Delphi

An online museum collection browser

UC Berkeley, School of Information

May 3, 2007

Authored by:

Olga Amuzinskaya

Adrienne Hilgert

Jon Lesser

Patrick Schmitz

Gerald Yu

Contents

Tables and Figures

1. Executive Summary

The Museum of Anthropology at the University of California, Berkeley, was founded in 1901 by Phoebe Apperson Hearst. The Museum collections comprise an estimated 2-3 million objects, spanning the world and several thousand years. Several years ago, a significant investment was made in a collection management system from Gallery Systems called, "The Museum System" or TMS. Their TMS database contains roughly 615,000 records, which represent nearly all the 2-3 million artifacts in the collection. The large discrepancy in the number of artifacts and records occurs because a single record can represent multiple artifacts. For example, a chess set of 32 pieces and a board constitutes a single record, but 33 artifacts. Of the 615,000 records, about 28,000 have high-resolution images associated with them, with new images added at the rate of a few thousand each month.

While the museum is relatively well known among researchers, a small public exhibition space (relative to the total collection size) coupled with exhibits that are only updated annually, fails to impress upon the general public the richness of the museum's collection. The only available access to the collection for the general public is the annually updated public exhibitions. Researchers can only access the collection with the assistance of a museum staff member who works with the researcher to formulate TMS queries and retrieve artifacts from the museum's storage facilities. This one-on-one approach is a serious strain on the museum's staff and severely limits the number of researchers who can be accommodated at any one time.

The main goals of our project are to expose the collection on the web and increase the visibility of the museum, emphasizing the breadth of the collection. We believe a focus on users interested in researching the museum's collection will yield an interface that is compelling for researchers, museum staff, and the general public.

To this end, our project work consisted of developing Delphi, a system for 1) exploring objects in the Hearst Museum collections by faceted browsing and searching and 2) creating sets of objects. Our system development spanned from conception to evaluation of a functional prototype and included 1) needs assessment, 2) user interface design, 3) ontology design, and 4) software implementation. Throughout the project, we followed user-centered, rapid prototyping, and agile methods of development.

1.1. Prior work & competitive analysis

There were three areas of the project for which we conducted background research on similar and related systems. For the ontology design and vocabulary, we considered a number precedents and theory; the section of this report on ontology development details this. For the technical aspects of the facet browser UI, we considered prior research and the few deployments available; this too is presented in the ontology development section. Finally, for the overall look and feel, as well as the user interaction modeling of the facet browser,

search capability and the support for defining and using sets, we considered a range of museum websites and related commercial sites (Our full notes are in appendix 6: Competitive analysis notes).

1.2. Scope and Deliverables

We identified our main goals, objectives and specific non-objectives and formalized these with the client. We divided the work into 4 sprints, following the “Agile” program management methodology. The work breakdown is summarized in Tables 2 to 5 below.

1.2.1. Goals and objectives

Goals	**Objectives**
Expose the collection to the web.	Design and deploy a prototype system that allows users to browse the collection.
Increase the visibility of the museum.	Attract user traffic by creating fun and easy ways to navigate the site.
Enable the museum to see and report any increase in website popularity.	Monitor and report user activities and areas of increased interest for the website.
Emphasize the breadth of the collections.	Enable efficient ways to traverse different types of collections.
Help the potential archive visitors to prepare for a visit with collection managers by being able to better formulate and communicate their search requests.	Allow potential archive visitors to search and browse multiple collections, drill down their search results, focus and narrow down their search queries. Enable them to reference objects of interest by their catalog numbers.
Emphasize the images that document the collection without violating the museum's ownership rights.	Design a browser that can traverse and efficiently present visual search results. Provide lo-res, watermarked images for viewing by all and, possibly, hi-res images for paying users.
Accumulate valuable information from the researchers and people "in the know" about the objects in various collections.	Provide support for community annotation of the image collection and other collections such as feedback, tagging, referencing, and favorites.

Table 1 — Goals and objective table.

1.2.2. Sprint 1: Project Scope, Needs Assessment and Tech Requirements

Work	Phases of Work	Date Estimate	Confidence Level
Project scoped	Individual sections - 70%; Summary - 10%; Proofreading - 10%	2-11-07	High
User Needs Assessment: Select target user categories. Understand their needs.	Scheduling - 10%; Interviews - 40%; Interview Reports - 30%; Assessment Report - 20%.	2-28-07	High
Provide technical support for developing the system.	Negotiations - 40%, documentation - 20%, set up - 40%	February	High

Table 2 — Sprint 1 planning.

1.2.3. Sprint 2: Design and Development

Work	Phases of Work	Date Estimate	Confidence Level
Design and develop the system.	Analysis, design, dev and debug - 70%; testing - 20%; deployment - 10%.	April, 1	Medium
Develop Interaction Design	Brainstorming and design - 20%; rapid prototyping - 60%; user testing - 20%	April, 1	Medium
Develop UI.	Prototype versions of the compelling functional user interface.	April, 1	Medium

Table 3 — Sprint 2 planning.

1.2.4. Sprint 3: More Design and Development

Work	Phases of Work	Date Estimate	Confidence Level
Test the system prototype.	Report the feedback received from the outside system developers, UI designers, researchers, and naive users. Schedule interviews - 10%; interview - 50%; analyze and report - 40%.	April, 22	Medium
Refine existing features.	Improve system modules - 50%; improve user interface - 50%.	April, 22	Medium
Implement additional features.	Analyze, design, develop, debug - 70%; test - 20%; deploy - 10%.	April, 22	Medium

Table 4 — Sprint 3 planning.

1.2.5. Sprint 4: Wrap-up and writeup

Work	Phases of Work	Date Estimate	Confidence Level
Present our work.	Project report - 80%; slides for the 5 min & 20 min presentation - 10%; ongoing/future work proposals for the museum – 10%.	May, 3	Medium

Table 5 — Sprint 4 planning.

1.2.6. Out of Scope

» We are not making any changes to, nor attempting to replace or supersede the current CMS (TMS). Therefore, we will not address accessioning/deaccessioning or updating object records.

» We will work with the database "as is". Any work to enhance the existing TMS database is out of scope.

» We will not create any original content for our system or the existing museum website.

» Features targeted specifically at users outside the core personae we develop during the needs assessment phase will be out of scope. For example, we do not anticipate educators will be a core personae, therefore features specific to educators are out of scope.

1.3. Initial feedback from client

Initial feedback from our client has been very positive. The staff are very impressed by the visual design as well as the functionality. They love the image zoom and explore widget, and are eager to play more with the sets functionality. They like the facets we defined, including the experimental ones like "Color". We will need much more experience with them before a more complete evaluation is possible, but at this point they seem quite happy with the Delphi browser.

We conducted some initial user testing of Delphi as well. In this small survey, the users enjoyed exploring the Delphi system. Although it will take 6 to 12 months to really evaluate Delphi in light of the larger museum goals, we believe that in addition to making the Museum collection accessible to a much larger audience, Delphi has achieved its goal of providing a fun, user-friendly way to navigate through the collection.

2. Overview of what was built

Delphi is an online browse and search system that can be used by researchers, students, and educators to search and browse through the diverse collections of the museum. Both searching and browsing functionality help a user to learn "what's there." The Delphi system provides features such as searching, browsing, viewing information about objects, viewing featured sets, creating personal sets of objects, and annotating the objects of the collection.

We called the system Delphi after the ancient oracle in Greece. We hope that the system will be a source of knowledge and learning to museum visitors. The name also appealed to the museum staff for the connection to anthropology.

The Delphi facetted browser provides a simple way to navigate through the collection contents. The browser view dynamically changes to show only categories relevant to the area of the collection that is being explored. This makes the browsing experience more visually streamlined and helps the visitor notice categories and relationships between objects that may not have been obvious otherwise.

2.1. Target User

Our main persona for whom we designed our user interface is Theresa Conant – a student/researcher who approaches a museum collection with specific search goals in mind, and who has at least mid-level search skills. Teresa is a not a professional researcher, nor does she have any particular expertise in the museum's collection. However, she has specific, definable needs that can be met with the museum's collection. While designing our first system prototype we were thinking about her needs and her expectations for the system. For the testing, we created scenarios to approximate what her possible tasks might be regarding searching, browsing, tagging, and creating, managing and sharing sets within an online museum browsing experience. We evaluated the performance of the interactive version of Delphi by performing the usability assessment tests with testers who had some characteristics which made them similar to Theresa. This process is detailed in the Needs Assessment section below.

2.2. Ontologies of Objects

The facetted browser is derived automatically from the underlying ontology, and so we spent considerable time considering the usability aspects of the structure and form of the facets, categories and general organization. A facetted design was chosen to make search easier and allow easy comparison of objects across culture, location, technique, this can lead users to discover some non-obvious relationships among the objects in the museum collection. Presentation (User Interface) of the faceted browser was targeted at non-experts, but can be successfully used by a broad range of users, including academic researchers.

Datamining and NLP techniques were used to transform the original museum metadata into an ontology that would reflect a common human experience of the world and provide a pleasing User Interface.

2.3. Implementing the Delphi System

2.3.1. Front Page

The front page welcomes visitors to the Delphi browser, showcases a few individual objects and pre-selected sets created by the Museum staff or other users, and provides access to main system features in a way that makes it easy even for the first time visitor to begin using the system. Figure 1 shows a screenshot of the front page.

Figure 1 — Delphi's front page.

2.3.2. Faceted Browser

The Delphi Faceted Browser (see Figure 2) provides a way to navigate through the collection contents by exploring any facet.

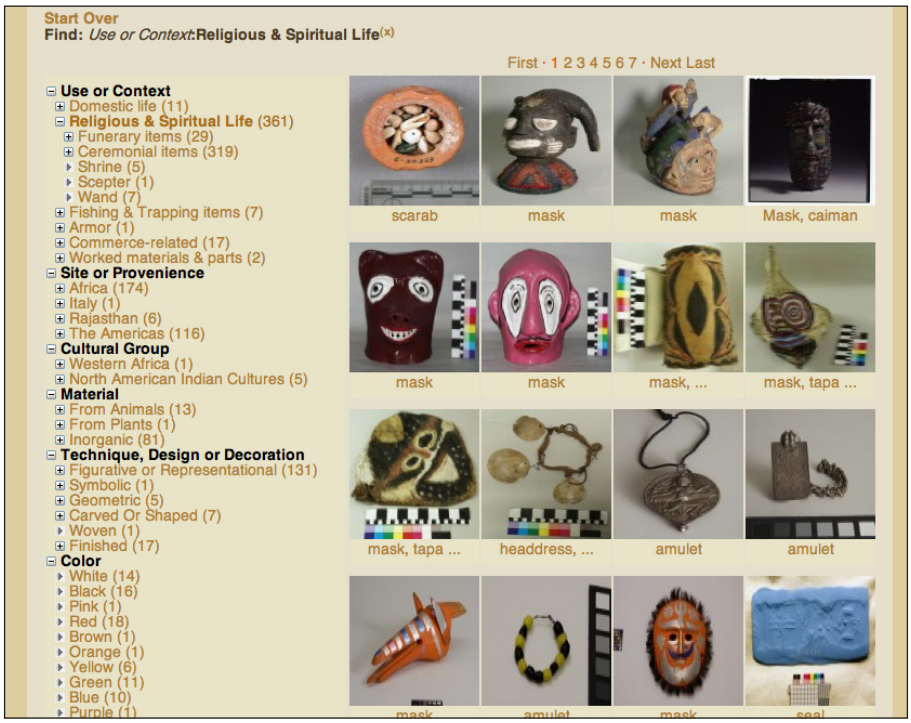


Figure 2 — Delphi's faceted browser.

» The browser view dynamically changes to show only categories relevant to the area of the collection that is being explored.

» The browser is built based on an ontology designed from the metadata in the Museum's current content management system. The Delphi system uses text-mining to associate ontology concepts with collections objects.

» The Help section describes how to browse and search the ontologies of objects presented by Delphi.

2.3.3. Object Detail

The Delphi user interface presents each object with descriptive metadata extracted from the Museum's database, with a zoom-and-pan explorer of high-resolution images, and a list of relevant categories for the object. Figure 3 shows an example of an individual object view.

