Title

Union Catalog of Art Images (UCAI Phase 1): Final Report

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

Publication Date
2004-02-01

Union Catalog of Art Images
(UCAI Phase One)

A Project of the University of California, San Diego Libraries

Final Report to The Andrew W. Mellon Foundation

by

Brian E. C. Schottlaender, Principal Investigator
Linda Barnhart, Project Manager

27 February 2004

CONTENTS

Appendices:

Accompanying CD: UCAI Documentation

EXECUTIVE SUMMARY

We are pleased to report that the UCAI Phase One project has met its goals, accomplishing them on time and within budget. Through the generosity of our partners, Harvard University and the Cleveland Museum of Art (CMA), we were able to significantly exceed the number of metadata records and thumbnail images initially estimated for the UCAI prototype. The UCAI team's key achievements include:

- Develop an innovative prototype database
- Develop data maps to VRA Core 3.0 in XML from three different dataset structures
- Convert 715,000 records and loading them into the prototype system
- Develop work unit and composite record concepts
- Articulate data standardization needs
- Make substantial progress on development of an automated clustering algorithm

The main goal of UCAI Phase One was the development of the technical infrastructure to support a union catalog of art images (UCAI) and in so doing, to demonstrate that a union database is technically possible. *This goal has been achieved by developing the prototype system and mapping, converting and loading 715,000 metadata records from three different sources.*

Secondary goals included:
(1) Demonstrate that visual metadata (i.e., thumbnail images) are a crucial element for image identification and record use (*Goal met; the thumbnail image has been a very valuable data element to UCAI staff in terms of work identification and record clustering*);
(2) Develop generic solutions for mapping diverse record structures to a common standard (*Solutions identified; tool to be built in Phase Two*);
(3) Articulate the difficulties in defining, identifying and understanding the relationships between work and surrogate records, and in working in general with hierarchical records (*Much work done, but more remains*;);
(4) Explore and define the issues surrounding the clustering of work and surrogate records, including determining algorithms for automated clustering (*Strong beginning, but more work remains).*

As the UCAI prototype developed, the staff spent time grappling with the barriers, on both a theoretical and a practical level, to creating a union catalog for art images. Some of these impediments, listed below, were known before the project started; others surfaced during the project. They include:

- Current cataloging environment does not support interoperability
- Image metadata is formulated to meet local needs
- Sufficient metadata may not be present to do what is needed for a union catalog

- Cataloging practices are inconsistent
- Authoritativeness of data is questionable
- VRA Core 3.0 has shortcomings
- Need for record synchronization between a union catalog and local catalogs
- Absence of unique identifiers for works
- Explicating the definitions of work and surrogate
- Difficulty in using legacy metadata

Phase Two will address the immediate next steps, however further work remains.

1. BACKGROUND

The Union Catalog of Art Images (UCAI) project began an eighteen-month period of research and development on April 1, 2002. The project was given a no-cost extension to continue development through December 31, 2003. Core team members included:

Linda Barnhart	Project Manager
K. Esme Cowles	Database Developer
Joseph Jesena	Data Administrator
Bradley Westbrook	Lead Designer
Tricia Rose	Design Specialist

Additional support at UCSD was provided by:

Brian E. C. Schottlaender	Principal Investigator
Vickie O'Riordan	Content Specialist
Gavin Hurley	Technical Implementation Supervisor (April 2002 – August 2003)
Chris Frymann	Technical Implementation Supervisor (August – December 2003)
Arnold Josafat	Administrative Assistant (April – October 2002)
Tina Nguyen	Administrative Assistant (October 2002 – December 2003)
Richard Caasi	Server Administrator
Lauren Silverman Wilson	Image cataloger
Sherri Panian	Budget support (April – October 2002)
Allyson Collins	Budget support (October 2002 – December 2003)

as well as by other senior administrators and support staff.

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

Sara Jane Pearman	Content Specialist, Cleveland Museum of Art
Frederick Friedman-Romell	Technical Specialist, Cleveland Museum of Art
Martha Mahard	Content Specialist, Harvard University
Ardys Kozbial	Content Specialist, Harvard University
Robin Wendler	Technical Specialist, Harvard University

The UCAI team held two meetings of the partners and invited consultants during Phase One. The first meeting was held May 20-21, 2002 and was designed as a kick-off meeting to explore community needs and system functionality. The second meeting, held March 3-4, 2003, brought the partners together to assess progress and confront the many questions that surrounded record clustering. Our invited experts included:

Sherman Clarke	New York University
Karen Coyle	California Digital Library
Laine Farley	California Digital Library
Tony Gill	ARTstor
Ed Glazier	RLG
Eric Li	ARTstor
Lauren Meserve	ARTstor
Emerson Morgan	ARTstor
Thomas Nygren	ARTstor
Guenter Waibel	RLG

The main goal of UCAI Phase One was the development of the technical infrastructure to support a union catalog of art images and to demonstrate that a union database is technically possible. This goal has been achieved by developing the prototype system and mapping, converting, and loading 715,000 metadata records from three different institutions. Secondary goals included:

(1) Demonstrating that visual metadata (i.e., thumbnail images) are a crucial element for image identification and record use. The need for a thumbnail as a data element was convincingly demonstrated in our clustering work, as data analysts found that thumbnails provided a quick and easy method to determine the appropriateness of works and surrogates within a cluster. By logical extension, thumbnails would quickly help catalogers accurately identify records that match the work being cataloged.

