs assessment for the technical infrastructure, needs assessment for the database content is also critical. Catalogers will not use a shared cataloging utility unless it has broad thematic coverage, significant niche coverage, and enough mass to convince users that their needs might be met. Acquiring appropriate content also needs to be balanced with the timing for rollout of the service. If catalogers try the service but find the database content inadequate, there is risk that they won't return to try again. If the service is not released until all appropriate content is in place, there is risk that alternate sources for the data will have been established.

Comparing extant database content with ideal database content is a difficult, and a largely manual, undertaking. There are challenges with trying to assess coverage of legacy records when there have been no consistent data content standards used within the visual resources community that match the broad categories used in scholarly literature. For example, art objects and cultural materials are often grouped in art history survey texts into categories by

- broad geographic regions (Western vs. Non-European)
- countries (Japan vs. China)
- time periods (Ancient vs. Middle Ages)
- styles (High Renaissance vs. Mannerism)
- work types (painting vs. sculpture)

In legacy data, these terms are found in a variety of elements (style/period, description, date, culture, subject) and in a variety of forms and combinations (England vs. English; Meso-American painting) thereby making it very difficult to match, group, and assess.

Often the categorizing data (e.g., Impressionism) were not present in the record at all. Once UCAI staff realized the enormity of the challenge of conducting text searching for gap analysis, we decided we could at best have our contributors assess their own content coverage and use that information to determine where the gaps were.

We looked to two other avenues for assistance with gap analysis. A master file for the core corpus of Western arts works that could match against the existing file in an automated way would help solve this problem. We are not aware of the existence of such a file, but establishing one (a "core image library") would be a useful next step by experts in the community. We also looked to our colleagues at ARTstor, a service for whom "gap analysis" would also be useful, for their approach and tactics. They investigated work done at New Grove that might be helpful in mapping records into various hierarchical categories, which could lead to better quantitative assessment of strengths and weaknesses.

4.2 Database Augmentation

UCAI Phase Two was quite successful in augmenting the prototype database, doubling the number of partners, and establishing a database of approximately 1.3 million records, the largest known compilation of image metadata records to date. Detailed statistics are provided in Appendix B.

As was done in Phase One, the UCAI team for Phase Two identified three very diverse datasets so that the project would benefit from solving the most complex mapping, ingest, and clustering problems. The six datasets that resulted from both Phases One and Two represented a variety of data formats including: relational database tables from Microsoft Access, Filemaker Pro, and FoxPro, as well as two different flavors of MARC21 and a local XML schema. As expected, this diversity of data formats provided a rich testing ground for the operability of tools and processes. While our software tools were not able to be used by the data contributors themselves, we nonetheless made significant progress by using a commercial mapping tool instead of the customized, hand-mapped approach used in Phase One.

Finally, the acquisition of three new datasets and the re-acquisition of fresh data from our Phase One partners proved to be incredibly useful as we explored the nuances of clustering duplicate records across the six collections. This additional richness improved our ability to discover clustering examples (both good and bad) and to develop new insights into the decisions catalogers make in creating metadata records.

4.3 Work on Database Tools

The bulk of the UCAI team's work in Phase Two was focused on software tools. The five areas that we concentrated on were: (1) further developing the "conceptual tools" related to works and images; (2) developing a new mapping tool; (3) generalizing the ingest tool; (4) refining the clustering algorithms; and (5) extending the merge tool beyond basic functionality.

4.3.1. Conceptual Tools

At the time of the Phase Two proposal, further development of "conceptual tools" seemed an urgent need for UCAI. As our work evolved during Phase Two, however, we found this was somewhat less necessary as a UCAI task for several reasons. The only definitions at that time in the visual resources community for works and images were those found in the data element set VRA Core 3.0, which UCAI found sorely lacking in specificity. The CCO data content standard has since released its own definitions of works and images as well as a preliminary list of relationship types between works and works as well as between works and images. In addition, the soon-to-be-released VRA Core 4.0 will put forth its own work/image definitions which were informed by UCAI's work, and it will incorporate the CCO relationship types into the value lists for the VRA Core 4.0 XML Schema.

For the purposes of the UCAI prototype and particularly for its clustering work, the UCAI team did find it helpful to create its own working definitions for works, images, and clusters (see Appendix C). If the underlying legacy metadata are not consistent or non-existent, however, there is limited practical applicability for these definitions.

All of these efforts are an excellent starting point for discussing these issues and for working towards standard practice to explicitly encode these relationships into catalog records, but much more work needs to be done. The community needs to come to consensus about whether the current work/image definitions and relationship types will work for visual resource records, and whether they are important enough to take the time to code in daily cataloging. In addition, the image community needs to compare its definitions with those being promoted in the larger cataloging community to identify if there might be any problems in how visual resources records would interoperate with records from other formats when they come together in a shared environment.

4.3.2. Mapping Tool

The Phase Two proposal described a new GUI mapping tool that UCAI would develop to allow non-programmers to specify the mapping of a dataset to VRA Core, the standard that forms the basis for the UCAI standard record. We envisioned a tool that would allow one of our Metadata Specialists (not a programmer) to specify the mapping directly and view the results in real time. UCAI staff was fortunate to find and test a satisfactory commercial mapping tool, MapForce by Altova, which suited our needs very well and allowed our software developers to focus in other areas instead. Our Metadata Specialists were able to use MapForce to convert from the dataset’s native (source) schema to the UCAI standard record schema and output the results into an XSL stylesheet that could be used in the pre-ingest process.