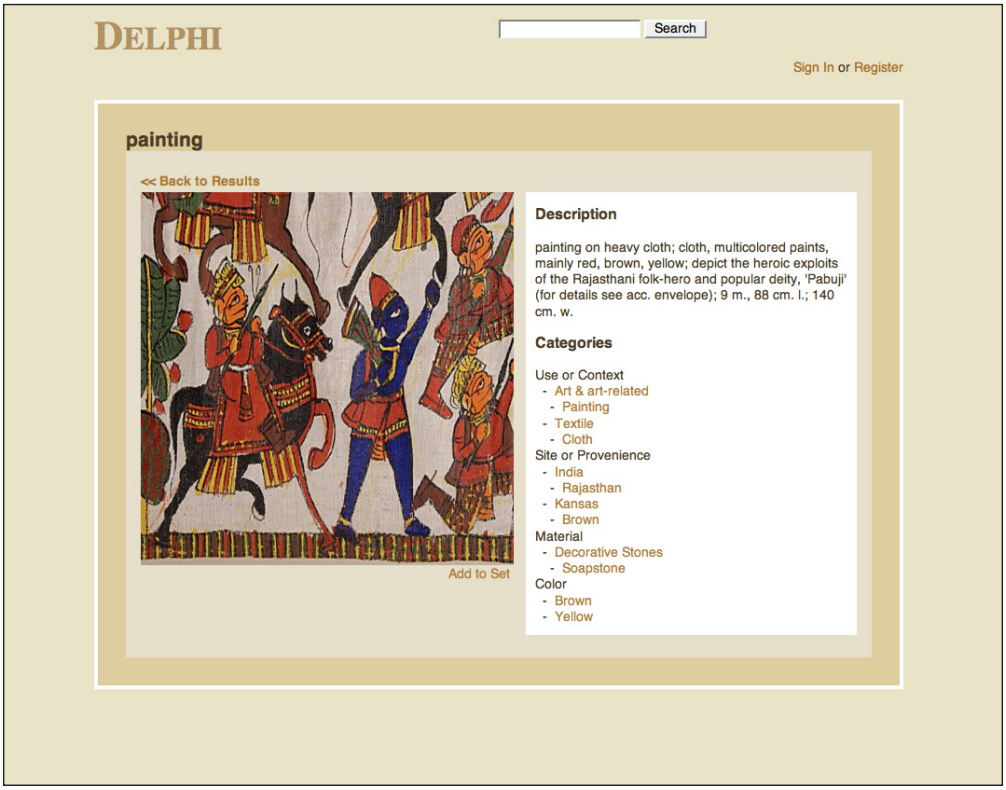


Figure 3 — Delphi's object detail page.

2.3.4. Personal Organization and Annotation Features

The Delphi system supports personal user organization and annotation of the collections by allowing visitors to create sets and tag individual items.

2.3.5. *Sets Feature*

A visitor can create and organize sets of objects of interest, and share their sets with others. This is an example of what a set looks like:

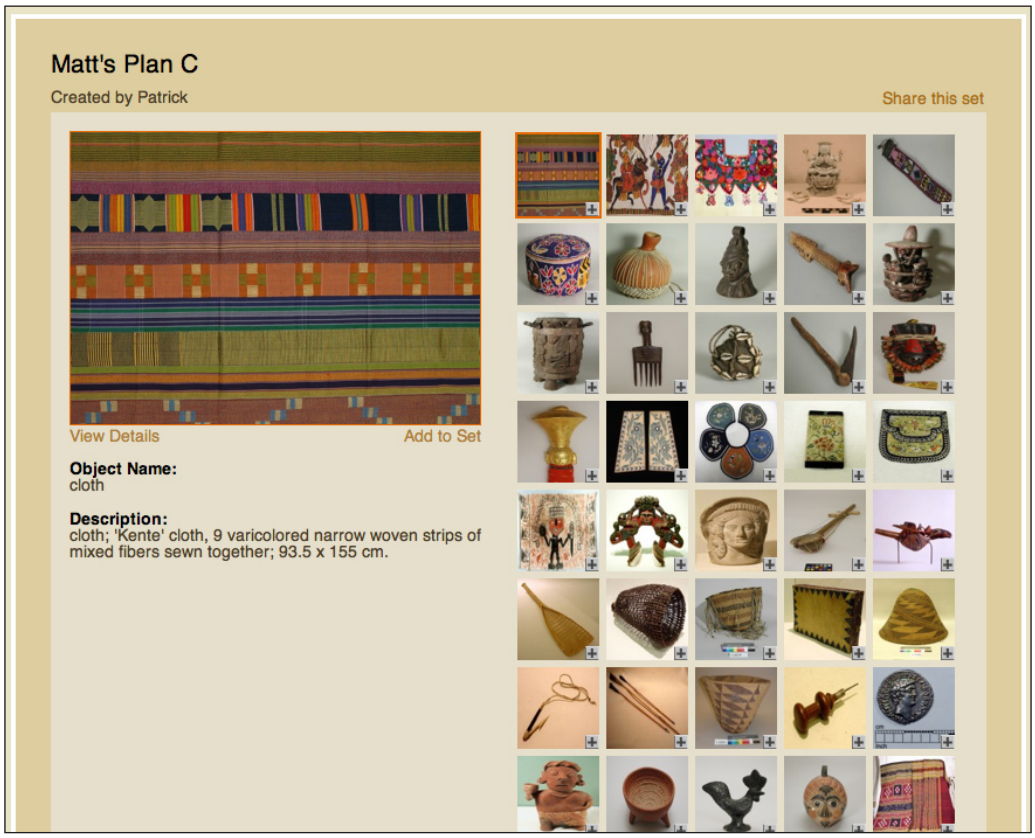


Figure 4 — Delphi's set view page.

2.3.6. *Tagging*

The tagging feature allows users to tag individual objects and search for tagged objects. This is an example of tagging provided on the Item View page

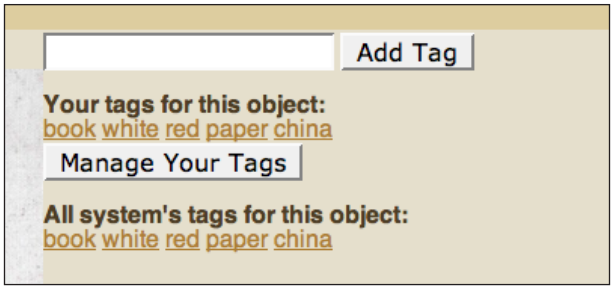


Figure 5 — Delphi's tagging features.

3. Needs Assessment

In our first sprint we conducted a month-long needs assessment to define our target users and what recommendations we should keep in mind as we moved into the design and implementation phases of the project. We concluded that the members of the museum-going public who had specific subject interests would benefit most from Delphi while having needs and feature requirements that overlapped the most with other user groups.

Our process included a data collection stage followed by an analysis stage. We wanted to understand the current practices of possible users of a museum collections browser. These potential users included:

1. museum staff such as collections managers, public programs coordinators, curators, and docents
2. experts in the anthropology/archaeology such as researchers and faculty curators,
3. educators
4. members of the general public who were actual or potential museum visitors.

At the outset, we understood that in the scope of the project, we would need to focus on one of these groups. However, we wanted to start with a broad scope and gather several points of view before narrowing down to the user group we could best serve within the scope of our project.

We collected data through qualitative, semi-structured interviews and observations where two or more group members would meet with a participant either at the participant's workplace (most preferred) or a local library or café if the workplace were not applicable or possible. The interviews and observations used a protocol of open-ended questions but also allowed for the exploration of points not specifically in the protocol; our intention was to understand as richly as possible both the participants' current practices with respect to searching and browsing museum objects and potential uses or a system like Delphi. Before starting a session, we made sure to obtain the informed consent of the participant to be interviewed and observed.

Based on the notes from our interview and observation sessions, we moved into a data analysis phase where we 1) identified several types of potential users and created personae for them to help focus our thinking, 2) did a task analysis of our potential user populations, and 3) created scenarios based on the personae and task analysis. We also conducted a competitive analysis of collections and exhibits browser sites for other museums. At the end of our analysis phase, we selected a target user group and generated a set of high-level design recommendations and a needs assessment report, both for use within our project group and for communication to our project sponsors.

3.1. Interviews

Over a month, we conducted 20 qualitative interviews/observations with museum-related individuals and potential and actual museum visitors. Our goal was to develop an understanding, from several points of view, of the current practices that related to searching, browsing, and viewing items at the Phoebe A. Hearst Museum of Anthropology. This information helped us establish a set of needs for a museum collections browser. Our participants came from several roles both within and outside the museum.

3.1.1. Summary of participants

- 13 museum staff
    - 3 collections managers
    - 2 registrars
    - 1 resident anthropologist
    - 2 public outreach/programs coordination
    - 4 docents/visitor services
    - 1 faculty curator
- 1 former graduate student (now working at the museum)
- 1 educator
- 2 anthropology researchers
- 3 potential/actual museum visitors

We recruited participants through 1) a sponsor-provided list of museum staff interested in talking with us, 2) direct recruitment at a museum event, 3) personal contacts, and 4) snowballing. As such, we were able to interview and observe participants with diverse roles and viewpoints. For participants from the museum staff, the main variation we saw was the extent to which the participant's role was public facing. For non-staff participants, the main difference was the level of expertise with the subject matter related to the museum's collections.

For data collection, we used semi-structured interviews and observations in sessions that lasted from 45 minutes to two hours. If possible and relevant, we conducted the session at the participant's workplace, such as at the museum or off-site facilities. Otherwise, we met the participants in local libraries or cafés. We conducted the sessions with two or more group members, where one member would take the role of lead interviewer while the other member(s) would be note takers. Before starting a session, we obtained the informed consent of the participant to be interviewed and observed and to have their statements be reported in aggregate. When allowed, we used an audio recorder to document the session and provide another stream of raw data. At the end of each session, we asked our participant for likely candidates for further interviews/observations.

During our interview/observation sessions, we asked a mix of open-ended and more structured questions and when possible asked our participant to demonstrate their work practices. The actual questions in the protocol varied by the role of the participant. For example, we asked museum staff questions about searching and browsing for items in the museum collection with their current tools, with an emphasis on their digital database and what fields were the most useful to them in describing an object. We wanted to know how much time and how many different search refinements it took them to "find" a specific object, if their searches yield successful results. We also focused on the types of people they encounter who are interested in the museum's collection and how frequently they visit. For non-staff participants, we asked a broad range of questions about their interaction with the collections and exhibits at the Hearst and other museums, both physically and online if applicable. We wanted to understand their familiarity with collections browsers, their current museum visiting practices, and what would attract them to museum related sites.

Generally, our participants also explored topics – not explicitly in our protocol – that were relevant to search, browsing, viewing, or working with the museum's collections. Between sessions, we revised our questions to better elicit this information.

3.2. Personae

In order to understand the users for which we were designing, we conducted two on-site contextual inquiries with Museum personnel, who were assisting visiting researchers at the time. Each lasted about 2 hours. In both cases, the Museum personnel were Collections Managers, who are responsible for providing access to the physical objects in their collection. One of their many duties is to perform searches of all relevant articles that the Museum may have, which the researcher is interested in studying. Searches are performed from several sources (card catalogs, accession files, ledgers, and the digital database) that each store a different set of information – when pieced together, these records provide a complete data set about the artifact, ideally including a digital image.