(2) Developing generic solutions for mapping diverse record structures to a common standard (the VRA Core). We deliberately chose diverse datasets, knowing that would force us to develop experience and expertise in dealing with visual resources records. We developed customized data maps for the three Phase One datasets, and have articulated data standardization issues and identified techniques that will assist in developing a more generic mapping and ingest tool.

(3) Articulating the difficulties in defining, identifying, and understanding the relationships between work and surrogate records, and in working in general with hierarchical records. We have made headway in this area but considerable work remains to be done. Working with Harvard's hierarchical records in combination with UCSD's and CMA's flat records was challenging. Separating flat records into work and surrogate records by analyzing and identifying inconsistently coded data elements was difficult and frustrating. Further development will be needed to define and describe the relationships between works (one work as a part of another, one work inspired by another, etc.)

(4) Exploring and defining the issues surrounding the clustering of work and surrogate records, including determining algorithms for automated clustering. The UCAI team

made substantial progress in this area and began experimenting with automated clustering algorithms. While we've made a strong beginning, more work remains to be done in this area.

2. ACHIEVEMENTS

We are pleased to report that the UCAI project has been on track, on time, within budget, and has met its goals. Through the generosity of our partners, we were able to significantly exceed the number of metadata records and images initially estimated for the UCAI prototype. Key achievements include:

Developing an innovative prototype database. After substantial analysis, the team developed an initial prototype based on the open source native XML database Xindice and the open source search engine Lucene. When fundamental problems with Xindice were discovered, the prototype was modified to also use Oracle's XML database and search engine functionality. The prototype systems have a Web interface with basic search and display functionality (including images) and the ability to view the originally submitted record in XML. Both the Xindice and the Oracle databases include 715,474 unclustered standard records and 260,531 related thumbnails.

Developing data mapping to VRA Core 3.0 in XML from three different dataset structures. UCAI data analysts designed a UCAI Standard Record, based on the VRA Core 3.0 data standard. We found that we needed to extend the Core to prevent unacceptable data loss, and have communicated our needs for Core extension to the VRA Data Standards Committee. UCAI analysts then developed customized maps from the three very diverse datasets—one MARC, one SGML, one relational—to the UCAI Standard Record.

Converting 715,000 records and loading them into the prototype system. Using the mapping structures created by the project analysts, UCAI technical staff developed an ingest system for record conversion and storage, and a process for indexing and retrieval. Links to thumbnail images and original records in XML are preserved.

Developing the concepts of the work unit and the composite record. Analysis of the datasets and the potential duplicate records has led the team to do some fundamental conceptual thinking to define the notions of works and related works represented by images. Such foundational work is important to display clearly the hierarchies and relationships between images, surrogates, and their various perspectives, views, and details. As the team struggled with large and conflicting amounts of data, the notion of a composite record began to take shape that could merge identical (or similar enough) values and make lengthy records comprehensible.

Articulating data standardization needs. Through careful analysis of the data, the UCAI team identified many problems inherent with non-standard legacy data of this magnitude. Some inconsistent data can be fixed through the mapping process and some can be fixed through normalization routines, but it was considered important to maintain the integrity of the submitted data and not to fall into the morass of fixing individual records within the UCAI system. Long-term goals are to enhance data interoperability among multiple datasets and to promote standardization within the visual resources community.

Moving toward a successful automated clustering algorithm. Identification of duplicate records as well as records for related works and surrogates is a substantial problem in datasets using differing or inconsistent content standards. The team designed a set of empirical and iterative clustering experiments to assess various clustering algorithms on various database subsets. We examined clustering using both textual elements (focusing on title, creator, and medium) and on image analysis (using the Levenshtein algorithm of content-based image retrieval).

3. Prototype Development

An overview of the UCAI prototype system follows, including system architecture, record structure, major processes, and prototype functionality.

a. Architecture

i. System

Initial specifications for the prototype focused on understanding user needs and expectations and designing a system to meet a crucial subset of those needs. Since this was a prototype, it did not need to scale to production load. This provided freedom to experiment with emerging technologies. The specifications therefore reflect fairly high-level requirements, without dictating lower-level technological decisions. The interface development was informed by the internal need for tools to review the output of our database processes—not by users. As both the developers and analysts used the prototype to review conversion and clustering output, reporting and statistical functionality was developed.

Because the initial data formats from the three partners were deliberately and substantively different from each other, the first technological decision was to convert all incoming data to XML in order to develop tools able to handle data from all sources in a uniform manner. This quickly led to the use of XSLT for most data processing. Java was chosen as the primary development environment, using the Xerces and Xalan XML processing tools. Java enables development in both a server environment (Linux) and in the developers' desktop environments (MacOSX and Windows XP).

With the basic technological choices made, the next step was choice of a database. Since we were committed to using XML, an XML database seemed like a logical possibility. There is a wide variety of native and relational-based databases available, so we decided to evaluate the capabilities and performance of a few leading candidates. After reviewing the available choices, we decided to focus on native XML databases, as these seemed to leverage our use of XML.

Tamino, the leading commercial product, was the obvious choice for a native XML database. Xindice, based on code recently donated to the Apache project, was the leading open source database. TeraText also looked promising as the only product to specifically advertise scalability and performance as key features. Early in the process, we realized that Xindice's built-in query capabilities did not scale well so we began to use the Lucene fulltext search engine as an alternate query engine. We also tested an alternative approach to storing XML: storing the XML files as text in a relational database.