4.3.3. Ingest Tool

The ingest tool was refined to work with the MapForce output, and successfully performed data standardization in a uniform way. Workarounds for some of the limitations of MapForce (handling missing elements, complex formatting operations) were also added to the ingest tool. We moved away from the highly customized processes of Phase One, and have substantially reduced the time and effort needed to ingest new collections. Another significant achievement of the ingest process was implementing the Union List of Artist Names (ULAN) which improved personal name clustering and retrieval through comparing ULAN identifiers in addition to non-standard name forms.

4.3.4. Clustering Algorithm

Clustering together multiple records that represent the same work was the most significant challenge. A system with multiple duplicate records is ineffective and inefficient to use, and the UCAI team took seriously the need to identify these duplicates using automated tools. Major work was done to refine the clustering algorithm to support additional elements (Date, Location/Repository, and Location/Site in addition to Agent and Title from Phase One). Extensive work on improving performance was also undertaken, resulting in reducing the time needed to cluster the entire database from 6 days to less than 15 hours.

Details of our findings regarding clustering are described earlier in this report. In terms of software development the team utilized an iterative process, making incremental changes to the clustering algorithm, running it against the full database, and then reviewing the results.

4.3.5. Merge Tool

In Phase One, a composite record concept was established for displaying a work unit or cluster. Its purpose was to create a synoptic view that would make it easier for catalogers to review and export data from UCAI. Once records are established as being part of the same cluster, data values were merged into a single work record (i.e., redundant values are de-duplicated while unique values are retained). In order to illustrate the concept, the UCAI team hand-crafted a composite record for a cluster of records related to Marcel Duchamp's work "The Bride Stripped Bare." This was then used to demonstrate the concept to various library groups for input.

In Phase Two, based on positive feedback, the team moved forward to build a merged record tool to automatically create the composite record. The tool used the basic ideas in the model with some additional functionality (see the examples in Appendix D). When a user does a search within the clustered database, the tool returns a list of clusters with the most basic identifying information such as: a thumbnail (when available); the number of records in the cluster along with the number of contributors represented in the cluster; and minimal descriptive metadata (title, agent, date, location).

When a single cluster is selected, the user is then taken to a page with three basic sections: Preferred Values, Merged Values, and Individual Records. The Preferred Values (title, agent, date) which display at the top of the page are chosen algorithmically by the computer based on those values used by the majority of the contributors to that cluster. Preferred Values are followed by Merged Values for each element. All unique values are displayed while redundant values are displayed only once, along with a number indicating how many contributors used that value. In this way, catalogers could choose terms based on consensus if they consider that a sign of authoritativeness. Some values in the Merged Values section may also link to controlled vocabularies (such as the Getty's ULAN) if the form of the agent name matches a form in the vocabulary. This lets the cataloger know if the term is an authoritative form of the name.

Merged values are followed by the list of Individual Records that make up the cluster along with links to the full version of those records if needed. Catalogers can download either the standard record or the native record.

Part of the composite record concept included a desire to group and display image information in a similar fashion to work information but the non-standard nature of this information made this impossible and the UCAI team settled on simply displaying all image title information as a single list following the list of Individual Records.

5.0 PRESENTATIONS AND COLLABORATIVE WORK

5.1 Presentations

Cowles, Esme. "The Problem with Duplicates" part of Metadata Strategies session, DLF Forum, November 2005

Cowles, Esme and Trish Rose. UCAI presentation at California Digital Library’s monthly Tech Talk, March 2005

Kozbial, Ardys. "Metadata: the view from my trench," as part of Session 5, Metadata: a view from the trenches, VRA Conference, March 2005

Kozbial, Ardys. "Shared Cataloging for Images (or, the Trials and Tribulations of Merging Datasets)," presentation at the RLG Art and Architecture Group at the ARLIS/NA Annual Conference, April 2005

Rose, Trish. “Beyond Google: A Union Catalog for Art Image Metadata” as part of Session 3, Preparing for Shared Cataloging: An Overview of Needs, Benefits, and Efforts, VRA Conference, March 2004

Rose, Trish. “Got Data: Now What?” part of Metadata Aggregation session, DLF Forum, November 2005

Rose, Trish. UCAI presentation, UCSD Summer Summit, July 2004

Westbrook, Bradley D. “Beyond Google: a Union Catalog for Art Image Metadata” as part of Session 13, Preparing for Shared Cataloging: An Overview of Needs, Benefit, and Efforts, ARLIS/NA Conference, April 2004

5.2 Articles

Barnhart, Linda. “The Union Catalog of Art Image Metadata” *Creating the Digital Art Library,* October 2005, pp. 74-85

Barnhart, Linda, K. Esme Cowles, Joseph Jesena, Trish Rose and Bradley D. Westbrook. “In pursuit of the Holy Grail: the Union Catalog of Art Images (UCAI) as shared cataloging utility” *Art Libraries Journal,* 29/4 2004, pp. 22-25

Barnhart, Linda, K. Esme Cowles, Joseph Jesena, Trish Rose and Bradley D. Westbrook. “Looking for Good Art: Web Resources and Image Databases. Part Two: General and U.S. Tools” *Searcher: The Magazine for Database Professionals,* October 2004, pp. 17-18

Westbrook, Bradley D. and Trish Rose. “Beyond Google: A Union Catalog for Art Image Metadata” *VRA Bulletin,* Winter 2005 Volume 31, Number 2, pp. 45-47

5.3 Meetings with Groups Doing Related Work

The UCAI team met with researchers from various other projects to learn from each other, share findings, and explore possible synergies, including:

ARTstor
California Digital Library American West Project
CLiMB project at Columbia University
Digital Library Federation OAI metadata aggregators
NITLE Advanced Search Technologies
OCLC Office of Research Metadata Switch Project

5.4 Standards and Committee Work

Much of what we learned about metadata aggregation for images and applying VRA Core 3.0 we were able to share and feed back into data standards work as members of the CCO Advisory Board, the Visual Resources Association Data Standards Committee, and the development team for VRA Core 4.0.