The contextual inquiries gave us a grasp of the sorts of tasks likely to be performed, as well as the different categories of users who might likely interact with our design. We identified many user "types," but focused on three as being more representative of the characteristics described to us:

» **general public:** someone who might discover objects of interest within the collection if it was easy to browse. *Low level search skills.*

» **student /researcher:** someone who approaches the collection with specific search goals in mind. *Mid level search skills.*

» **museum staff / curator:** someone who regularly performs searches, additions, and corrections to records in the collection, and knows the important dimensions that relate to each other. *High level search skills.*

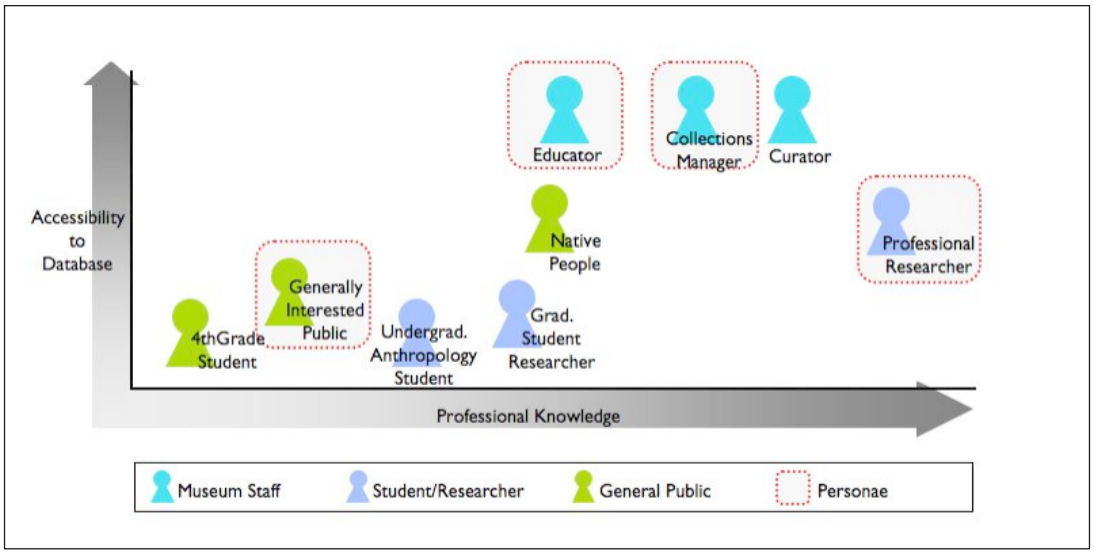


Figure 6 — Personae plotted by knowledge and access.

We conducted three interviews with museum personnel who could articulate the interests of each category, because of their specific role within the Museum. Our participants included an educational outreach person, a recent grad student from JFK University, who earned an M.A. in Museum Studies, and the curator of the current gallery exhibition.

Each interview lasted approx. 2 hours, and we asked a mixture of open-ended and structured questions that differed slightly depending upon their role. All of the questions revolved around searching and browsing for items in collection, with an emphasis on their digital database: what fields were the most useful to them in describing an object, how much time and how many differnet search refinements did it take them to "find" a specific object, did their searches yield successful results, and so on. We also focused on the types of people they encounter who are interested in the Museum's collection, and how frequently they visit.

From these interviews we were able to glean that searching through the database currently is a cumbersome process, which requires a highly trained individual to achieve results. In addition to the categories we were previously able to identify as being "important" users to consider, we also learned that a variety of other people use the collection's artifacts, or interact with the collection in critical ways. This includes scholars affiliated with a research institution, non-affiliated researchers, native peoples, the Museum Registrar, and all manner of students with varying degrees of sophistication -- from elementary school kids, to highschoolers, to undergrads. From this wide variety of users, we combined several into four distinct Personae, who were grouped together by their goals.

3.2.1. Description of personae and goals

It's important for you to remember as you read the following descriptions that personae are not real people – they are a 'composite' of characteristics from our target audience, that we discovered through the interview process. Personae are important to construct, because they represent an archetypical person with whom we can identify, and who we take into consideration at every decision point along the way. What would this person want? The persona has specific goals that we're trying to help him/her achieve with our system. Narrowing the focus of our design with one person in mind, who comprises significant traits of many, helps us increase the overall functionality and satisfaction for this set of users.

In a discussion that followed our interviews we came up with 12 different personae and wrote the name of each persona on a post-it note. We then arranged and discussed the personae until we had what turned out to be a close approximation of an affinity diagram. The personae were grouped by similarity and arranged along a spectrum of "hardcore-researcher" to "casually-interested public." Once arranged, we picked four personae that we felt were most important and sufficient to represent the remaining personae.

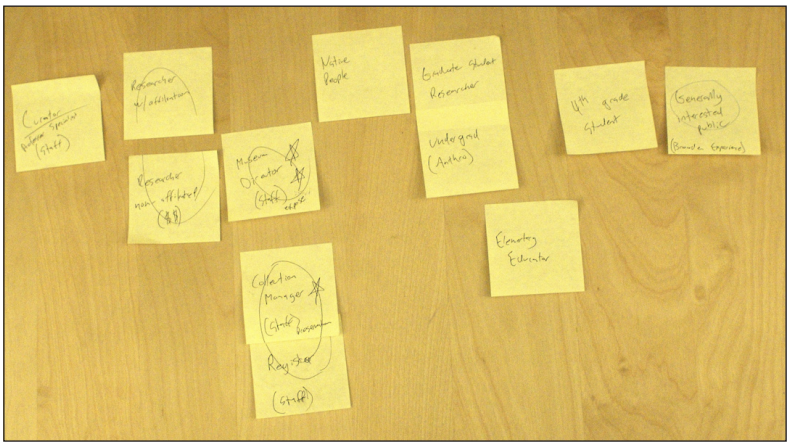


Figure 7 — Planning with post-it notes.

3.3. Persona 1: Alan Prewitt, Ph.D.

Age: 56

Category: Researcher

Occupation: Professor

Figure 8 — Alan Prewitt.

Alan Prewitt is a full professor in the Department of Anthropology at the University of California, Berkeley. Alan received a BA in Anthropology from Colorado College in 1975 and Ph.D. from Columbia University in 1982. He joined the faculty at Berkeley in 1986. Professor Prewitt is a cornerstone of his department and is well liked by graduate and undergraduate students. His course "Great Sites and Lost Tribes: The Romantic Element in Archaeology" is a favorite among Anthropology undergraduates.

His primary research revolves around the Lakota, Dakota, and Nakota cultures of the American west. While in the past Alan published four-to-six papers a year, he's recently put most of his energy into a forthcoming book. Alan has worked on a dozen or so projects that heavily involved the collections of the Phoebe Hearst Museum on campus over the course of his tenure at Berkeley.

Alan spent two weeks over the summer with his wife Melinda hiking and camping in South Dakota. Alan's wife Melinda Prewitt is a sociology professor at UC Berkeley. Alan and Melinda have a 22 year-old son, Michael, and two dogs, Lewis and Clark.

3.3.1. Quote

“The Kabwe skull holds special significance for me. In 1975, I wrote my very first undergraduate paleontology research paper on this skull.”

3.3.2. Goals

» Finish Political Organization of Native North Americans article for submission to “American Anthropologist”.

» Use images to illustrate his forthcoming book, The Last Days of the Sioux Nation

» Find some interesting visuals for his undergraduate course, “Great Sites and Lost Tribes: The Romantic Element in Archaeology”

3.3.3. Justification

Interviews with the Museum staff revealed one of the primary users of the collection were researchers affiliated with colleges and universities. Alan represents one such researcher with an interest in Native American collections, which make up a substantial portion of the total Museum collections. One researcher we were able to talk to was working on a book about Native American baskets. She was photographing many of the Museum’s baskets for publication in an upcoming book.

3.4. Persona 2: Joyce Reisner

Age: 33

Category: Museum Staff

Occupation: Collections Manager

Figure 9 — Joyce Reisner

Joyce Reisner is a Collections Manager who specializes in the Asian region, which includes artifacts from India, Japan, Mongolia, China, and some parts of Indonesia and the Philippines. Joyce earned her Bachelor's degree in Anthropology at UC Berkeley. A favorite professor steered Joyce into pursuing a Master's degree in Asian Studies, which is still her passion.

She has been working at the Museum for about 4 years, and experienced the "conversion" firsthand, when the Museum records "went digital." Well, at least that's the ultimate goal. Only a few collections (North America and Egypt) are fully represented in the database with complete descriptions and images. Artifacts in other collections are represented in the database with just an object number and a storage location number.

The Museum only has one available workstation from which visiting scholars can run database queries, and it's constantly in use. Researchers come with long lists of items to find (sometimes to the Museum staff's chagrin). On the plus side, it's not unusual for researchers to have additional information to contribute about an object, which can add content and richness to their future body of knowledge. Unfortunately, there's no consistent place

to capture that data. It's a good thing Joyce has such a good memory for details! Although it's not part of her job description, she sometimes is the conduit of putting scholars in touch with each other, when she notices their complementary research pursuits. It's ironic that with such finely-honed research skills, they're still dependent upon Joyce's available time and esoteric knowledge to help them find relevant material in their area of interest.

With her one-to-many relationships and the responsibilities of handling a large collection, Joyce is kept busy from early morning to after hours.

3.4.1. Goals

» Preserve the artifacts in her collection, which means limited public handling

» More resources available for visiting scholars to perfom their own searches

» Have a consistent place to add new details about an object

» Include images in the digital database, so that researchers can browse images first, then narrow their artifact requisitions before approaching Collection Managers

» Have a life outside of the Museum (which means working normal hours).

3.4.2. Justification

This persona closely models what we observed when we interviewed several Collections Managers at the Hearst Museum.

3.5. Persona 3: Sally Grant

Age: 35

Category: Museum Staff

Occupation: Museum Education and Public Relations

figure 10 — Sally Grant

Sally is a Museum Educator at the Phoebe A. Hearst Museum of Anthropology. Sally double majored in Anthropology and Education as an undergraduate before pursuing an MA in Museum Studies at J.F.K. University. Her thesis was titled, “University Museums and the Development of a Campus Audience.” She strongly believes that visiting museums helps kids to develop and grow.

Sally loves children, and is always supportive when elementary and middle school teachers contact her to setup a field trip to the Hearst Museum. She tailors the information she gives them to guide them toward exhibits that would be most relevant to what the students are learning. However, since she doesn’t have access to the actual collections, it’s often a difficult and time-consuming proposition for her to gather the appropriate objects for them to see.

She wants more people to come to the Hearst Museum, and is always working on publicity materials to capture the public’s attention. Creating a strong online presence is one of her

main projects this year.

In her free time, she visits the websites of other museums and galleries to compare their education programs and public relations efforts. She especially likes the SFMOMA website because they have a well-designed virtual gallery which enables visitors to see the depth of the collection.

3.5.1. Quote

“the future of museums is interactivity”

3.5.2. Goals

- Expose the Museum’s awe-inspiring collections to the general public, educators and students
- Make the Hearst Museum website an exciting, enjoyable and interactive experience
- Work collaboratively with teachers
- Efficient way to save object details, associated study materials and notes from past field trips, so that they can be easily accessed in the future.

3.5.3. Justification

One of the goals of the Museum is to increase their visibility. Our system could potentially be a valuable asset that enhances the public face of the museum. This was confirmed when we interviewed the person in charge of public outreach at the Hearst Museum. She also saw a role for our system in her own work when she prepares materials for visitors, and answers their questions about the Museum’s exhibits and collections.

3.6. Persona 4: Theresa Conant

Age: 24

Category: General Public

Occupation: Aspiring Young Designer and the Academy of Arts Fashion Student

Figure 11 — Theresa Conant

Theresa Conant's been interested in fashion since the age of 10. Her first job was working in a designer clothing store in San Francisco, mainly because she got great outfits at wholesale price. But because of her responsible work ethic, strong accounting skills and excellent communication with customers, Theresa was soon promoted to store manager. As such, she's acquired an uncanny ability to spot the next trends, which has made her store very successful. In her spare time for fun, she made a few designs of her own, and much to her surprise, her customers clamored to purchase them!

She still works part-time at the store, but is focusing all her energy at the Academy of Arts, where she pursues a degree in fashion design. Once she graduates, she's determined to launch her own line of clothing. She already has a large and eager clientele, based on her extensive customer contacts and networking ability.

Theresa looks for inspiration in fabrics, on the Internet, in nature, on the news -- and carries a small sketchpad wherever she goes to jot down ideas. Lately she's been obsessed with the idea of 'cultural heritage,' and found one intriguing image on the Hearst Museum

website when she was searching for unique textile designs.

Theresa would love to spend more time visiting museums and galleries, but her time is extremely limited between her job, her studies and her creations.

3.6.1. Goals

» To find time-saving tools that make her busy schedule more manageable

» To create beautiful things that stand out from the ordinary

» To represent herself in her designs

» To customize her life

» To get inspiration and ideas from other cultures around the globe

» To share her special items with others

3.6.2. Justification

One of the Museum's goals is to increase access to its collections. Our interviews with Museum staff confirmed that they are very interested in increasing access to the general public as well as researchers. Defining the general public is a nebulous task. We believe the members of the general public that use our online system will approach it with a vast diversity of very specific motivations. This persona attempts to capture one such motivation.