We developed basic query tools for each of the databases and tested the performance of several different queries. We found a lot of variance among the different databases. Tamino, Xindice (using native query engine) and storing the XML files as text in a relational database were ruled out on performance grounds; they didn't scale well enough even to use for a prototype. TeraText and Lucene/Xindice offered roughly equivalent performance, with each performing all queries in 5 seconds or less.

Since Lucene/Xindice and TeraText performed equally well, other factors were considered in making the final decision. Lucene/Xindice was free and open source; TeraText was expensive and proprietary. Xindice implemented the emerging standard XML:DB interface (a well-designed, object-oriented API); TeraText required using a Z39.50 client or using their proprietary scripting language for development. Xindice was much easier to setup and configure and ran on more platforms (e.g., Linux, MacOSX). Using Lucene/Xindice left open the option of purchasing another commercial product later.

This latter possibility proved to be prescient. Lucene/Xindice performed well, but two problems emerged: unreliable memory management, and character encoding problems. The memory management bugs resulted in the server running out of memory or slowing down, frequently requiring restarting the database or the entire server. The character encoding bug resulted in all modified characters being garbled when retrieved from the database. While tolerable in a prototype system, these problems made Xindice questionable going forward.

When we revisited the state of XML databases, we found that Oracle's XML database functionality had been greatly improved in the most recent version. The improved functionality combined with our confidence in Oracle as a database platform led us to switch to using Oracle as our primary database. This was relatively easy since the software was developed in a modular fashion. In the long term, we believe Oracle will provide possibilities for scalability and stability that would not have been feasible with Xindice.

ii. Record types

At this time, four record types have been established for use in the UCAI database. They are institutional records, native records, standard records, and composite records.

UCAI institutional records are used to collect and store data about repositories contributing records to UCAI. There is one institutional record for every contributing repository. Presently, the institutional record is made up of ten elements that identify the institution (name, address, telephone), the person designated by the repository to serve as the contact for UCAI, and the terms, if any, that govern use of the contributed records. In addition, institutional records are referred to from the composite record in order to provide quick and ready contact information should a viewer of the database want to acquire more information about a particular record or resource. Eventually, institutional records could, with additional data elements, be used to preset repository profiles for interacting with UCAI. For instance, a repository might wish only to see records from a selection of repositories represented by records in UCAI. The institutional record could be used to set such a preference.

While institutional records are used for managing information about repositories, the other three record types are used for managing and processing the resource metadata contributed to UCAI. **UCAI native records** are XML versions of an institution's visual resource metadata before any processing is done by UCAI staff. The native record contains all data that appears in the original contributed record, including element names in their native format. Because some of these data are not used in the UCAI standard record, native records are retained in the UCAI database and are linked from the corresponding UCAI standard record in the event a viewer wants to determine how information was mapped to the UCAI standard record, or to confirm if the original data were modified in any way or omitted through their mapping to the UCAI standard record.

The **UCAI standard record** (see Appendix B) is the architectural backbone of the UCAI database. Also an XML record, it is the basis for the clustering and merging processes and for constructing the composite record. The standard record is comprised of the 17 data elements defined in the VRA Core 3.0 element set. In order to accommodate administrative data and preserve data granularity in some native records, the UCAI standard record added to the VRA 3.0 Core one additional element and twenty qualifiers. For example, the element "Subject" has no specified qualifiers in the VRA Core 3.0 element set, but in the UCAI standard record it can be further subdivided by "Personal Name," "Corporate Name," "Topic," "Geographic," "Period," and "Authority." Defining specific qualifiers allows the richer granularity of subject headings—for example, the 6xx fields in MARC records—to be preserved and utilized in UCAI.

The UCAI standard record is used for both works and surrogates. A work record describes an original object or resource—a sculpture, painting, or piece of pottery. A surrogate record describes a representation of that object or resource. Typically, the representation is a visual rendering in the form of a slide or digital image. In many instances, the distinction between work and surrogate is blurred in the contributed record metadata and, to the degree possible, must be distinguished in UCAI.

A work record and its corresponding surrogate records comprise the UCAI **work unit,** a cluster constructed from UCAI standard records using an automated clustering algorithm. In the UCAI prototype, each work unit must have at least one work record; it may have any number of additional surrogate records. The UCAI team will continue to explore in Phase Two the best way to link and display the various relationships between work units, as well as organizing mechanisms for the often voluminous surrogate records.

The **UCAI composite record** is assembled from the UCAI work unit through a record merging algorithm. The composite record is intended to simplify the display of complex and lengthy records to make them intelligible and useful to catalogers. As duplicate records are merged into a composite record, the most common value for a given element is retained and displayed, with the unique values accessible through a link. This shortens the display, for example, of multiple date fields, some identical and some with minor (or major!) variations. Because data elements for the composite record may be drawn from different work records, the UCAI composite record is a unique concept, distinct from the OCLC master record model (one record with all of its data elements chosen to be the master) and the RLG primary cluster model (records clustered together with one entire record given primacy). In the UCAI database, the model is applied to both work records and to surrogate records.

The presence of variant data and the absence of authority control necessitate a procedure for determining a preferred or display value--the value that is displayed in the composite work record. UCAI is considering two different options. One is to use the data value that is first submitted to UCAI until it is rejected and/or modified by the community of visual resource catalogers. The second is to use the most frequently-occurring value in the work unit. Each option presents concerns about accuracy and utility of the selection for any given work unit. The final option will become apparent when the clustering work nears conclusion and development of the merging process begins.