These activities demonstrate we were not working “under the radar” but very much out in the public sharing our findings with the community, using their feedback to advance our goals, and helping to advance data standardization needs.

5.5 Partner Meeting

The UCAI team held its Partner Meeting on February 17-18, 2005, for the purpose of getting feedback on dataset mapping, the newly clustered database, and the merged record display. This meeting was very successful in helping the team develop different strategies for clustering works of known and unknown creators, and for identifying both conceptual and practical issues.

6.0 ACKNOWLEDGEMENTS

UCAI Core team members included:

Brian E. C. Schottlaender	Principal Investigator
Linda Barnhart	Project Manager
K. Esme Cowles	Database Developer
Joseph Jesena (through May 2005)	Data Administrator
Ardys Kozbial (July 2004-)	Metadata Specialist
Tricia Rose	Metadata Specialist
Bradley Westbrook (through May 2004)	Lead Designer

Additional support at the University of California, San Diego (UCSD) was provided by:

Luc Declerck

Administrative support: Allyson Collins, Nhu La, Tina Nguyen, JoAnn Raimond, Frieda Romero, Colby Sellman

Content expertise: Steve Illot, Vickie O’Riordan, Greg Reser

Technology support: Richard Caasi, Declan Fleming, Chris Frymann, Juliana Gorney, Ron Joyce, Ed Totman

In addition, our dataset partners provided a high level of commitment and participation:

Heather Glaser	University of Pennsylvania
Trudy Jacoby	Princeton University
Martha Mahard	Harvard University
Rebecca Moss	University of Minnesota
Sara Jane Pearman	Cleveland Museum of Art
Robin Wendler	Harvard University

APPENDIX A: Projected Usage for a Central Utility

We performed a rough estimate of the amount of data processing and application usage we expected a central utility to need to be able to handle.

Data

The key factors to scalability for the batch processes (ingest, clustering) are the number and size of the datasets. We assumed six datasets, based on the six datasets we had already acquired, and assumed that roughly the same number would be needed by a central utility. We estimated that each dataset would contain approximately 200,000 records. It should be mentioned that differences in cataloging practices mean that records from different institutions are not equivalent for processing purposes, since hierarchical and/or rich records may contain many times as much data as flat and/or sparse records. Based on these figures, we calculated that the ingest and clustering processes would need to be able to process each record in a short amount of time in order to be practical for a central utility.

Usage

Estimating application usage contained more unknowns, but we estimated three factors that we thought would be critical: the number of potential users, the adoption rate, and the rate at which an average user would use the system.

Potential Users

We consulted the 2000 Carnegie Classification of Higher Education Institutions (http://www.carnegiefoundation.org/Classification/CIHE2000/Tables.htm) to determine that there were approximately 3,200 institutions. We assumed that Research Universities and Masters Universities would likely have multiple cataloger FTE, while Baccalaureate Colleges and Associate's Colleges would have one or partial cataloger FTE. This allowed us to calculate that roughly 3,500 catalogers in the United States who would be potential users of a central utility.

Adoption Rate

We estimated the possible rate of adoption to be between 5% and 20%, based on funding level and participation in professional organizations.

Average User

We consulted Susan Jane Williams' Per Unit (Image) Cost and Labor analysis (http://www.vraweb.org/forms/UnitCost2.pdf) to estimate that it would take the average user 10 minutes to catalog a single item.

Based on these figures, we estimated that, depending on adoption rates, at a minimum 18 to 73 works would be cataloged per minute without the use of a central utility. We then estimated that a number of different searches and record views/downloads would be performed for each work cataloged and calculated that between 75 and 300 searches and between 650 and 2500 record views would be performed each minute on a central utility.

Actual Performance

Our software easily met the estimated ingest and clustering performance requirements, ingesting at a rate of 0.02 seconds per record, and clustering at a rate of 0.06 seconds per record. In practical terms, our software was able to ingest roughly 900,000 records in 4.5 hours, and cluster the entire database in 15 hours.

Testing interactive application performance is much more complicated. We looked at the queries used by UCAI staff and partners when testing the system and developed queries and estimated the number of page views for an automated load-testing system. Our web application's response time scaled linearly with the number of users, and handled our estimated volume of traffic with average page generation times under one second. However, many other factors (including network latency, usage patterns, database size, and browser rendering times) affect actual performance.