After many long discussions we chose to closely align our design with the Theresa Conant persona. Theresa is a not a professional researcher, nor does she have any particular expertise in the Museum's collection. However, she has specific, definable needs that can be met with the Museum's collection.

3.7. Task Analysis

In our initial design process, it was essential that we studied our users overall goals, and how they related to the tasks they performed each day. Conceptualizing a model of what users want to accomplish and their associated information needs, and then ranking them in levels of priority (high, medium, low) for each user, helped us to see which tasks were most important, and therefore which needs should take precedence in our interface design. Breaking down individual goals and needs into a task analysis also makes it possible to identify areas of high priority alignment across users.

As we see below in the chart, our task analysis showed organizing sets to be the most important task, following searching and browsing. We therefore concentrated many of the subsequent features and functionality we designed around giving users this ability.

	Professor: Alan Prewitt	Collections Manager: Joyce Reisner	Museum Educator: Sally Grant	Designer: Theresa Conant
Goal 1: Collecting research material	High	Medium	Medium	High
Searching for specific objects in the collection	High	High	Medium	Low
Browsing objects in the collection	Medium	Low	High	High
Visiting the museum: exhibits	Low	Low	Low	High
Examining physical objects in the collection	High	High	Low	Medium
Organizing materials into a set or sub collection	High	High	High	High
Goal 2: Broadening personal experience	Low	Low	Medium	High
Visiting the museum website	Low	Low	Medium	High
Visiting the museum: exhibits	Low	Low	Medium	High
Participating in educational programs	Low	Low	High	High
Goal 3: Bettering the museum experience for others	Low	High	High	N/A
Providing information about objects in the collection to requesters	N/A	High	High	N/A
Contributing information about objects in the collection	High	High	N/A	N/A
Helping outside researchers	N/A	High	Low	N/A
Helping teachers plan visits	N/A	Low	High	N/A
Goal 4: Increasing visibility of the museum	Medium	Medium	High	N/A
Highlighting specific artifacts and collections	High	High	High	N/A
Exposing the collection to the web	Low	Medium	High	N/A
Advertising and public relations	N/A	Low	High	N/A
Designing educational programs	High	N/A	High	N/A

Table 6 — Task analysis.

3.8. Scenarios

Scenarios are developed in order to test the validity of proposed design concepts, and also to test whether we have provided the necessary components to complete a task. Each scenario is a small vignette that presents our persona with a routine task that they must

accomplish in the context of using our Delphi system. Creating realistic scenarios is an exercise that forces us to go through the daily activities performed from each of our personae perspectives. This in turn helps us produce efficient interaction design and useful functions, while eliminating unnecessary features.

We developed three scenarios based on our Theresa Conant persona and her goals.

3.8.1. Scenario 1: Theresa wants to check out the previous Native American Baskets exhibition.

The semester is midway through, and she is busy with several course projects. "It's awful", Theresa says, "that the professors always tell us to visit as many museums as possible to get inspired, but rather than give us enough time to go there, it's endless assignments!" She marks her calendar with all the events that she wants to go, but due to the busy semester, she rarely ends up going. This time, she missed the huge Native American Baskets exhibition that just ended yesterday.

However, this time she remembers the Delphi system provided by the Hearst Museum of Anthropology in which all the current and previous museum exhibitions are shown. So she visits the website. On the front page, she sees the big emblematic picture of the exhibition and a link to the set of the objects. Every object has several high quality pictures that were taken from different perspectives, and along with a detailed description. She zooms in to get a closer look at the details. She takes particular note of the materials, stitches, and weaving that she may be able to incorporate in her own designs.

She is excited that even at the 2 A.M. in her room, she can view all the details of the objects that were shown at the exhibition.

3.8.2. Scenario 2: Theresa makes her own collection for the later use.

Theresa is taking a "Costume Design" class for this semester. Preliminary fashion sketches and samples of proposed materials must be turned in tomorrow for a project midterm grade. As a first-year student at the Academy of Art, Theresa has a lot to prove. The competition is intense, and if she wants to make a name for herself, she has to come up with something really memorable, something that has meanings on several levels. Theresa has been looking for a concept that could even become her "signature."

She leafs through her sketchpad again for some inspiration. The Native American patterns she sketched from a few weeks back are simple, yet very strong. This could work, but she needs to see a few more designs and expand upon them in her own style.

Unfortunately, she's supposed to fill inventory tonight at her store, and it's already two in the afternoon. Theresa remembers that the Hearst Museum over in Berkeley has one of the largest Native American collections in the country. Well, between work and rush hour traffic, she doesn't have time to visit the exhibit now! She goes online to the Museum's website

and clicks on Delphi. She quickly browses through their North American collection, first looking at Native American textiles, then moving onto Native American baskets. Every time she sees something she likes, she adds the object to her collection. It hasn't taken her long at all, and she's already got more than 50 images in her set.

3.8.3. Scenario 3: Theresa shares her collection with her project members.

One of Theresa's school projects requires her to work in a group. As a part of a contextual study, three group members decided to gather and share images of patterns that will be applied to their work. Theresa remembers the set of Native American textiles and baskets that she created the other night.

She visits the Hearst Museum website, clicks to Delphi and logs in to her account. There are 3 sets that she created herself. She emails the "Native American" set to herself, and also to the two other group members with a short message saying, "I think this could be good reference material for our project". Theresa's group-mates receive the email from Delphi and peruse the set of baskets.

3.9. Target user

Amateur Researchers / "Seekers:" We have decided to design our system to accommodate the needs of the general public, which the Museum wants to attract. The system can also be successfully used by researchers, students, and educators.

Key Target User Qualities:

- Task oriented / time Sensitive: Have a goal in mind. If they do not succeed in finding their item, they will abandon search and look elsewhere.
- Off-and-on museum goers: Usually come to the museum to see an exhibit that has personal appeal. Want to learn "the story" for an item that caught their eye.
- Internet savvy: Active Internet searchers.
- Seekers, not contributors: Do not see it as their mission to leave information behind for others. May want to make comments for themselves – free associations, reminders, "field notes," etc.
- Interest in seeing online pre-made "Collections: Want to see the objects that have already been organized along some dimension. Interested in seeing exhibits created by authoritative museum stuff/researchers. Interested in learning a story that these stuff/researchers can tell about an item or a collection. Many said that what has meaning to them "is when they can compare and contrast the same object across cultures."
- Images are a must: Pictures are THE most important item to have. Want to see images that are relevant, and not have to wade through a lot of content that doesn't have an associated image.

» Low tolerance for "pain:" Have little interest in seeing data from database which is "dumped" into the record. For most of our participants, this is a negative, exhausting experience, which would make them go elsewhere

» Public vs. private: They would like to have a privacy option for themselves, their sets, and their tags.

3.9.1. Justification of the Target User

We focused on the amateur researchers, because their core requirements satisfy a "sweet spot" of needs across the various personae we have studied, and so provide value to a broad range of users. The amateur researchers are the general public representatives who are looking for specific information in our collection. They are the people who may have some knowledge of the subject (but are not specialists) and have a specific goal (or sometimes a long-lasting interest) that the browser will help them to achieve (see also section X.Y "Description of Our Target User"). We believe that we might realistically count on a reasonable number of such users for our system.

We did not choose curators, educators, or collection managers as our target user for the following reasons:

» Educators are out of the scope because designing for them is a separate project by itself. Their requirements for the system are too tightly connected to the specifics of their occupation and designing specifically for their needs may influence the design of the system a way that will make it less appealing for the general public.

» Researchers value the reputation of their data source extremely highly and are very discriminating about other people's opinions. They, however, would not mind checking out "what is available" occasionally, and will be interested in finding images and copying them or creating their own sets out of them for their work.

» Collection managers, like educators, require specifically designed system that is customized to their work functions. For the public outreach goal of these users, Delphi would provide them with an opportunity to do so as well as to find, create and store some additional, rich (even if somewhat unreliable) metadata about collections/objects, even if it is in "one more system."

» Faculty curators: have interest in having their work recorded in some way. In the past at other museums, this has been in the form of a comprehensive "coffee table" book that encapsulates most of the work on an exhibit. However, their involvement in making this giant document is more at a management level (deciding what content goes in) rather than actually producing the actual book.

3.10. High level recommendations

Target user population is the interested/directed general public (eg amateur researchers).

Site content:

- Images are a must
- Enable seeing an image from different angles, being able to rotate, and seeing enlargements of the portions of images to show texture are great.
- A balance of info/image must be achieved to be positive and satisfying experience, return visits, and avoidance of the unpleasant user experiences.
- Watermarks will be tolerated as long as they do not destroy the visual appeal.
- Showcase the collections created by others. For example: as a good entrée into browsing or viewing the most popular objects or blogs about objects.
- Content (and search results) presentation: If possible, include context beyond just the object. Content that puts object in context is how general public "experiences" an object in an anthropology museum (very different than an Art museum).

User Activities

- Search
  - Keyword search a must!
  - Minimize search time
  - Provide "guidance" for searching.
- Creating sets and sharing them.
  - Many would want their sets to be private
  - Some would want their sets to be shared among a group of specified users if they are working on a group project
  - Some would want their sets to be viewed by the general public
- Annotations and tags
  - Include features for authoring annotated collections.
  - 1st: Ability to comment on individual objects.
  - 2nd, ability to comment on groupings
  - Ability to send URL, with comments
- Privacy: Default user activity to private
- Profile: Not interested: the museum browser does not lend to interest in chatting with others general public representatives about museum objects
- Registration: Is acceptable for the users who are really interested in additional features provided by the system.

4. Design Process

Our project's design sprint followed the user-centered design process as closely as possible, where we took the findings of our needs assessment to ideate, rapidly prototype, and evaluate Delphi's features over several iterations. The efforts in this area were roughly broken down between work on the sets functionality and work on the rest of the Delphi site, which included the facet browser.

Our design phase began while data was still being collected in our needs assessment, as we started to see common threads from different participants in our preliminary analyses. In particular, we placed more emphasis on the sets functionality and made it the primary design effort of group members in Professor Marti Hearst's I213: User Interface Design and Development class. We started with design sessions where we brainstormed features and their initial forms and produced rough sketches and descriptions of our ideas. From these, we started to actually design Delphi, creating interaction flows, mockups, sitemaps, and other artifacts that helped us better define how Delphi would function from the user's point of view.

We then created paper prototypes or screen walkthroughs of features and site pages. We tested these with users to determine the validity of our design direction and also to gain feedback, which we used to refine our ideas and designs. Our design process also included other evaluation methods such as heuristic evaluation and peer design reviews.

As our designs moved toward high fidelity interactive prototypes and actual implementation, we continued with the cyclical process of a) designing, b) building, and c) evaluating. Our user testing included a pilot usability study of Delphi's sets and browsing/searching functionality. At each major stage of design, we made sure to get user feedback to validate or challenge our interfaces. We would then take the lessons learned into the next iteration.

4.1. Facetted browser design process

The design of the browser pages went through numerous iterations from initial sketches to the current version. Throughout the process of designing these pages, we kept in mind a number of principles and constraints such as the need for the page to display results with and without images and the heuristic of having 4-10 facets visible as well as 3-15 subcategories within each facet. The principles driving the design of the facetted browser are further discussed in the section on Ontology implementation. We also incorporated feedback from users or peers at each iteration.

Our initial sketches focused on defining the page elements and exploring possible layouts. We decided that each page should have a facetted browser, search results with thumbnails, and some navigational elements like breadcrumbs. We also decided to use a two-column layout with the facetted browser on the left and the browse results on the right. While we

considered a one-column layout with the facetted browser above the results, we determined that this layout would have issues with horizontal scrolling and the visibility of results if the facetted browser took up too much screen space.