We expect to merge metadata for surrogates within a work unit in the same manner. This would be very useful for work units composed of a large number of duplicate records for the same surrogate. However, we are not sure that the surrogate metadata is of a sufficient quality for matching duplicate records for the same surrogate. For a great number of cases, surrogate information is present in the record only as a type statement (e.g., “slide”) and some supplementary title information.

b. Processes

i. Mapping and converting

The intellectual process for mapping and converting each partner’s data required considerable time. Although the differences in record structures did not come as a surprise, we did not anticipate the degree of diversity across institutions nor the disparity in content standards relating to how data were recorded.

Structural differences – flat vs. hierarchical records. Like many institutions, both UCSD and CMA take a flat record approach to cataloging in which both the work and surrogate information are conflated in a single record. Harvard has taken a less common hierarchical record approach which, like VRA Core, puts the work and surrogate information into separate records. Within the Harvard dataset, this structure results in a cluster. In some cases (approximately 4%), Harvard records contain a third level of hierarchy in which collection information is recorded.

Because the predominant structure within the visual resources community follows the flat record model, it made sense for the UCAI team to devise a technique for splitting flat records into works and surrogates to be able to build work units. The hierarchical record model is less common, and the underlying principles for forming clusters (at Harvard, or at any other institution using this approach) may or may not be the same as those for UCAI. Because it was necessary to homogenize, we began by flattening the Harvard records. This had the effect of equalizing the datasets and did not privilege Harvard's pre-established work clusters. This technique also allowed the Harvard surrogate records to be members of work units outside of their Harvard-established groups. We plan to test the opposite approach—leaving the Harvard hierarchical work units intact and matching the other records to them—in Phase Two.

Depth of information. The depth of descriptive information from each partner institution varied widely. In some cases, records were extremely minimal. In other cases, records were so extensive that we needed to extend VRA Core 3.0 elements and qualifiers so we would not lose important data. The UCAI standard record was established to identify the elements and element qualifiers being used within the UCAI system. As mentioned above, and in addition to the standard VRA Core 3.0 elements, some elements were added for administrative purposes, such as tracking the date the record was imported into UCAI and recording the unique Record Identifier. Qualifiers were added to some of the VRA Core elements (e.g., Vital Dates and Nationality to the Creator element) in order to carry over contributor data that did not fit into the existing VRA Core 3.0 qualifiers.

Content Standards. Contributed records were extremely diverse in how data were recorded and formatted. Even within a single institution data were often not recorded in a uniform fashion. These differences were most problematic in the fields needed for clustering, such as title, creator, and dates. The following two record examples are for the same surrogate:

David /det. upper torso	*David ¾ view*
1623-4	*1623-1624*
Bernini, Giovanni Lorenzo	*Bernini, Gian Lorenzo*

Note that one institution records its image titles at the end of the work title separated by punctuation while the other records its image titles in a field separate from the work title. For the purposes of clustering and display we appended image titles to the end of their work title field. We sought to normalize dates where possible so that the above dates, for example, would be read as equivalent. Creator names presented an even greater challenge. We did not want to change partner data to a "preferred" spelling of a name (indeed, how would we determine the preferred spelling?). We determined to resolve this problem through synonym rings, in which various spellings of a creator name could be linked and considered equivalents. A search on any variant would retrieve all records for that creator. The efficient creation of such rings would involve the incorporation of controlled vocabularies into the UCAI system. Because this is a complex and potentially time-consuming feature to implement, we plan to address this issue in Phase Two.

As part of the mapping and conversion process we also recognized we would need to do a certain amount of data standardization within fields that would otherwise present impediments to our clustering and merging processes. We first categorized the types of standardization needed in order to determine the best way to deal with these problems. Ideally, standardization should be dealt with in the contributor’s system and then re-imported, but often the problems that we identified with the data were not always considered problems in the contributor’s system.

Standardization Types and Examples:

- Parse: When there was no 1:1 relationship between a contributor’s field and the UCAI Standard Record field, a manual review of the data was necessary to identify the appropriate UCAI element(s) and move the data. For example, a single field from the contributor might contain data about culture, subject, type, and period—data that need to be manually reviewed and parsed out to appropriate separate fields. While this approach was feasible for a prototype, it may not scale for a production system.

- Normalize: Reducing variations of an expression to a preferred form. For example, CMA data reveal 41 different entries for US and American. Ideally, these should be standardized to a single term based on an authority, such as the TGN. In a few cases we normalized data to a single term to assist with clustering. Because of concerns about changing contributed data and the workload involved, we did not think it was a good use of our time to normalize all variants.

- Remove extraneous data: Removing text and format characters that have no relevance outside the contributing institution (e.g. brackets, codes, cataloger’s initials). We manually reviewed all data for these problems. If the information was coded consistently, we were able to create programs to automatically remove it.

- Correct mis-spellings and factual errors: Analysis of contributed data inevitably revealed misspelled terms (e.g., Picassi) and factual errors (e.g., Da Vinci as an American artist) were identified. We decided not to correct these problems, as we had neither the time nor the resources to do so. Such problems could be corrected in the contributor’s system and re-imported into UCAI, or could be left for community correction.

Techniques and Tools. In order to make the standardization process more efficient and consistent we relied upon two automated techniques: algorithmic and dictionary. The algorithmic approach enacts a program or script that automatically checks and fixes data for identified problems. The advantages of this approach are that it is more efficient than fixing individual records, and is in principle repeatable and extensible to multiple datasets. However, one of its weaknesses is that the variance in data content and structure across institutions requires algorithms to be derived for each contributing institution.