APPENDIX B: Statistical Analyses

Record and Thumbnail Counts

Contributor	Original Format	Records					Thumbnails	
		Collection	Work	Image	Total	%	Images	%
CMA	Relational	0	101,691	172,875	**172,875**	13%	61,681	14%
Harvard	XML	5,669	272,689	260,276	**538,634**	40%	85,740	19%
Minnesota	Relational	0	8,227	14,035	**22,262**	2%	23,305	5%
Pennsylvania	MARC	0	97,094	165,059	**165,059**	12%	60,138	14%
Princeton	Relational	0	56,274	143,798	**200,072**	15%	25,528	6%
UCSD	MARC	0	136,771	232,511	**232,511**	17%	188,808	42%
					1,331,413		**445,200**	

Element Population

Element	0	%	1	%	Multiple	%
Agent	320,215	35%	402,390	44%	190,670	21%
Contributor	0	0%	913,275	100%	0	0%
Culture	722,818	79%	102,491	11%	87,966	10%
Date	165,839	18%	400,072	44%	347,364	38%
Description	594,264	65%	249,006	27%	70,005	8%
ID Number	510,959	56%	299,554	33%	102,762	11%
Location	258,887	28%	401,085	44%	253,303	28%
Location.Repository	383,399	42%	400,490	44%	129,386	14%
Location.Site	483,507	53%	328,092	36%	101,676	11%
Material	407,768	45%	482,455	53%	23,052	3%
Measurement	544,155	60%	308,065	34%	61,055	7%
Native ID	0	0%	913,275	100%	0	0%
Record Type	0	0%	218,962	24%	694,313	76%
Relation	478,920	52%	76,144	8%	358,211	39%
Rights	892,645	98%	20,630	2%	0	0%
Source	913,275	100%	0	0%	0	0%
Style/Period	802,785	88%	108,529	12%	1,961	0%

Subject	361,177	40%	205,609	23%	346,489	38%
Technique	907,541	99%	5,390	1%	344	0%
Title	15,760	2%	448,003	49%	449,512	49%
Work-Title	21,449	2%	781,103	86%	110,723	12%
WorkType	346,524	38%	388,450	43%	178,301	20%

Core Elements Populated, By Contributor

Core elements: Agent, Title, Date, Site, Repository, Object Type, Material

	UCSD	**Harvard**	**Minn**	**Princeton**	**Cleveland**	**Penn**
0	18,761	0	0	68	5,750	97
1	54,911	0	0	133	19,840	1,339
2	72,037	446	0	3,704	11,094	9,973
3	63,776	3,880	14	9,497	6,430	20,216
4	21,289	29,794	1,345	13,516	25,579	49,874
5	1,368	94,201	1,834	26,254	39,362	58,781
6	0	110,696	2,602	1,933	52,118	24,023
7	0	39,341	2,432	1,600	12,611	756

Well-Populated Records, By Contributor

Records with at least Title, Date and Site or Repository populated.

	UCSD	**Harvard**	**Minn**	**Princeton**	**Cleveland**	**Penn**
Known	26,379	159,672	5,414	31,032	64,994	76,765
Unknown	10,082	63,219	2,791	9,965	48,687	40,986
Total	**36,461**	**222,891**	**8,205**	**40,997**	**113,681**	**117,751**

Agent Population and ULAN Coverage

	Records	**%**
Known (ULAN)	382,677	42%
Known (non-ULAN)	190,529	21%
Unknown	340,071	37%
Total	**913,277**	

APPENDIX C: UCAI Definitions

These definitions were created for the purpose of clarifying concepts within the UCAI team and may or may not conform to definitions in the broader cultural resources community.

A **cluster** is a set of records which describe a work and its related images. A cluster must be composed of at least one work record but can contain any number of image records including none. Also referred to as a **work unit.**

An **image** is a visual representation of a work, either in whole or in part. An image record describes a representation of a work that can be used in certain situations (e.g., when the work itself cannot be experienced firsthand) as a substitute for the work. Images may include photographs, drawings, slides, and many other formats. Also referred to as a **surrogate**.

There is one **institutional record** for each institution contributing to UCAI. Institutional records are used for collecting and storing directory-like information from a contributor to UCAI. This includes institutional address and contact information.

Merged records are used for displaying the records in a cluster in a compressed, economic fashion. Merged records are automatically created with the UCAI merged record tool. The merged record combines non-unique values (to reduce redundant display), preserving the number and identity of records utilizing that value. Unique values are displayed. Merged records have three sections: Preferred Values, Merged Values, and Individual Records. Also referred to as a **composite record**.

A **native record** is data submitted from the UCAI partners but transformed into XML. This initial XML document is stored as the native record. Typically the native record corresponds in all ways to the submitted record in terms of element names and values. However, those element names will be encapsulated in XML tags. In instances where the original data is submitted as tables from a relational database, it may be necessary to merge the tables or elements in them into a single table or record structure and then use the merged structure as the basis for the UCAI native record.