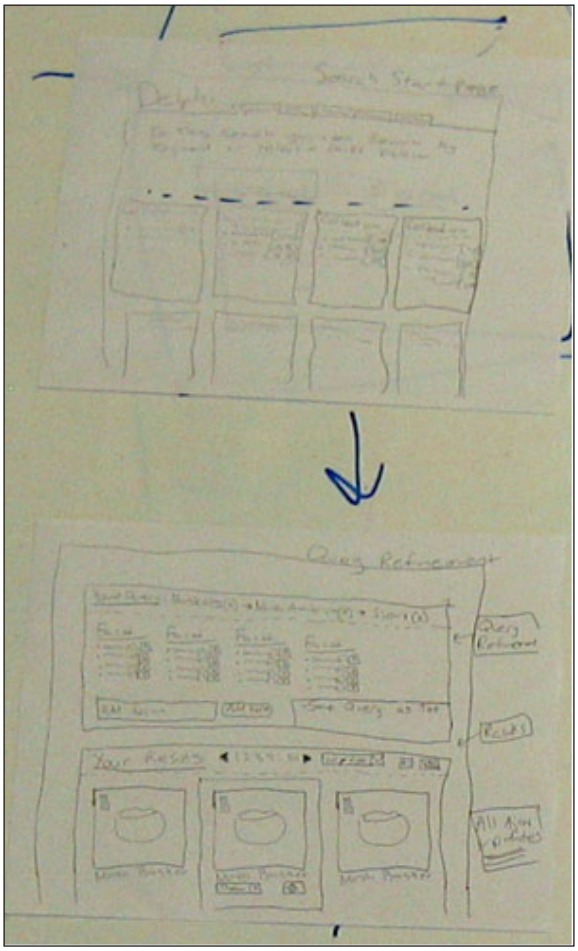


Figure 12 — Sketches: one-column layout

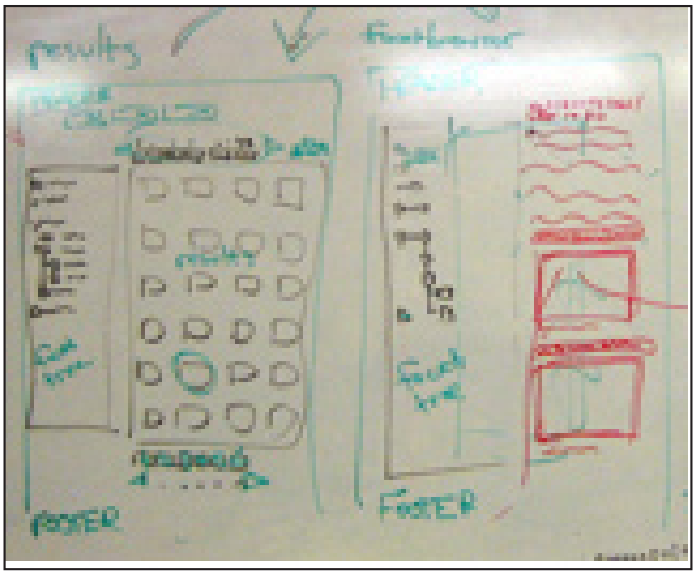


Figure 13 — Sketches: two-column layout

We then created several versions of onscreen HTML mockups. At this stage, we began to think about the different states of the page such as when the user had not selected any categories in the facet browser. We created a page to represent this state that included instructional text as well as a sample search result. Our rationale for including these elements was the fact that a user may need to given help on how to use a facetted browser as opposed to one with mutually exclusive categories.

We evaluated these mockups with two users to assess general usability issues with the design. At this stage, we left out a mockup of the actual facetted browser but did ask the users to comment on what kinds of features they would want in one. While the overall look and layout were generally liked, users also pointed out several areas for improvement, mainly in the areas of feedback, navigation, and the usefulness of page elements. In particular, users felt that they would want feedback in the facetted browser about the number of results in each category, beyond just the number of results in the current query. They also said that text labels in the results grid would be useful for giving a better idea of what the returned objects were, especially if they were not familiar with objects from anthropology and archaeology. Also, they wanted a mechanism like breadcrumbs to remove query terms. They also felt that the instructional text and sample result on the initial browser page could be confusing to users and were of limited help in explaining how a facetted browser worked. They preferred to learn the functionality by using the browser and preferred help text to be off screen on a separate dedicated help page.

Figure 14 — HTML mockup: initial browser page

Figure 15 — HTML mockup: results page

Using the feedback from the mockups, we then iterated to an interactive prototype. We also made changes to the interface as features such as the facetted browser became available for integration.

On the initial browser page, we reduced the instructional text and eliminated the sample result. Though users expressed a preference for a dedicated help page, we still felt that the facetted style of browsing was still unfamiliar enough to most users that some text would still be helpful in getting novice users started. Also, as a result of performance considerations and our focus on showing visually appealing results, we made the default behavior of the browser to return only results with images, while giving an option to return all results.

Figure 16 — Interactive prototype: initial browser page

On the results pages, we changed the layout of the results grid to incorporate four-column instead of three to increase the visual density of the results and thus reduce the amount of scrolling needed. Underneath each thumbnail, we also included a text label with the name of the result. We also implemented breadcrumbs, which displayed the history of selected query terms incorporated a mechanism for removing the category from the query. We also provided a link for removing all query terms in one click to start over the browse session.

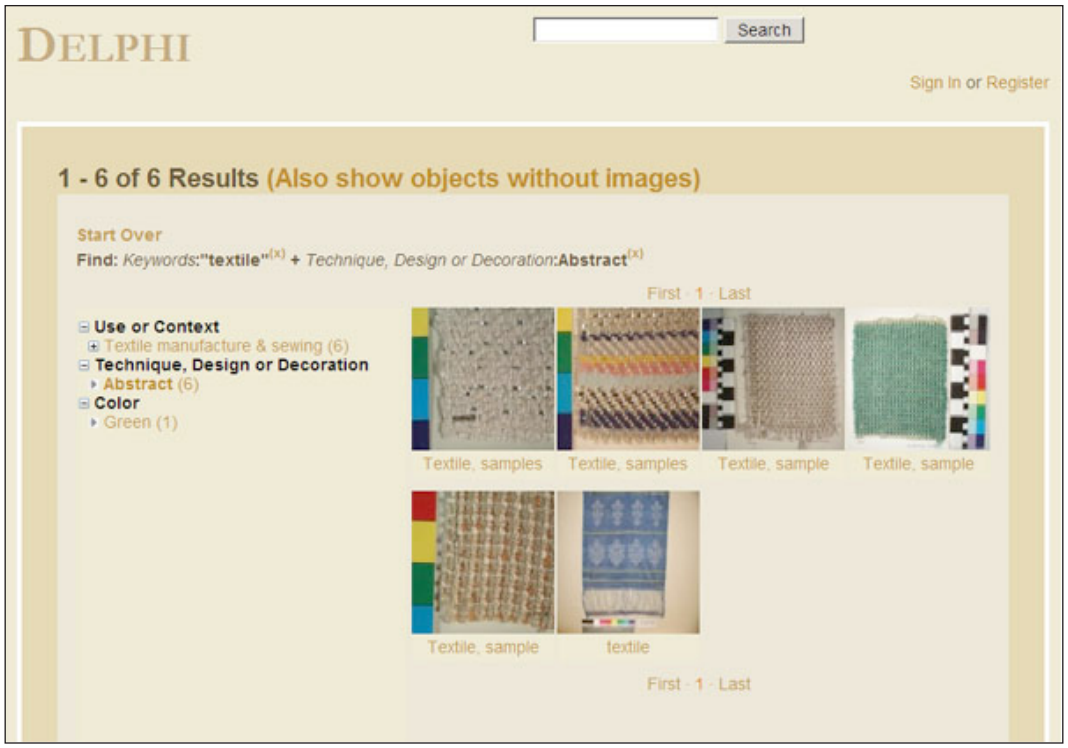


Figure 17 — Interactive prototype: results page

5. Ontology Development

A significant effort within the Delphi application to the PAHMA museum was the development of the facetted ontology. The ontology was used both to categorize the museum collections and as well to generate the user interface presenting a facetted browser for exploring and searching the museum collections online. Initially, we had assumed that the existing metadata in the collection management system (CMS) made use of controlled vocabularies (CV's). We anticipated a modest ontology development scope that would require a certain amount of translation or cross-walk/mapping, but would focus primarily on generating a pleasing and easily understood interface. We see our primary audience more as the enthusiastic public than academics or other expert researchers, and so ease of use is paramount. However, in discussion with museum staff and examination of the existing CMS metadata, it emerged that there was minimal use of CV's and that the object annotations were largely free text, conforming only somewhat to standard language. Moreover, the use of fields (tables and columns) in the database was inconsistent.

We resolved to expand the scope of the project to include a text mining component that would follow current semantic search models to categorize the objects in the collection against the ontology, producing the metadata necessary for facetted browse and search. The rest of this section describes the design of the ontologies, the text mining tools, evaluation and future work.

It is worth noting some related work in this area. Several research groups have addressed the issue of photo search (in the context of information management for non-professionals). Dumais et al. [2] provide a text-search interface that then supports many pivots and sorts to explore a document collection. While they do not leverage a facetted ontology, their pivot model effectively presents facets that reflect common user experience. The Haystack project [6] explored the use of dynamic hierarchies to refine searches of media based upon both personally authored and automatically extracted annotations. However, they do not describe an application with anywhere near the scope we are dealing with. Flamenco [13] demonstrates a faceted browsing solution for image collections, and includes tools to help extract the facet metadata that underlies the UI. We built upon many of the ideas in Flamenco, although we took a somewhat different approach to defining the ontology. In particular, many of the concepts were either fairly specific to the domain (e.g., specific types of baskets, types of ancient coins, and names of Native American tribes), or were so general as to defy simple word labels (e.g., figurative designs featuring a flower). Because of this, we were not at all confident that WordNet or other thesaurus resources would be very useful. We built upon some earlier work presented in [9] that explored facetted browsing of photo collections and in [10] that investigated quality measures for ontologies used in facetted browsing applications. In addition to these academic resources, we also considered many enterprise websites that present collection browsers or commercial applications of facetted browsing. These informed our design principles for supporting the browser UI.

5.1. Choice of facets

We were informed by the general principles of facet design described in [8, 11] and elsewhere, such as selection of facets that are as orthogonal as possible. However, this was not always straightforward. For example, it proved very hard to come up with an organization or grouping of *cultures* without resorting at least in part to *geography*.

The ontology is the basis of the user interface for browsing and query refinement. As such, a design goal for the ontology was that it produce an easy to use and understand UI. This led us to a general goal to have on the order of four to ten facets. Fewer would likely not be enough for such a large collection, and more becomes unwieldy to use. In addition, for a pleasing UI, we felt that we should have a relatively narrow graph shape. Our target was to have between three and fifteen siblings at any grouping level. We knew that this was at best an indirect route to a nice UI, since it is nearly impossible to predict how many sibling categories will actually be matched for any given set of query results. Nevertheless, we hope that this improves the usability of the ontology (exhaustive usability testing has not been done, and we do not yet have sufficient experience with users to draw definite conclusions on this).

Ranganathan [8] proposed a model for facets that while oriented to book-like objects, still captures the ideas of substance and activity around each object. Soergel also discusses facet design and describes commonly encountered facets of *Things or Objects, Materials, Properties, Processes,* and *Goals or Purposes* ([11], p94). We drew upon these as well as recent work in cultural heritage access [1] that emphasizes facets reflecting a human experience of the world over formal academic models. We also had to acknowledge the metadata available in the collections management system, and the implicit facets in these metadata. At the same time, taking a text-mining approach opened up some new possibilities for facets such as *Color*.

After much discussion, we opted to present facets that describe Who, What, Where, When, How and Why:

1. Who: **Culture** presents the culture associated with an object. These are grouped as much as possible under super-culture groups, but in many cases the only reasonable grouping was by general geographic region.
2. What: **Material** describes what the object is made of (but not what it is in a functional sense - see also #6 below).
3. Where: **Location** follows common precedent, borrowing from gazetteers. However, we introduced groups for states (e.g., “Southwest”), etc., to follow our self-defined guidelines for producing a pleasing UI.
4. When: **Time Period** describes named periods in history for various objects in the collection. We need to convert period names to date ranges (either using or producing a temporal gazetteer as in [7]) to get a uniform dimension across which all objects can

be compared. This work remains to be done, and so no *when* facet appears in the first version. This is a high priority for the next update.