The second automated technique is the dictionary approach. Once problems are identified, a dictionary of values is created with rules for how to change each value. This approach brings precision to changing data values. That precision can also be considered a weakness because it may not be flexible enough. In addition, the labor-intensivity of developing and implementing a dictionary of synonyms is of concern.

ii. Clustering

Grouping together all metadata records for a work and its surrogates is one of the most fundamental challenges for UCAI. In the visual resource community, a work is an object—typically unique—such as a painting, a sculpture, a building, a tribal mask, or a tea set. A surrogate is a representation of a work. Typical surrogates include slides or photographs and, more recently, digital images. A metadata record is a description of a work, its surrogate(s), or the work and its surrogate(s) together.

The UCAI group began to look at clustering as a means of compensating for differences in descriptive practice, of bringing together works widely dispersed due to variant practices, and of saving the cataloger time in browsing through redundant displays. There is an identifiable corpus of standard masterworks common to many, if not most, image collections, and we expect to see records for these from each UCAI contributor. Even within a single collection, there are often multiple nearly-identical records representing multiple nearly-identical slides. To identify these in an automated fashion (despite dissimilar and sometimes conflicting data values) and bring them together (at the same time sorting out related works, details, perspectives, and surrogates) has been UCAI's biggest challenge.

Efficient and accurate clustering is greatly aided by data standardized in content and format. The converting and mapping processes UCAI applied to ingested metadata were designed to standardize the format of the metadata and to normalize, where possible, the data values that would affect clustering. For the most part however, inconsistencies and errors in data values were allowed to stand, as they were seen to be characteristic of metadata produced by a community that has not had the benefit of community-wide data standards. Moreover, normalizing the data values would have required additional resources.

To build work clusters, or what we call a "work unit" (at least one work record and any number of surrogate records), the UCAI staff opted for a two-part, hierarchical clustering method. In the first part, all individual metadata records were compared and grouped together using a series of steps collectively referred to as the clustering algorithm. The clusters achieved at the end of this process are work units, but their parts, the surrogate records, are not organized in any way. The second aspect of the clustering process, then, is to organize the surrogates. The result of the second part of the process is what we call an "articulated" work unit, a work unit in which the surrogate clusters are formed and organized.

In order to build the UCAI clustering algorithm, the UCAI staff identified objective measures that would allow us to determine whether changes to the algorithm were improvements or setbacks. We set up a series of four carefully controlled experiments so we would be able to gauge progress, and used a subset of the full database so that system performance and data analysis would be manageable. One key area that we discussed at length was defining and using system-calculated similarity measures for comparing the records and constructing the work and surrogate units. Title was considered to be the single most important identifying feature for a work, followed by creator, date of creation, classification number, and type. For surrogate clustering, surrogate title (when it existed) was considered most important, followed by type, creator, date of surrogate, and source (location whence where the image was taken or captured).

After subsequent reflection, it was decided that the classification number would not be particularly useful since one of the data sets did not use classification numbers and the other two, while employing the same base classification system, applied it differently.

We also determined that the bibliographic source for the image (the book title, edition, and page number for copy-stand photographic reproductions, for example) would not be a good clustering measure, despite our hopes to the contrary, because the recording of source information was highly idiosyncratic in form and inconsistent from record to record. Using this element for clustering would require substantial work for little return, given that only a small percentage of the images included source information.

The UCAI team also recognized the potential value of the object identifier that, because of its uniqueness, could expedite both work identification and clustering. Unfortunately, this data was not present. However, if institutions can be persuaded to make these identifiers public, and catalogers would be willing to research and code this data element in their records, it would be an invaluable mechanism for associating works and surrogates.

It was clear that there were several ways to build algorithms for building work clusters. Since it was not clear which method would be most efficient and accurate, four different possibilities were tried. They are:

1: Based on a flat file of metadata, an algorithm that groups by titles first and then filters the title groups using, in sequence, the similarity measures of creator, date of creation, and type.

2: Based on a flat file of metadata, an algorithm that groups by creator first and then filters the creator groups using, in sequence, the similarity measures of title, date of creation, and type.

3: Based on a flat file of metadata, an algorithm that applies simultaneously in a round-robin comparison the similarity measures for title, creator, date of creation, and type.

4: Based on the hierarchical data set procured from Harvard, an algorithm that would cluster UCSD and CMA metadata records according to the work units already established in the Harvard data set (unique UCSD and CMA records--not matching Harvard's work units--would have to be clustered using an algorithm built in part 1, 2, or 3 of the experiment.)

The UCAI team will evaluate these experiments objectively to determine which produces the best clustering results. We strive for the ideal of no inappropriately dispersed work units and, conversely, no inappropriately conflated work units.

Two teams of a designer and a programmer each were assigned to develop the first and second algorithms. Each team constructed its algorithm in segments and iteratively. A baseline title and creator grouping was done first. The groupings were then refined by adding parameters to the matching command. Some of those parameters included: restricting match to the first few words of the title, separating work title from surrogate title where possible, removing from consideration articles and common prepositions, devising synonym relationships, ignoring word order, permitting inexact matching of words (e.g., log vs. dog). Parameters were added to the matching command one at a time and their effect measured by processing a test data set, analyzing the resulting clusters, and comparing the results to earlier results.

In the early builds of the algorithms, the test data set was composed of approximately 1500 records all having the string "Aphrodite" or "Venus" in either their Title or Title.Largerentity elements. In subsequent builds, the algorithms were applied to an approximately 10,000 record set united by the presence of the strings "mother," "Madonna," "Mary," and "virgin" in each record's Title or Title.Largerentity elements; and to two random sets of records, one including 25,000 records and the other 50,000.