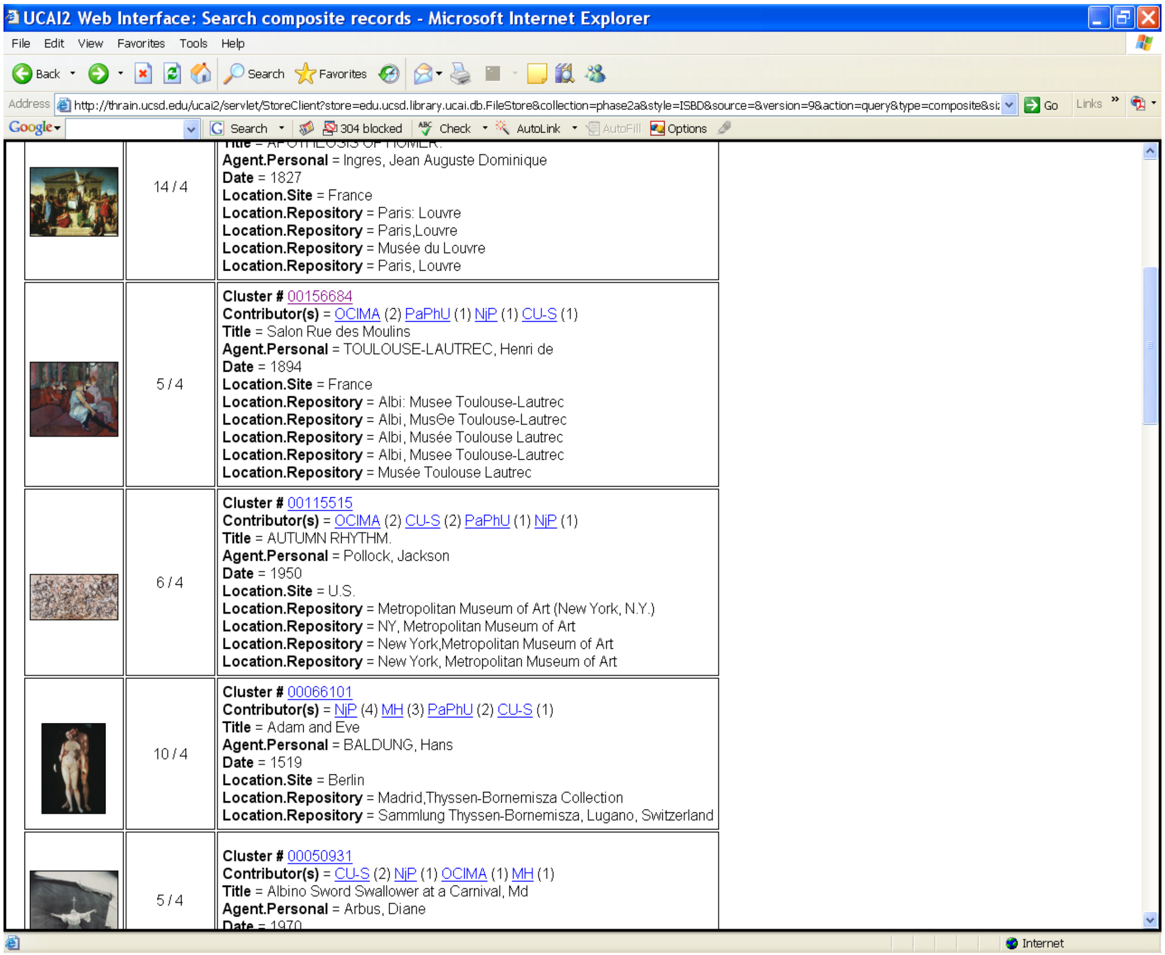
The UCAI **standard record** is the basis for all UCAI database processes such as clustering, merging, and constructing composite records. The UCAI standard record is an extension of the VRA Core 3.0 element set. The UCAI standard record maintains the primary VRA Core 3.0 elements, but has secondary attributes added to several of the elements: title (collection), creator (vital dates and nationality), description (work and group), subject (period and authority), and source (location). The UCAI standard record includes (or is used for) both work records and image records.

A **work** is a unique entity such as an object or event. Examples include a painting, sculpture, or photograph; a building or other construction in the built environment; an object of material culture, or a performance. Works can have parts which may be cataloged as works themselves but have a whole/part or hierarchical relationship with the larger entity.

APPENDIX D: Screen Shots of Merged Records

List of clusters resulting from a search

Each cluster includes information needed for selection such as a thumbnail: number of records/number of contributors; minimal descriptive metadata