5. How: **Technique, Design or Decoration** and **Color**: these are somewhat unusual for ontologies, in that they model how objects are made or finished rather than what materials are used or how the objects are used. AAT [3] does model some aspects of this, but in much less detail. In particular, we mine motifs and designs and organize them by design types and specific motifs. Although this area is somewhat sparse due to the amount of original metadata that provides this information, we think this may be expanded by visitors via a tagging model (see also Future work, below). Initial feedback indicates that even with partial coverage, these facets present interesting ways to explore and understand the collection.
6. Why: **Use or Context:** We adopted this in favor of a more traditional What facet in which the nature of the objects is commonly modeled. Some traditional subject indices such as LCSH [5] and Dewey [4] tried this at least to some extent, as in the ***TX - Home economics*** heading of LCSH, and the ***600 - Technology*** subtree of the Dewey classification. This facet required considerable discussion and revision as we sought to emphasize the Why over simply What, and still be sensible to most users. This facet still has some issues as a result - some common objects may not be in obvious places. For example, the ***Mask*** concept is located under ***Religious & Spiritual Life -> Ceremonial Items***, which may not be an obvious place for many users to look. This is mitigated by a search mechanism that lets user type in keywords to locate concepts (as well as free text tokens in the original metadata).

We also considered and rejected several other facets, including Religion and Language. This decision was partly because of the sensitivity of the topics, and partly because these facets did not provide sufficient additional power of resolution above and beyond the other facets.

5.2. Design of ontology structure

Although the project scope increased to support text mining, the basic structure of the ontologies was unaffected. Based upon our experience with facetted search [13, 9]) as well as the theoretical grounding [8, 11], we had resolved to use a facetted ontology. These make search easier as they can model a large number of queries efficiently and in a relatively easy to understand manner. The value of facetted ontologies has also been characterized as their ability to easily and cleanly model polyarchy. Given the extent and breadth of the museum collections, it was important that visitors be able to compare objects across cultures, locations, techniques, etc. In addition, the facetted browsing interface provides an excellent overview of the collection by indicating the number of results in various categories of each facet (see Figure 17).

Figure 18 — Facetted browser showing results counts per category for query on "basket."

Once the primary audience had been identified as non-experts, we also resolved that the language and organization of the ontology facets should be generic rather than jargonistic. We eschewed Latin names for plants and animals (although these are supported as synonyms), as well as jargon from anthropology and/or from the domain of museum collections management. By the same token, we wanted the *organization* of concepts to be accessible to a broad range of users, rather than following strict scientific rules. We favored common, folk categories over formal Linnaean taxonomies. For example to organize animals used as design motifs, we used a simple grouping into *land animals* vs. *sea animals* (including fish, mollusks and marine mammals) vs. *birds*.

The simplified organization may put off some academic researchers who would like a scientifically based organization, despite the general utility for a broad range of museum website visitors. Nevertheless, we discussed the idea of providing alternate facets to different audiences. In this scenario, researchers would self-identify in a profile page, and would then be presented with an altered UI based upon different facets. A variant on this idea would provide additional detail below the level provided for most visitors, and generally only of interest to researchers (e.g., specific species of a type of shell used or detailed provenience (location) information). In the general case, any subgraph of an ontology could be associated with certain roles, and the UI would select facets or filter categories for visitors identified to a given role.

For the initial deployment, we have concentrated on the primary audience (and so a single

set of facets), but may explore these personalized ontology options in a later phase of the project.

5.3. Design of ontology formalism and expression

When the project scope changed to include text mining, this dictated a significant change in the expression of the ontology. We had to adopt something that looks more like a categorization ontology or thesaurus. This is expressed in XML for easy manipulation and interchange. For example, we built XSLT stylesheets to produce HTML views of the ontology. Because of the relatively lightweight nature of the application at this point, we did not need to develop a formal DTD or XML Schema definition, but rather use an ad-hoc XML syntax. The schema models the facets as standard trees with narrower terms grouped under the respective broader terms. In addition, however, the schema includes rich support for declaring hints and rules to support the linguistic processing of the corpus. We began with a traditional categorization schema but then expanded this to accommodate entailment rules. The result is a relatively simple and easy to edit structure that produces a very rich set of potential index terms used by the concept matching tools (discussed below). Figure 18 presents an excerpt of the ontology to illustrate the schema.

```
<taxonomy id="TechniqueDesignorDecoration" title="Technique, Design or Decoration">
  <noiseToken value="small"/>
  <noiseToken value="minute"/>
  ...
  <reduce from="carved ivory" to="carved"/>
  <reduce from="carved wooden" to="carved"/>
  <heading id="Undecorated" title="Undecorated">
    <synonym value="plain"/>
    <synonym value="No decoration"/>
    <excl value="plain weave"/>
    <excl value="plain-twined"/>
  ...
  <heading id="FigurativeOrRepresentational" title="Figurative or Representational"
    nomatch="true">
    <!-- These are really fallbacks if we do not latch anything underneath -->
    <synonym value="figurative"/>
    <synonym value="mask"/>
    <!-- These will all apply to all "token" elements under this subtree -->
    <prefix value="carved "/>
    <prefix value="carving of a "/>
    <prefix value="figure of a "/>
    <prefix value="wooden "/>
    ...
    <!-- token name with each of the following suffices appended -->
    <suffix value="-shaped"/>
    <suffix value=" decoration"/>
    <suffix value=" motif"/>
    ...
    <heading id="HumanShapes" title="Human Shapes">
      <synonym value="anthropomorphic"/>
      <synonym value="anthropo-"/>
      <synonym value="devil"/>
      <excl value="dust-devil"/>
      <excl value="devil peak"/>
      <suffix value=" bust"/>
      <suffix value=" effigy"/>
```

```
        <token value="human"/>
        <token value="man"/>
        <token value="woman"/>
        <token value="child"/>
        <token value="bride"/>
        <token value="groom"/>
        <token value="dancer"/>
        <token value="musician"/>
      </heading>
      <heading id="AnimalShapes" title="Animal Shapes">
        <synonym value=" zoomorphic"/>
        ...
        <heading id="Bird" title="Bird" nomatch="true" astoken="plural">
          <heading id="Crane" title="Crane" nomatch="true" astoken="plural"/>
          <heading id="Pelican" title="Pelican" nomatch="true" astoken="noplural">
            <token value="pelican"/>
            <token value="pelicans"/>
            <token value="pelikan"/>
          </heading>
          ...
        </heading>
      </heading>
    </heading>
  </taxonomy>
```

Figure 19 — Excerpt of ontology illustrating schema

The excerpt is from a facet that models design motifs used in objects. Each facet is defined by a `taxonomy` element under the document root. Each `taxonomy` has `heading` children for the top level categories, and `heading` elements nest to present narrower concepts. Each `heading` has both an XML `id` value (for internal reference within the schema) as well as a display string in the `title` attribute. The ***Undecorated*** heading provides a simple example of a concept that has several `synonyms` and `exclusions`, used to improve the text-mining recall. The `synonyms` are related to Soergel's "lead-in terms" in that they function as alternate ways of indicating the given concept. The `exclusions` are terms or phrases that contra-indicate the concept. These provide a means to *latch* a given concept using a common token that has homographs or polysemes, by providing the exceptional contexts for the *other* meanings. Especially in a collection with a focused domain such as this museum collection, it is often easier to identify and describe the predominant usage with simple rules and add exceptions, than to fully characterize the predominant context(s). In addition, the text mining performance is often better if there are relatively few exceptions to check among many possible matches.

Another simple feature of the schema is the designation of *noise* tokens (in the `noiseToken` element), analogous to "stop words" in other systems. These are tokens with no utility for disambiguation (e.g., "frag" and "fragments" in the Materials facet). It should be noted that these are defined on a per facet basis. Thus while "small" has no utility in disambiguating design motifs, it might be significant for the ***Use or Context*** facet. The noise tokens are elided from the original metadata strings before we attempt to latch concepts. The reduce element provides a related feature that simplifies common n-grams. These rules are for cas-

es where one of the tokens cannot be removed in all cases without impacting the categorization, and are just a more conservative form of noise token removal. In the example above, removal of all instances of "wooden" would preclude matching things like "wooden bird" as a bird motif in a carving. However, since "carved bird" will latch the concept as well as "carved wooden bird", we can safely remove the extra word ("wooden") in that context.

The simple synonyms are fine for many materials and some specific concepts with distinct names, such as place names and names of cultures. However even with exclusions, the synonym tokens are insufficient for many common concepts, especially those involving polysemic tokens. To address this shortcoming, we added support for entailment patterns that function as synonyms but are often entire phrases rather than just terms. The model combines tokens designated for each concept with prefix or suffix strings to produce synonym-like strings. For the most part these phrases look like those used in entailment models for knowledge extraction (which was the inspiration for the approach). The `prefix` and `suffix` lists are valid for the entire subgraph, and so common ones are declared for broader concepts (e.g., ***Figurative or Representational*** in the excerpt), and are added to any `prefix` or `suffix` lists defined for narrower terms (e.g., "bust" and "effigy" under ***Human Shapes***). In the excerpt of Figure 2, the tokens "man", "woman", "dancer" et al. are combined with each `prefix` and `suffix` to produce entailment phrases including "carving of a man", "figure of a dancer", "human bust", et al. This exploded set of synonyms is combined with explicit synonym declarations (e.g., "anthropomorphic") for text mining, and is still subject to any declared `exclusions`. This approach yields a rich set of latching strings that increases recall, but is still relatively easy to manage in the ontology definition schema.

One last feature that we added to the schema is the ability to define additional inference rules that represent (often implicit) semantics in the domain. For example, if an object is made of gold, we understand that the object will have a golden color. Because this is generally implicit in our understanding of the world, it is unlikely to be expressed in the annotations for an object. In order to enrich the metadata resulting from categorization, we can model semantics like this using an `implies` element. In the example described, ***Material:Gold*** implies ***Color:Golden***. We have not taken full advantage of this yet, due to the labor intensive process of modeling these semantics. We would also like to make this more flexible to support rules like:

{UseOrContext:Mask `together with` ***Material:Wood)*** `implies` ***Technique:Carved***

Although we have not done as much as this as we could or would like to, we plan to add this soon. Since the process of mining the metadata for concepts is relatively inexpensive to run, we can continue to refine the ontologies, enriching features like implies, and the system will improve over time (see also the Future Work section, below).

5.4. Sources used for vocabulary designs

We drew upon a number of sources in the design of the vocabularies. The museum made some limited use of the Getty Thesauri, and AAT [3] in particular. We considered this and other traditional sources for inspiration. In addition, as mentioned above, we evaluated the organization and facets used in many libraries [5, 4]. However, many of the actual concepts we included came directly from the corpus. We extracted a dump from the collections management database as comma-separated values, split the free text for each value into phrases and tokens, and then ran simple usage statistics on these. We computed the usage on a column by column (i.e, field by field) basis to understand both the vocabulary as well as which columns should be mined for which facets (described below in the text mining and techniques section). We then worked through the lists in descending order to gather up terms and patterns. This was a somewhat painful process, but yielded an ontology that is representative of the collection and its idiosyncrasies.

We also drew upon the domain expertise of the museum staff. For example, the collection managers and other staff (all generally trained in anthropology) are familiar with the many cultures (and associated names) represented in the collections. However, where they are often content to work with large flat organizations of these cultures (e.g., a simple list of about 400 Native American cultures), our constraints for the UI required us to develop some deeper and narrower organization of the cultures. Initially, their tendency was to use a geographical organization based upon political regions (e.g., U.S. states), at the same time that they noted the perils of doing this (notably the fact that many cultures did not "belong under" a single state, but spanned a region. We spent some time talking about facet design and keeping the facets orthogonal to one another, but there was no accepted grouping of all cultures into nice neat super-cultures, and so we fell back on geography in some cases. To mitigate the misalignment problems between cultural range and modern political designations, we opted for more vague geographical regions like "Northwest U.S." This is not entirely satisfactory, but works reasonably well as a compromise.

Finally, we drew upon our common sense understanding of certain domains, especially to organize concepts for public rather than academic access. For example, we favor common names over the Latin designation for the many mollusk species represented in the collection, and we use common usage categories like ***Precious And Semi-Precious Stones*** to group ***Amber, Garnet, Ruby***, et al. rather than precise geological/gemological taxonomies.