The experiment is still ongoing with construction of the first and second algorithms approaching conclusion. Although the experiment is not complete, several discoveries have been made in the iterative building of the first and second algorithms. Chief amongst these were several types of impediments to accurate clustering that are present in the data. For title clustering, impediments include inconsistent internal punctuation, variant internal prepositions, inconsistent use of diacritics, inconsistent word endings (e.g., -s, -ed, -ing), and unclear demarcation between work and surrogate title information. Impediments in the creator segment of the clustering also included inconsistent use of diacritics, in addition to inconsistent use and positioning of title information and variations in the fullness of name.

Many of these kinds of impediments were accounted for in the title clustering algorithm by adjusting it to ignore initial articles and internal prepositions and other common words, by normalizing strings to remove diacritics from the matching process, or discerning parameters by which, in a majority of records, surrogate title information could be separated from work title information. Other kinds of impediments could not be so easily adjusted for. These included the problems of word order, synonyms, geographic term as the first word in the title, inconsistent use of date of creation, and outright input error such as typographical errors and misspellings.

The project team considered each of these impediments singly. As mentioned above, the group decided it would have to accept input errors, as it was beyond the scope of the project to identify and correct all such errors. Also, the team concluded that exposing such errors would help contributing partners understand the importance of implementing effective quality assurance processes in their production of metadata. The team experimented with constructing rudimentary synonym rings within titles. The results were favorable in that such devices made it possible to cluster titles such as “Aphrodite crouching” and “Venus crouching” or “Venus and Cupid” and “Venus and Amor.” However, the team finally concluded that constructing synonym rings did not scale to the overall database, as it would require identifying all synonymous relationships needing to be accounted for as well as developing a process or accommodating new extensions to the relationships as new metadata was ingested into the database. As a consequence, the team decided to accept the split clusters due to the presence of synonyms.

Dealing with the date of creation was more problematic. The team thought variations in date of creation statements could be managed by normalizing all dates and then matching on certain dates and date spans. However, a large number of clusters (15%) contained records with either conflicting dates or no dates at all. This forced a record to split off from a cluster in which it was grouped by virtue of title and/or creator. These discoveries forced the team to conclude that the date of creation was a much weaker similarity measure than it had originally speculated.

As the size of the experimental subsets increased, it became clear that some automated tools or processes would be necessary to perform the analysis of a dataset in order to characterize and compare it with another dataset. The three measures determined to be most useful were:

(1) compression/dispersion of a dataset as a measure of clusters created per total number of records in the dataset,

(2) compactness of a cluster taken as a measure of average distance from a 100% match within a cluster, and

(3) isolation of a cluster taken as a measure of its exclusivity from other clusters in the dataset.

Taken together, the three measures appear to be a relatively strong set of measures for characterizing the quality of the clustering process.

iii. Merging

In order to improve the display of clustered work units, we developed a preliminary merging algorithm to create rudimentary composite records. The algorithm chooses preferred values for the title, creator, and date fields. The values are chosen by making a list of all values in a set of records, comparing each value to every other value, and choosing the value that has the highest average score (referred to as the "confidence value"). This confidence value is used as a rough indication of cluster uniformity to analyze the performance of the clustering software. It is also used to identify clusters that require manual review to identify areas where the clustering algorithm needs improvement.

Currently, the preferred value for each field is selected in isolation. This often results in the values being selected from different records, rather than identifying a single representative record. Display of the merged records currently consists of a header listing the preferred title, creator and date values, and a list of records in tabular format.

c. Functionality

It should be pointed out that functionality of the UCAI prototype was designed for the UCAI programmers and data analysts, not eventual end users.

Ingest, mapping and converting. The prototype contains Java classes and XSL stylesheets for converting the three datasets provided by our partners: MARC, VIA SGML, and FoxPro XML. The software processes a file of bibliographic records, converting each record to XML, and merges any accompanying accession or artist records. This merged XML format, which maintains the structure and semantics of the source data as faithfully as possible, is stored as the native record. Each native record is converted to VRA Core 3.0 format and stored as the standard record.

In addition to the mapping from the source format to VRA, a number of standardization routines have been developed. These include removing data which are meaningless for a union catalog context and mapping values from general fields (e.g., Description) to more specific fields (e.g., Technique) based on the value's content.

Query. A Web application allows users to query the database with fielded searching, fulltext searching, boolean operators and wildcards. The search results can be presented either in relevance order or in ID order. The ID, contributing organization, title, creator, and date values (and thumbnail image if available) are displayed for each record. Detailed views of each record are available, or the search results can be paged through or downloaded as a single XML file. In addition, the query can be refined by changing any of the search terms, or by adding new terms.

Browse. A simple browse interface is provided by listing the indexed terms in each VRA category (Title, Creator, etc.). The list of terms for each field (or for all fields) can be paged through, and a search can be performed to retrieve the records matching any term.

Record Display. The detailed display of an individual record lists all of the standard VRA data elements, and the administrative metadata such as the ID, contributing organization, thumbnail image, and conversion date. In addition, both the standard and native records can be viewed as XML files.

Clustering. As part of our investigation of clustering techniques, we have developed two clustering applications. Both perform the same task—grouping similar records—but differ in the order that different fields are processed and the specific algorithms for comparing titles and creator names. The clustering applications process a database of standard records, perform a search for records with similar titles and creators, and compare the source record to the search results. The comparison scores are used to group matching records into clusters, which might then be sub-divided into finer-grained clusters. Preferred values for the principal values (title, creator, and date) are then chosen.