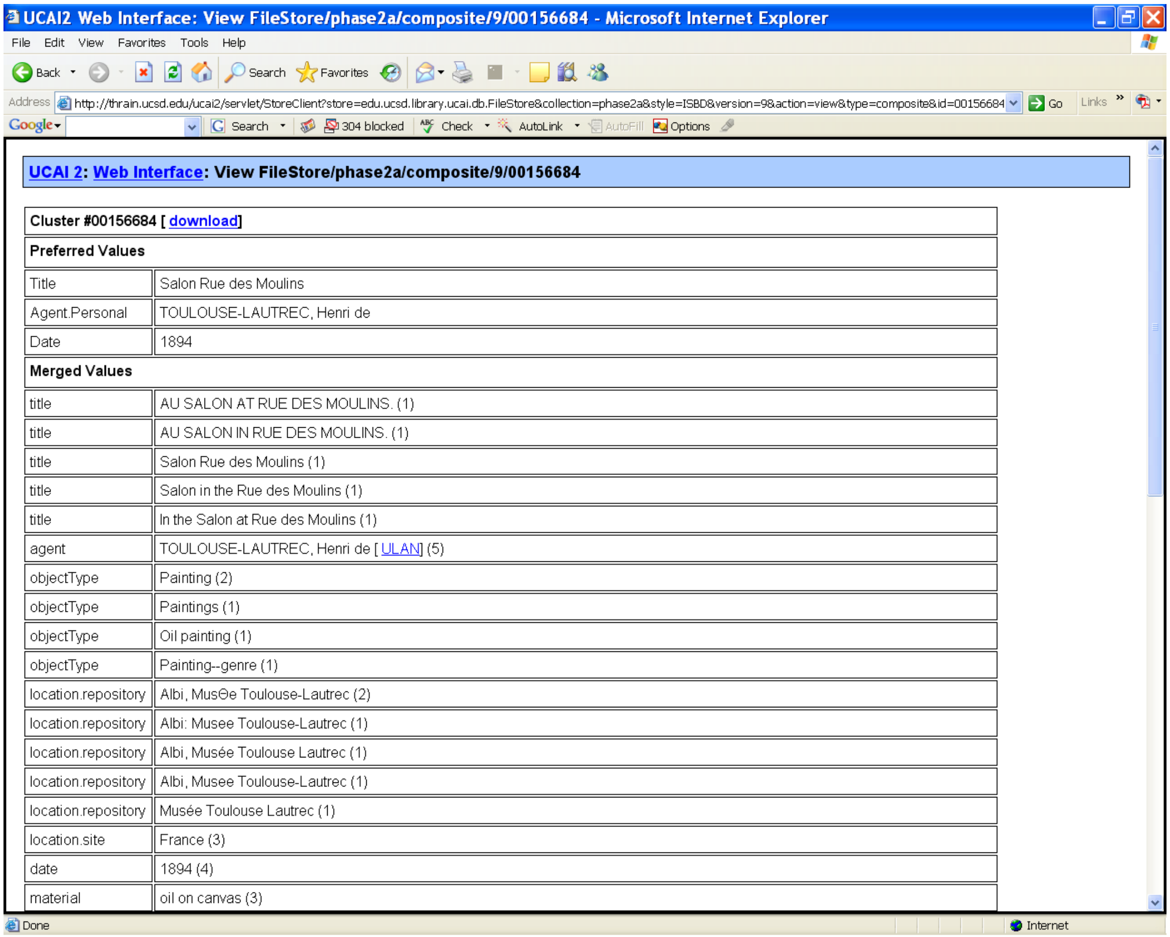

UCAI2 Web Interface: Search composite records - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address
Go
Links
Google
Search
304 blocked
Check
AutoLink
AutoFill
Options
Internet

http://thrain.ucsd.edu/ucai2/servlet/StoreClient?store=edu.ucsd.library.ucai.db.FileStore&collection=phase2a&style=ISBD&source=&version=9&action=query&type=composite&si

	14 / 4	**Title** = APOTHEOSIS OF HOMER. **Agent.Personal** = Ingres, Jean Auguste Dominique **Date** = 1827 **Location.Site** = France **Location.Repository** = Paris: Louvre **Location.Repository** = Paris,Louvre **Location.Repository** = Musée du Louvre **Location.Repository** = Paris, Louvre
	5 / 4	**Cluster #** 00156684 **Contributor(s)** = OCIMA (2) PaPhU (1) NjP (1) CU-S (1) **Title** = Salon Rue des Moulins **Agent.Personal** = TOULOUSE-LAUTREC, Henri de **Date** = 1894 **Location.Site** = France **Location.Repository** = Albi: Musee Toulouse-Lautrec **Location.Repository** = Albi, MusΘe Toulouse-Lautrec **Location.Repository** = Albi, Musée Toulouse Lautrec **Location.Repository** = Albi, Musee Toulouse-Lautrec **Location.Repository** = Musée Toulouse Lautrec
	6 / 4	**Cluster #** 00115515 **Contributor(s)** = OCIMA (2) CU-S (2) PaPhU (1) NjP (1) **Title** = AUTUMN RHYTHM. **Agent.Personal** = Pollock, Jackson **Date** = 1950 **Location.Site** = U.S. **Location.Repository** = Metropolitan Museum of Art (New York, N.Y.) **Location.Repository** = NY, Metropolitan Musuem of Art **Location.Repository** = New York,Metropolitan Museum of Art **Location.Repository** = New York, Metropolitan Museum of Art
	10 / 4	**Cluster #** 00066101 **Contributor(s)** = NjP (4) MH (3) PaPhU (2) CU-S (1) **Title** = Adam and Eve **Agent.Personal** = BALDUNG, Hans **Date** = 1519 **Location.Site** = Berlin **Location.Repository** = Madrid,Thyssen-Bornemisza Collection **Location.Repository** = Sammlung Thyssen-Bornemisza, Lugano, Switzerland
	5 / 4	**Cluster #** 00050931 **Contributor(s)** = CU-S (2) NjP (1) OCIMA (1) MH (1) **Title** = Albino Sword Swallower at a Carnival, Md **Agent.Personal** = Arbus, Diane **Date** = 1970

Details of a single cluster (top of page)

Preferred values are list at the top followed by the Merged values (unique and de-duped values) for each element