5.5. Text mining tools and techniques

We built a simple tool that scans configuration files, imports and exports vocabularies, and performs various analysis and mining activities on the database dump from the museum collection management system. Since we want to make the code for this (along with the rest of the Delphi infrastructure) available to the community as an Open Source project, we resolved to use a portable application framework and settled on Java. Although the UI

support in Java is poor, the text mining application does not require much user interaction. The heart of the application focuses on XML and string processing and some data structure manipulation, support for all of which are fairly good in Java. In addition, we had some support classes from an earlier Java project that we wanted to leverage.

The tool is currently very rough and needs to be redesigned for easier use, but the main components of the actual processing are there and working. Figure 19 illustrates a process flow diagram for the text mining application, and the primary system modules. The modules are oriented to the associated configuration and data files, and some basic operations that can be performed:

1. A Dump-Column Configuration module reads an XML configuration file that describes columns in the DB dump (see Appendix), and provides the information to other modules (see step 1 in Figure 19). Each column in the dump is identified by name (requiring that the dump file have the names in the first row of output). For each column, the configuration specifies token separator rules (since the functional syntax of each column varies with use), global noise tokens to be ignored and other global parse cleanup (e.g., mapping embedded newlines in strings to spaces). In addition, the file specifies which facets should be mined from that column and specifies the estimated reliability for facet concepts that are latched in that column. For example, there is a provenience column in the Hearst museum metadata that (theoretically) specifies where the object was originally found. We mined this column for concepts in the Location facet, and with very high estimated reliability. Because humans are not consistent or always well trained in using a metadata system, location information is sometimes found in a column labeled "Medium." Therefore, we also mine the "Medium" column for the Location facet, but we use a lower estimate for reliability because of the increased risk that false matches will creep in from that context. The reliability only impacts the ordering of search results, and we have not yet evaluated the effectiveness of the reliability estimates.
2. A Vocabulary scanner module reads in vocabulary dump files from a metadata database (where this exists), and will export the vocabulary in our XML schema (see step 2 in Figure 19). We used this functionality at the beginning of the project, since the museum staff were more comfortable working in File Maker Pro (FMP) database tools than editing XML. Once the vocabularies/ontology began to settle, and especially as we added more decoration for the linguistic processing, we needed the expressive power of our XML schema and so stopped converting from the FMP dumps. We may need to address the issue of importing our ontology into the collection management system (CMS), and/or syncing our ontology structure and vocabulary to that used in the CMS. An XML ontology scanner module reads in the XML ontology file, builds the internal ontology data model, expands the entailment phrases, and assembles a set of hash tables for fast processing. This module leverages some core classes that represent a facet and a taxonomy node respectively.

3. A MetaData Reader module reads the data dump (a CSV file) and handles the global linguistic cleanup described in the dump-column configuration (see the first parts of step 3 in Figure 19). It provides various utility methods to access the column data (by column name or index). This module must also handle issues such as merging multiple rows in the dump for a given object (because the dump file is often a large OUTER JOIN of tables, some duplication of information can result; we require only that the dump be sorted by object ID, and then we merge actual rows into logical rows that merge all the actual rows but elide duplicate information).

The main application module contains all the logic to integrate these services and perform the text mining process (see step 3 in Figure 19, although this description provides slightly more detail). This process consists of the following basic steps, which are run for one facet at a time over each logical row, and for each column. We could run multiple facets at once (and somewhat reduce total scan time), but we tend to work on one facet at a time and so this works better in practice (there are also some issues with noise token handling that arise if we run all facets for each row). The steps involved are:

1. Noise tokens are removed. This just a string replace operation performed on the input string for the column.
2. Strings are tokenized into phrases (using the configured token/phrase separators), and as well into words.
3. Reduce operations are performed to align common patterns (as described above, this is analogous to stemming, but much simpler).
4. The resulting phrases and individual words are checked in the hash tables. For phrases with more than a few words, we also assemble n-grams of individual words and check these against the hash tables. We only build n-grams up to a maximum length, reflecting the length of the entailment phrases used in the ontology (the current default is 4). We may need to refine this and allow configuration of this maximum, and whether to use this fallback at all (depending on the facet). We are still evaluating the effectiveness of this approach, but based upon a simple analysis it does seem to increase recall without sacrificing precision.
5. Where matches are found to the hash tables, exclusions are checked. Because many of the exclusions are phrases rather than simple tokens, and may not even involve the matching token, the exclusions for each matched concept are checked as a raw string compare against the full string for the column.
6. All remaining matches (those not excluded) are assembled into a set. We process all the columns for a given object and add matches to the set as we go. If two columns both indicate a given concept (which does happen), we use the higher reliability value (from the column configuration) for the category association.
7. Once all columns have been mined and the set assembled, we emit the category associations as rows in an output MySQL dump file. The resulting file can be very quickly imported into the application database for the website.

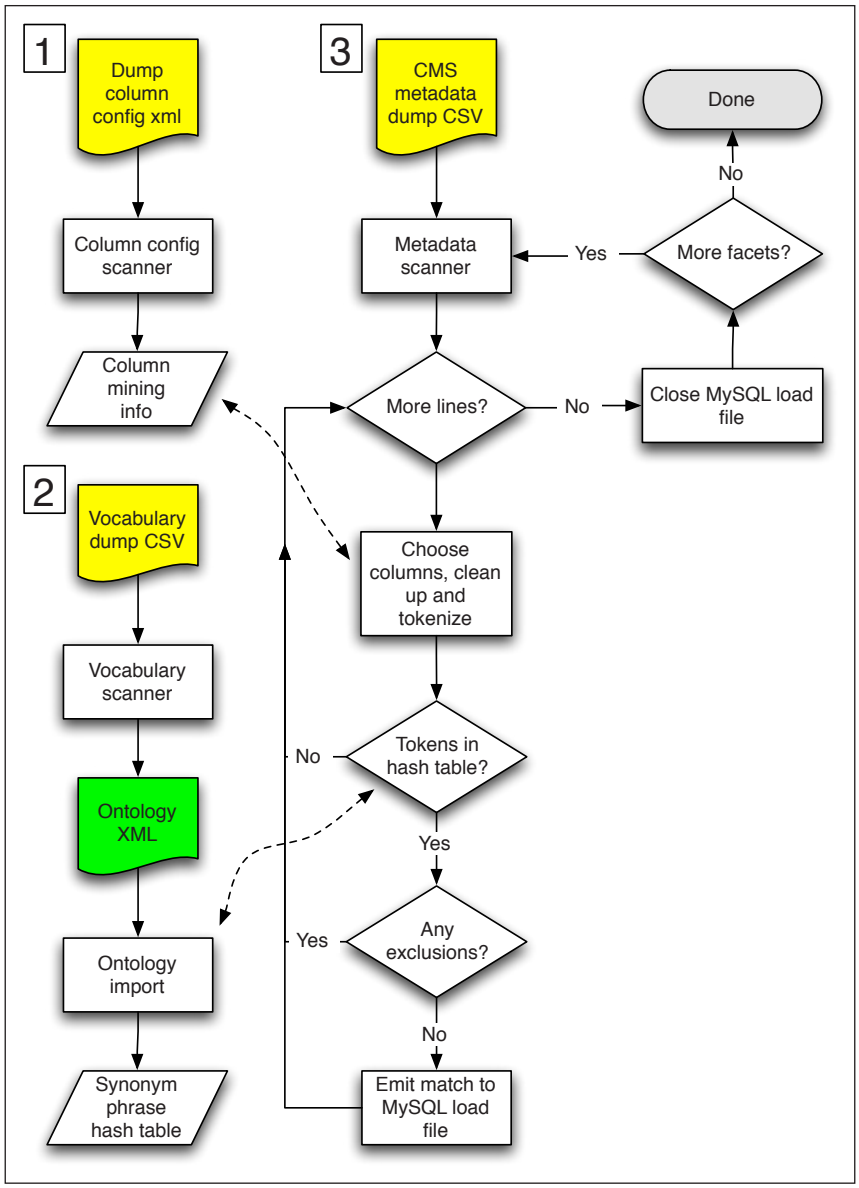


Figure 20 — Data Mining Flow Diagram.

We made a conscious choice to avoid parsing in a traditional NLP sense, to simplify the code and to achieve faster performance (not to mention that we would need an alternate language model for the writing style used in these annotations). We have also chosen to expand synonyms to include plurals and other grammatical forms rather than stemming the input tokens. Again, we think that the tradeoff of larger hash tables against code complexity and mining speed is a good one for this application.

5.6. Evaluation metrics

Any project like this, especially when used for information search applications, begs the question of how well the ontology works in practice. The traditional measures used are *precision* and *recall*, and sometimes utility. Soergel [12] also describes measures of *discrimination* and *novelty*, but notes that for all of these metrics a number of problems arise in practice:

> Requirements for recall and precision vary from query to query, and retrieval performance varies widely from search to search, making meaningful evaluation difficult. Standard practice evaluates systems through a number of test searches, computing for each a single measure of goodness that combines recall and precision, and then averaging over all the queries. This does not address a very important system ability: the ability to adapt to the specific recall and precision requirements of each individual query. The biggest problem in IR evaluation is to identify beforehand all relevant documents (the recall base); small test collections have been constructed for this purpose, but there is a question of how well the results apply to large-scale real-life collections.

Given the very large and diverse collection we are working with, and the uneven (often sparse) original metadata we are mining, these problems of evaluation become particularly acute. Nevertheless, we have a number of ideas for evaluating the results that we hope to pursue:

1. We want to produce reports for all column text that does not get latched to a concept in the ontology, and then produce statistical usage reports for these items in much the same we did for the original analysis of the metadata when we were assembling the concepts for the ontology. We hope that this will show us the terms we are missing, and so guide additions to the concept set and/or the linguistic hints on existing concepts.
2. We hope to leverage the visitor community to both assess and improve our metadata. The website supports tagging of objects in the collections. We would like to request of certain enthusiasts within the community that they use certain specific tags to indicate incorrect or missing category associations. In addition, we will analyze the general patterns of tag annotations for missing concepts and other implicit feedback on the ontology and the categorization tools.

Nevertheless, we have some basic measures of the quantity of concepts we have extracted, and other simple quantitative measures. These are presented in Tables 1 and 2, below.

Facet	Use or Context	Location	Material	Technique, Design or Decoration	Color	Cultural Group
4-Apr-07	130,400	562,400	118,800	23,800	10,200	27,600
5-Apr-07	260,656	2,973,112	367,200	42,558	11,200	111,200
21-Apr-07	417,131	2,186,152	374,199	57,023	11,092	85,284

Table 7 — Number of objects categorized per facet, for different software versions.

Table 7 presents the number of category associations found in the entire collection of over 600,000 objects. Note that we have a location for a significant majority of the collections, but that some facets (notably Color) are much more sparsely represented. Especially given the relatively straightforward task of latching color concepts in this collection, we believe this reflects the original metadata, and illustrates the constraint that we can only mine what is already there.

Some other points are worth noting. Between 4th and 5th April, we added the logic to infer broader concepts, which accounts for the significant jump in values. ***Color*** is a relatively shallow facet and so is less affected by this. Between 5th and 21st April two main changes happened with partially offsetting effects on the numbers. First, we fixed a bug that was occasionally creating duplicate associations. This accounts for the significant drop in associations for the ***Location*** facet, and the modest drop for ***Color***. Second, we did a significant revision of the ***Use or Context*** facet, and made some improvements to the ***Materials*** facet and the ***Technique, Design or Decoration*** facet. The new ***Use or Context*** facet latched many more concepts, even as duplicates were removed. We saw a modest increase in the match count for ***Material***, which would have been somewhat larger were the duplicates not removed at the same time. We saw a more significant increase in the count for ***Technique, Design or Decoration***. While difficult to interpret in isolation, these figures nevertheless show the impact of revising and improving the ontologies. We believe this is comparable to increasing the recall of the system, and so translates to useful metadata for the end user.

Number of categories in all facets	7,149	100%
Number of categories matched	5,522	77%
Number of unmatched categories	1,627	23%

Table 8 — Ontology latching effectiveness.