The Web interface allows viewing an index of the clusters and a detailed view of each cluster with the preferred values and all of the individual records in the cluster. In addition, statistics about the number of clusters, number of records in each cluster, and degree of confidence in the preferred title value are displayed.

Reports. The prototype contains several reporting and statistical features. As mentioned, the clustering process generates statistics about cluster composition. There is a Web application to view the basic statistics for a database, such as the number of native and standard records contributed by each institution. In addition, there are several applications to analyze the database contents or selectively retrieve data for further analysis, such as generating a profile of which fields are populated, which records lack titles, or how many unique values are present in a given field.

4. IMPEDIMENTS TO A UNION CATALOG

In developing the UCAI prototype, project staff learned a great deal about the barriers that stand in the way of record sharing in the visual resources community. We identified the following factors:

Current cataloging environment does not support interoperability. Today's universe for visual resource catalogers is a strongly rooted original cataloging culture, with similar work being redundantly repeated across institutions. However, though the work is similar, the data structures, syntax, and semantics differ widely from one site to another. Recently, there has been progress toward interoperability through the promulgation of the VRA Core data standard and the Cataloging Cultural Objects (CCO) guidelines, but acceptance and actual change in practice have been slow. Upcoming programs at the VRA and ARLIS/NA national conferences on “Preparing for Shared Cataloging”—in which UCAI staff will participate—will help raise awareness in the community about the need for standards and consistency. A union catalog of art image metadata in itself is a resource that will promote, if not force, interoperability.

Image metadata is formulated to meet local needs. Within an institution, data elements are chosen and designed to meet local needs, and data values are determined based on local expertise and needs. For example, one institution might use only the English form of an artist’s name, while another uses the form (with dates) in the native language of the artist; each is valid in its own universe, but makes wider sharing of data more difficult. Because record sharing has not been possible, the community has no incentive to standardize practices. The work on UCAI, VRA Core, and CCO will reinforce within the community that there can be common solutions (which can accommodate local needs) that can be leveraged for the greater benefit of all. A “mindset of sharing” needs to be nurtured in the visual resources community.

Sufficient metadata may not be present to do what is needed for a union catalog. Some records, particularly surrogate records, have minimal data, with missing elements and/or unknown values. The UCAI team and the community need to adjust their expectations for a union catalog from legacy metadata. Certainly the metadata will be expanded and adjusted over time, but it will not be possible to achieve perfect records or perfect clusters.

Cataloging practices are inconsistent. The biggest barrier to effectively mapping and clustering UCAI data had to do with cataloging practices that were inconsistently applied within institutions. Descriptive and coding practices need to be consistent over time, both within an institution and (ideally) between institutions. Inconsistency is a problem that is not unique to visual resource catalogers; it plagues every database in every institution. However, for a variety of reasons, inconsistent practices seem to be a particular problem for this community. At least one of the partner institutions did not have a cataloging guide for us to refer to in performing data analysis. Even when a guide did exist, the recorded data often differed from the stated practice. As a result, the mapping process relied very heavily on detailed analysis of a significant number of records in order to assess the primary patterns of cataloging. The team feels consistency in cataloging practice is one of the most important issues in the visual resources community today. As stated by Karen Coyle in 2000, "If your content is good, if it is consistent, you have what you need to feed different record formats or different systems. If your content is not based on standards and if your coding of the content is irregular, no record format can save you." (http://www.kcoyle.net/marcdead/marcdead.html)

Authoritativeness of data is questionable. Because image collections are responsive to local needs and local expertise, we expect that questions will arise about the authoritativeness of the metadata in a shared environment. Unlike the bibliographic world, where one can transcribe a title page and record a publication date, metadata elements are more fluid in the image universe. There will no doubt be “trust issues” for certain records or categories of records. For example, if we pursue the composite record concept, and choose the most frequently occurring elements for initial display, will these elements be the “correct” or “true” data? How does one encode expertise (or the latest research finding) about a specific element into a record? Should a union catalog privilege the metadata from the museum that holds the art object? Can the multiple purposes for which each record will be used be served by a single set of data values?

VRA Core 3.0 has shortcomings. This data dictionary is a good starting point, and it is useful to have the community reach agreement about what would comprise a common metadata structure. Developers of VRA Core 3.0 intentionally left issues of specificity and system implementation up to the local institutions. Implementing VRA Core 3.0 in a union catalog environment revealed weaknesses in two areas. First, it lacks specificity. UCAI staff needed to add both elements and qualifiers so the richness of contributed records would not be lost. The UCAI standard record structure in Appendix B provides details about our additions, information that has been provided to the VRA Data Standards Committee (developers of the VRA Core standard). Second, a formal XML schema needs to be prepared so that the Core can be consistently implemented and validated. UCAI staff have been actively involved in this effort.

The need for record synchronization between a union catalog and local catalogs. For a shared source of cataloging copy to be effective, there must be efficient automated ways to add, update, cluster, and merge records. It would be a more effective long-term strategy, for example, to have catalogers fix errors in one system and have the record transferred and overlaid in the other (rather than make the same correction in two places). However, there are complex data standards issues that will need to be resolved before such synchronization can be developed. In some cases, it will be appropriate for data to be corrected locally and imported into UCAI. In other cases, it will be appropriate to correct data only in the UCAI system. Mechanisms will have to be put in place so that data standardization done only in the UCAI system will not be overwritten with records from the contributors. These issues will have to be resolved before UCAI can move to production.