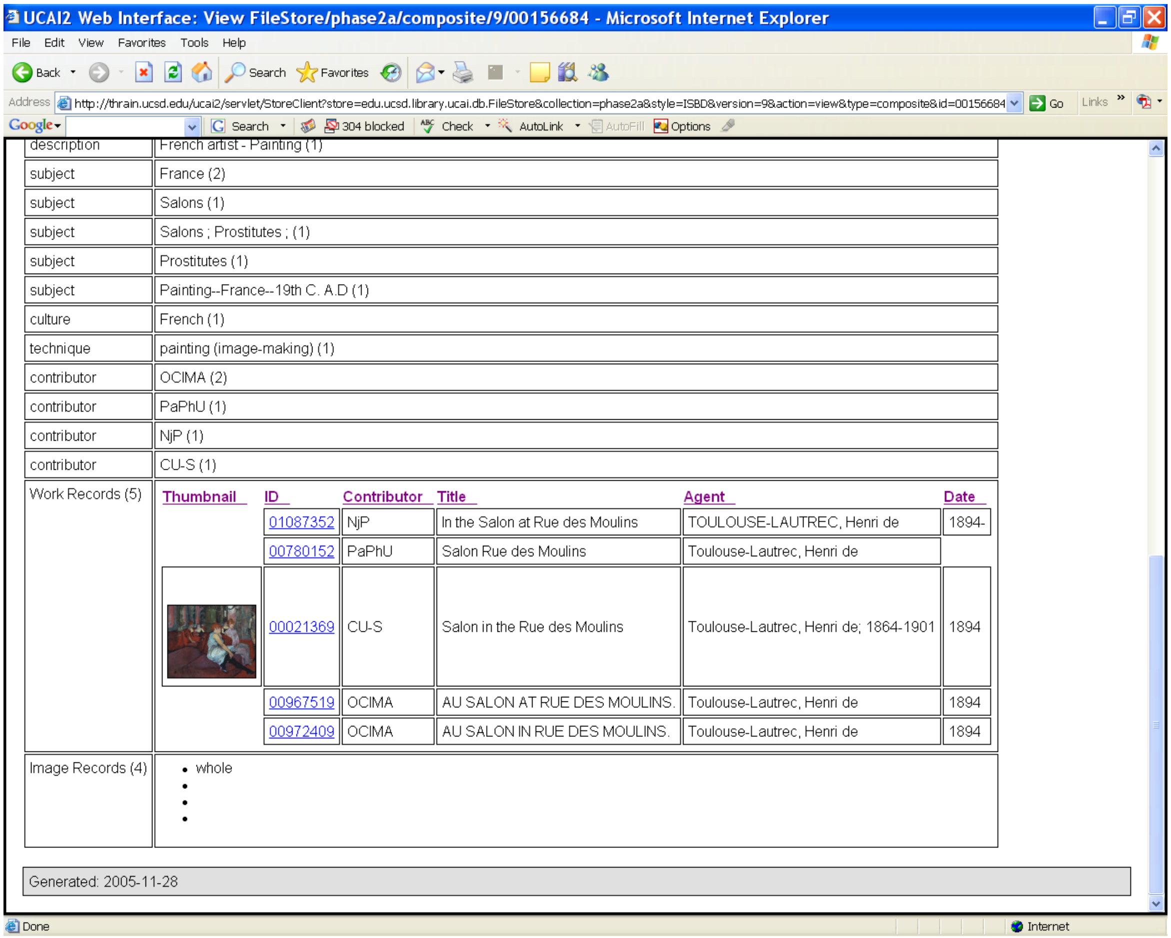

UCAI2 Web Interface: View FileStore/phase2a/composite/9/00156684 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://thrain.ucsd.edu/ucai2/servlet/StoreClient?store=edu.ucsd.library.ucai.db.FileStore&collection=phase2a&style=ISBD&version=9&action=view&type=composite&id=00156684
Go
Links
Google
Search
304 blocked
Check
AutoLink
AutoFill
Options
Done
Internet

UCAI 2: Web Interface: View FileStore/phase2a/composite/9/00156684

Cluster #00156684 [download]	
Preferred Values	
Title	Salon Rue des Moulins
Agent.Personal	TOULOUSE-LAUTREC, Henri de
Date	1894
Merged Values	
title	AU SALON AT RUE DES MOULINS. (1)
title	AU SALON IN RUE DES MOULINS. (1)
title	Salon Rue des Moulins (1)
title	Salon in the Rue des Moulins (1)
title	In the Salon at Rue des Moulins (1)
agent	TOULOUSE-LAUTREC, Henri de [ULAN] (5)
objectType	Painting (2)
objectType	Paintings (1)
objectType	Oil painting (1)
objectType	Painting--genre (1)
location.repository	Albi, MusΘe Toulouse-Lautrec (2)
location.repository	Albi: Musee Toulouse-Lautrec (1)
location.repository	Albi, Musée Toulouse Lautrec (1)
location.repository	Albi, Musee Toulouse-Lautrec (1)
location.repository	Musée Toulouse Lautrec (1)
location.site	France (3)
date	1894 (4)
material	oil on canvas (3)

Details of a single cluster (bottom of page)

Shows Merged values followed by the individual work records that make up the cluster. Clicking on the ID column will take you to the full record for the contributor. Values for the image records are listed last.


UCAI2 Web Interface: View FileStore/phase2a/composite/9/00156684 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
Go
Links
Google
Search
304 blocked
Check
AutoLink
AutoFill
Options
Done
Internet

http://thrain.ucsd.edu/ucai2/servlet/StoreClient?store=edu.ucsd.library.ucai.db.FileStore&collection=phase2a&style=ISBD&version=9&action=view&type=composite&id=00156684

description	French artist - Painting (1)
subject	France (2)
subject	Salons (1)
subject	Salons ; Prostitutes ; (1)
subject	Prostitutes (1)
subject	Painting--France--19th C. A.D (1)
culture	French (1)
technique	painting (image-making) (1)
contributor	OCIMA (2)
contributor	PaPhU (1)
contributor	NjP (1)
contributor	CU-S (1)

Work Records (5)

Thumbnail	ID	Contributor	Title	Agent	Date
	01087352	NjP	In the Salon at Rue des Moulins	TOULOUSE-LAUTREC, Henri de	1894-
	00780152	PaPhU	Salon Rue des Moulins	Toulouse-Lautrec, Henri de	
	00021369	CU-S	Salon in the Rue des Moulins	Toulouse-Lautrec, Henri de; 1864-1901	1894
	00967519	OCIMA	AU SALON AT RUE DES MOULINS.	Toulouse-Lautrec, Henri de	1894
	00972409	OCIMA	AU SALON IN RUE DES MOULINS.	Toulouse-Lautrec, Henri de	1894

Image Records (4)

- whole
-
-
-

Generated: 2005-11-28

APPENDIX E: Technical Description

Data Flow

The first step in the UCAI data flow is to ingest data submitted by contributors as native records, which are XML records that follow the contributor's system as closely as possible. For contributors already using XML (or MARC, which can easily be converted to XML using publicly-available tools), this process is trivial. For contributors using relational databases, however, this process is more complicated. Typically, several tables are used for different types of data (works, images, artists, etc.), and are exported from the database as separate XML files. These files must be merged to preserve the relationships between the different types of data. Also, an XML schema must be written for the merged XML format, in order for it to be recognized by MapForce.

Once a contributor's data are ingested as native records and an XML schema is developed, MapForce (2003 and 2004 versions) is used to generate an XSL stylesheet that transforms the native record into a standard record.

The ingest tool reads native records from the database, uses the MapForce XSL stylesheet to transform them into standard records, and then stores the standard records in the database. In addition, some complex operations are performed both before and after the MapForce stylesheet is applied, to handle complex operations that are difficult in XSL, and to handle limitations of MapForce. Some of these operations include looking up ULAN identifiers and adding them to Agent elements, checking for available thumbnails and adding their URLs to the records, formatting LCSH-style subject headings as strings, and removing extraneous punctuation and other unwanted formatting characters.

Clustering

The clustering tool reads standard records from the database and exports data values into tab-delimited text files. This is done to reduce memory overhead when working with hundreds of thousands of records in memory at one time. The exported Agent names are clustered using ULAN identifiers if available (and last names if not), partitioning the database into Agent groups. Each Agent group is partitioned by Title. These groups are then partitioned using a combination of Date, Location.Repository and Location.Site -- comparing only the values that are present in both records that are being compared. Finally, the final groups are combined into merged records which contain all the standard records and a summary of the frequency of unique values for each element.

All of these tools use a standard interface for storing and accessing data. Implementing drivers for multiple database systems is straightforward. We have developed interfaces for three database systems: (in Phase One) a driver for Xindice (an open source native XML database), Oracle XMLDB, and for storing XML as files in a filesystem hierarchy. For databases that do not provide query functionality (or with query functionality that is too slow or otherwise undesirable), the Lucene search engine library is used to provide a generic query capability. For each type of data, a Lucene index is created that contains three types of indexes. First a fulltext index is created containing all text from a record in a single element. Second, a fielded index is created containing each VRA element in its own Lucene field. Lastly, several elements are indexed in a non-tokenized index used for sorting and browsing; this type of index does not tokenize (i.e., break into words) values.

A web application allows users to search, browse, and list records. It uses Java servlets for the main application code, and XSL for display. It uses the same interface for accessing data and performing queries as is used by the batch tools. Currently, records may be viewed and downloaded in XML format, but no other export functionality was developed.

Search Engine

We used the Lucene search engine for the query capabilities of our web interface. Lucene had a number of advantages, including being open source, written in Java (and therefore portable), easy to implement, and fast. Because Lucene was independent of the underlying database, it could be implemented once and then reused, providing consistency across different databases. And Lucene provides a sophisticated query syntax with wildcards, booleans, and other features with which users are familiar.

However, as the project progressed, we became increasingly aware of Lucene's drawbacks. Because it requires a separate index, the indexing process makes the ingest process slower. It does not have robust support for fielded, structured data, and does not provide a network-accessible query service like virtually all database software does. In addition, possibly because of the limitations of statistical analysis of sparse metadata records mentioned in the clustering section, the default relevance scoring did not work well for UCAI.

Source Code

Java: 13,600 lines of code, including:

- Clustering: 3,900 lines of code
- Database: 4,300 lines of code
- Ingest: 1,850 lines of code
- Test: 400 lines of code
- Web Interface: 1,850 lines of code

Shared Java code (mostly utilities shared with other projects): 4,500 lines of code
XSL: 5,300 lines of code, including:

- MapForce-generated stylesheets: 2,600 lines of code
- Web Interface stylesheets: 2,700 lines of code

Tools used

All tools were written in Java (Sun JDK 1.3 - 1.5, Apple JDK 1.3 - 1.4), using a number of open source tools and libraries:

- Apache Tomcat application server (versions 5.0 and 5.5)
- DOM4J XML parsing library (version 1.4)
- MARC4J MARC and MARCXML processing library (all versions up to b7)
- BerkeleyDB lookup table library (versions 1.5-1.7)
- Apache Lucene search engine library (versions 1.2-1.4)

Because all software is written in Java and XSL, it can be used on many different platforms. During the project, we used Linux (2.4 and 2.6 kernels, various RedHat and Gentoo distributions), MacOSX (10.2 - 10.4), and Windows XP. Typical hardware used for batch processing and web application:

Rack-mounted Linux server

- 2.6GHz P4 Xeon
- 1GB RAM
- Linux 2.4.x

PowerMac

- Dual 2.0GHz G5
- 1.5GB RAM
- MacOSX 10.4

However, the batch tools and web interface run acceptably on a PowerBook (purchased at the beginning of Phase One) used for several demonstrations:

- 867MHz G4
- 768MB RAM
- MacOSX 10.4