Absence of unique identifiers for works. Because it is often difficult to identify works from either the available metadata or images (imagine 200 similar silver teasets), a unique identifier for that work would play a vital role in helping distinguish them. Often the identifier is the only way to establish if two records are referring to the same work. For UCAI, clustering could be simplified if works and surrogates could match on object identifiers. Historically the visual resources community (other than museums) has not seen this as a particularly significant piece of information to record about object surrogates. It is therefore important for the community to adopt an object identifier standard, to share identifiers, and for catalogers to record this information.

Explicating the definitions of work and surrogate. While there are brief definitions of “work” and “image” (what UCAI has been calling surrogate) as part of VRA Core 3.0, they are inadequate for UCAI’s purposes. More detailed definitions are needed to have an implementable union catalog. For example, there is one class of surrogates that are "visual representations of a work," such as a slide, a photograph, or a digital image. There is also another class of surrogates which could be described as details of the work (the smile of the Mona Lisa, her eyes, the artist signature, etc.) Is there coding in the records to identify these types of surrogates? Are there ways of organizing them in record displays so they will make sense to the user? What are the types of relationships between surrogates (and between works), and how can they be described and displayed? UCAI may make some preliminary decisions in order to move ahead without waiting for community analysis or agreement.

Difficulty in using legacy metadata. Given the barriers described above, is it possible to bring diverse legacy metadata sets together in a union catalog? Although sometimes starting fresh with new standards sounds attractive, it is neither realistic nor practical; institutions will not discard years of investment in their existing records. UCAI staff has reached the conclusion that legacy data can be usable, and can form a good starting point for a community-managed and supported database. However, significant resources will be required to make legacy metadata operate within a union catalog environment. Compounding the problem is the need for a "critical mass" of image metadata, with an appropriate breadth of coverage, in order to give catalogers an incentive to use a shared utility as a resource. We are convinced that there is a critical need for data sharing in this community. We see this as a "chicken and egg" problem; which comes first, standardized practices or a shared system/service? UCAI staff feels that having a system/service in place will drive standardized practices.

5. OUTSTANDING ISSUES/NEXT STEPS

Phase Two funding has been approved and will address the following next steps:

- Develop a set of production-quality tools that operate on a large standardized set of legacy metadata
- Assess technical needs for a production environment
- Assess database content needs
- Augment database content
- Refine the existing database processing tools and develop new tools, including developing further the "conceptual tools" related to works and images; creating a new mapping tool; generalizing the ingest tool created during Phase One; refining the clustering algorithms; and extending the merge tool beyond basic functionality

Appendix A
Statistical Profile of the UCAI Prototype

	Cleveland Museum of Art	**Harvard University**	**University of California, San Diego**	**TOTAL**	**Percentage**
Record count	118,019	373,895	223,560	715,474	100%
Thumbnail count	10,047	57,345	185,152	252,544	35%
Empty creator	53,036	70,979	82,791	206,806	29%
Empty titles*	652	0	0	652	0.1%
Empty dates	13,705	54,583	35,975	104,263	15%

*There are approx. 5,500 records with “Untitled” as their title and around 20,000 other records with “Untitled” as part of a larger title. There are also approx. 2,000 records with the thumbnail filename as the title.

Appendix B
UCAI Standard Record (VRA Core 3.0 Extended)

The combination of elements and qualifiers comprises the entire element set for the UCAI standard record.

There are three types of elements: R= Record descriptor, W= Work descriptor, I= Image (i.e., surrogate) descriptor. Each element will be of at least one type, but some elements can be of two types.

Element Set: Elements in red are local additions to VRA Core 3.0.

Element	Qualifier	Type
Record.UCAI_ID	None	R
Record.Source	None	R
Record.Creation_Date	None	R
Record.Thumbnail	None	R
Record.Native_ID	None	R
Record_Type	None	R
Type	None	W, I
Title	Title.Variant Title.Translation Title.Series Title.Larger Entity Title.Collection	W, I
Creator	Creator.Personal Name Creator.Personal Name.Work Creator.Full_Personal_Name Creator.Corporate Name Creator.Corporate Name.Work Creator.Corporate_ Name.Corporate_Subordinate_Unit Creator.Vital_Dates	W, I

	Creator.Vital_Dates.Work Creator.Role Creator.Role.Work Creator.Attribution Creator.Nationality Creator.Nationality.Work Creator.Associated_Titles	
Date	Date.Creation Date.Design Date.Beginning Date.Completion Date.Alteration Date.Restoration	W, I
Technique	None	W,I
Material	Material.Medium Material.Support	W,I
Measurements	Measurements.Dimensions Measurements.Format Measurements.Resolution	W,I
Description	Description.Work Description.Collection	W, I
Subject	Subject.Personal_Name Subject.Corporate_Name Subject.Topic Subject.Geographic	W

	Subject.Period Subject.Authority	
Culture	None	W
Style/Period	Style/Period.Style Style/Period.Period Style/Period.Group Style/Period.School Style/Period.Dynasty Style/Period.Movement	W
Relation	Relation.Identity Relation.Type	I
Source	Source.Location	I
Rights	None	W, I
Location	Location.Current Site Location.Former Site Location.Creation Site Location.Discovery Site Location.Current Repository Location.Former Repository	W,I
ID_Number	ID Number.Current Repository ID Number.Former Repository ID Number.Current Accession ID Number.Former Accession	I