Table 8 illustrates one of the challenges ahead of us, and the need for deeper analysis of the ontology and text-mining support. Only 77% of the categories we defined in the ontology were successfully latched within the collection. Some of these may be inclusions from standard vocabularies like gazetteers for ***Location***, in which case the lack of matches is

understandable and harmless. However, we need to investigate whether we have defined categories that are lacking sufficient or correctly specified linguistic hinting to properly latch the concepts during text-mining.

One additional metric is the performance of the text mining software in terms of compute resources required to run. We ran the text-mining application on a modest workstation (Intel, 3GHz CPU with 1GByte RAM) and on a modern business laptop (Intel Dual core 1 GHz Mobile CPU, 1GByte RAM). The two platforms are roughly comparable and we saw similar performance numbers on the two. The Java-based application can load the ontology, expand the synonyms and build the associated hash tables in a few seconds. Mining a 315 MByte database dump of 615,000 objects across all facets (7K+ categories expanding to tens of thousands of synonym terms and phrases) takes roughly a half an hour. Loading the dump files into a MySQL database running on the laptop takes about five to ten minutes. Loading it on a quad-core Xeon server takes only a few minutes. Our decisions to favor processing speed in the design of the ontology and mining model seem to have paid off well.

5.7. Conclusion

We built a practical ontology that combines linguistic features for text mining with design constraints to yield a pleasing facetted browser UI. We combined bottom up analysis of the original metadata corpus with exemplary models of organization for public-accessible indexing and our own sense of the audience to construct facets and a concept organization that we believe will be useful to a broad range of visitors. We described the resulting principles we followed to design an ontology that supports a facetted browser UI. We designed a novel approach to specifying entailment phrases for use in text mining, and implemented an efficient text mining engine based upon the ontology specification. Although we have not had sufficient experience to judge visitor response or perform a thorough evaluation of ontology quality, we believe there is sufficient positive feedback to merit continued development. We look forward to integrating a model of community maintenance of the ontology, using lessons learned in social media and other commercial systems, to make the system continually improve in a sustainable manner.

6. Website Implementation

6.1. Requirements negotiation

In January of 2007, we began a technical requirements negotiation with Michael Black of the Hearst Museum. In order to maximize code portability we wanted to develop on a common LAMP (Linux, Apache, MySQL, PHP) stack. To keep our development options as open as possible, we requested the latest versions of Linux, Apache, MySQL, PHP as well as SSH access. To accommodate the Museum's collection we estimated about 300 GB of storage space would be necessary to house both the database and object images.

In February 2002, we were offered space on a server located in UC Berkeley's Information Services and Technology data center. The server specification met or exceeded all of our requirements. However, late in the project we encountered storage space limitation due to unexpected use of the server by another project team using the server.

Hardware	CPU	4 x 3.0 Ghz
	RAM	4GB
	HDD	538GB
Software	Linux	Fedora Core 5
	Apache	2.2.2
	MySQL	5.0.27
	PHP	5.1.6

Table 9 — Specifications of our development server.

6.2. System architecture diagram and description

We designed Delphi to be a supplement to, and not a replacement of, a museum's existing collection management system. Delphi's purpose is to expose the collection on the web and foster community around museum objects. Purposely out of scope is support for things such as accessioning and deaccessioning, data entry, loan and exhibition management, etc. We came to this decision in part because the Hearst Museum already licensed a product to perform these functions (TMS by GallerySystems). It was also our assessment after talking with members of the Hearst Museum staff that creating a gallery management system would require more museum experience and time than was available to our team.

Delphi is composed of five main modules: Authentication, Sets, Facet Browser, Tags, and Front Page. These modules are used in conjunction with database and presentation abstraction layers to generate Delphi's pages. The objects in Delphi's database come from TMS, the Museum's collection management system. The metadata dump from TMS is processed

through ontology tools we developed that output cleaned object metadata and facet vocabularies. Figure 20 shows how Delphi's modules work within a larger system to build and serve pages to users.

The Delphi site pages are implemented with HTML, CSS, and JavaScript on the client-side, and PHP, Smarty templates, and a database on the server-side. The pages were architected such that markup, behaviors, and presentation are modularized as much as possible. For example, the markup uses Smarty tags, JQuery, and an imported style sheet to separate PHP, JavaScript, and CSS from the HTML. This made modifying the code base easier, as most changes of a specific kind - such as layout - could be made in one file instead of across several.

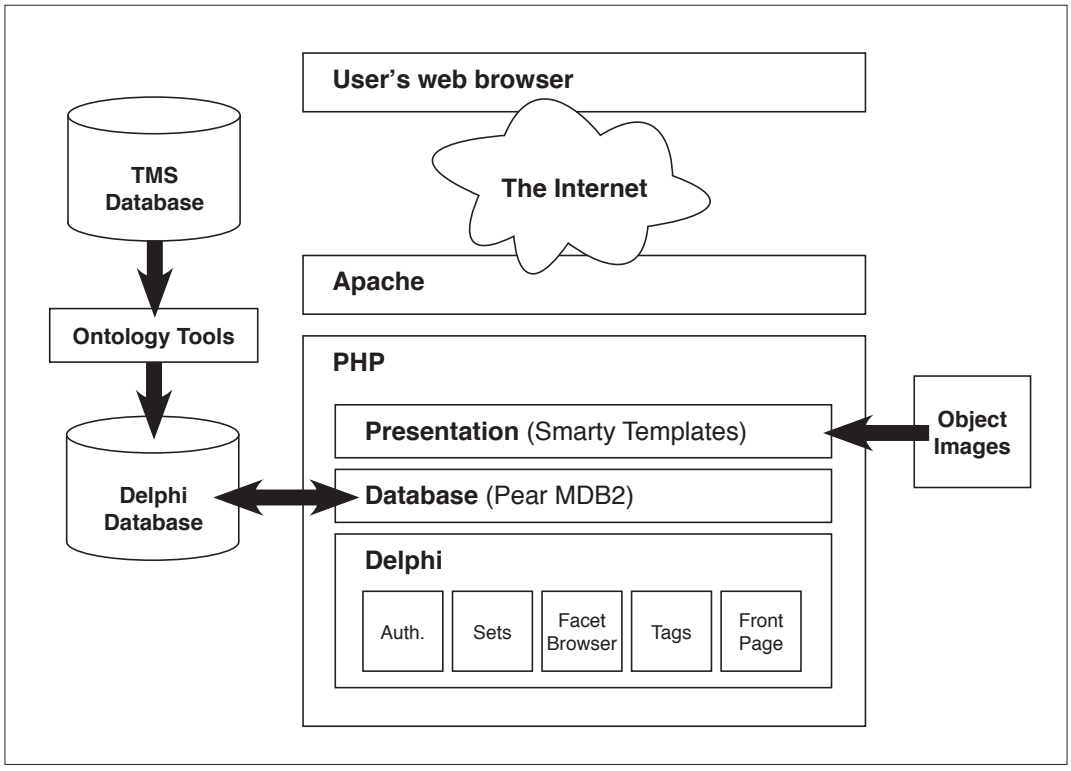


Figure 21 — Delphi system architecture.

6.3. Description of site pages

Figure 21 shows all the pages within Delphi. Following the diagram is a description of each page along with a screenshot.

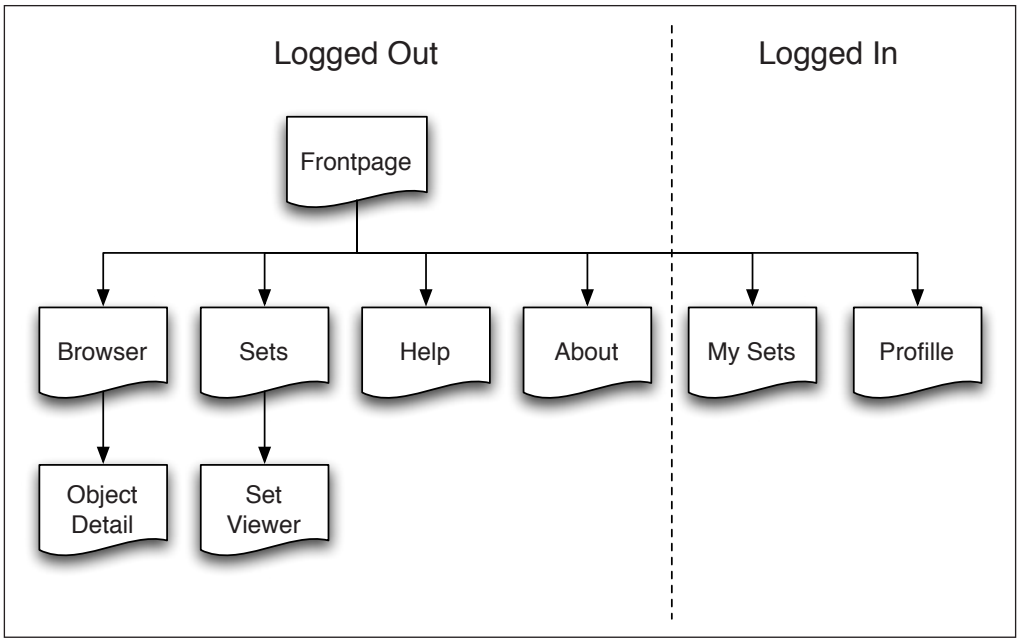


Figure 22 — Delphi site map.

6.3.1. Front page

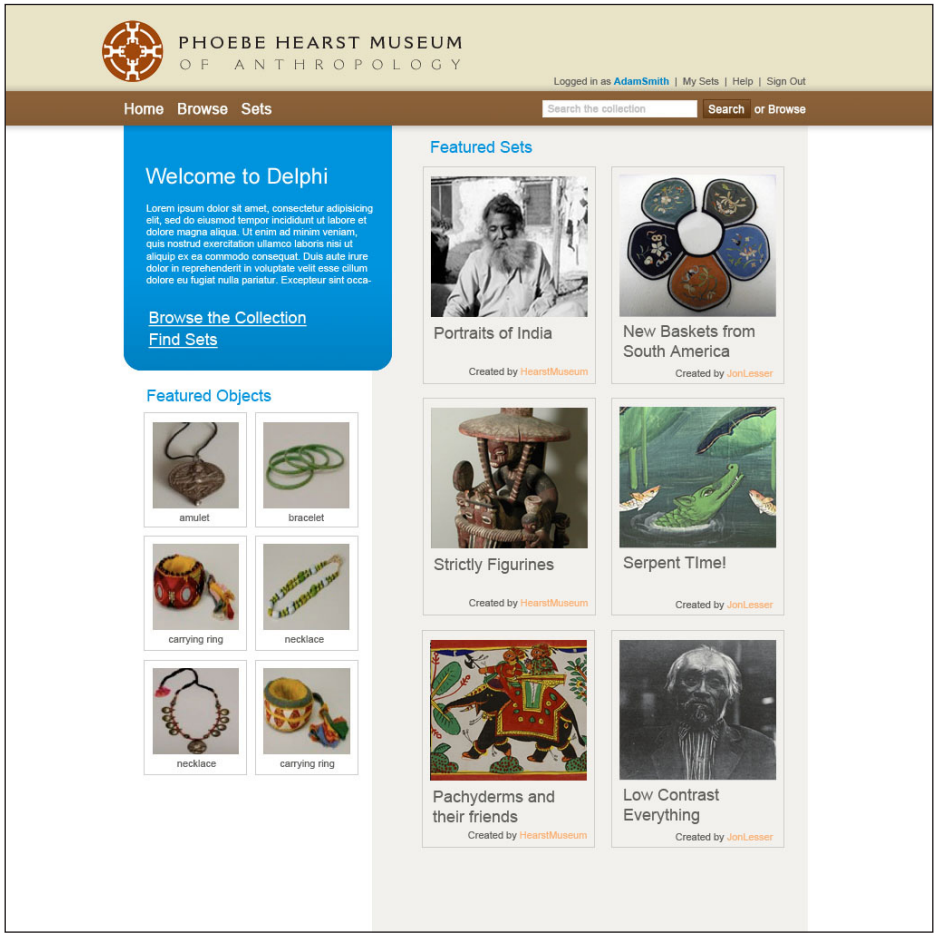


Figure 23 — Delphi's front page.

The front page introduces the user to Delphi with some welcoming text and a diversity of ways to dive into the collection. There are three primary elements: featured objects, featured sets, and links to search or browse the collection.

6.3.2. Browser

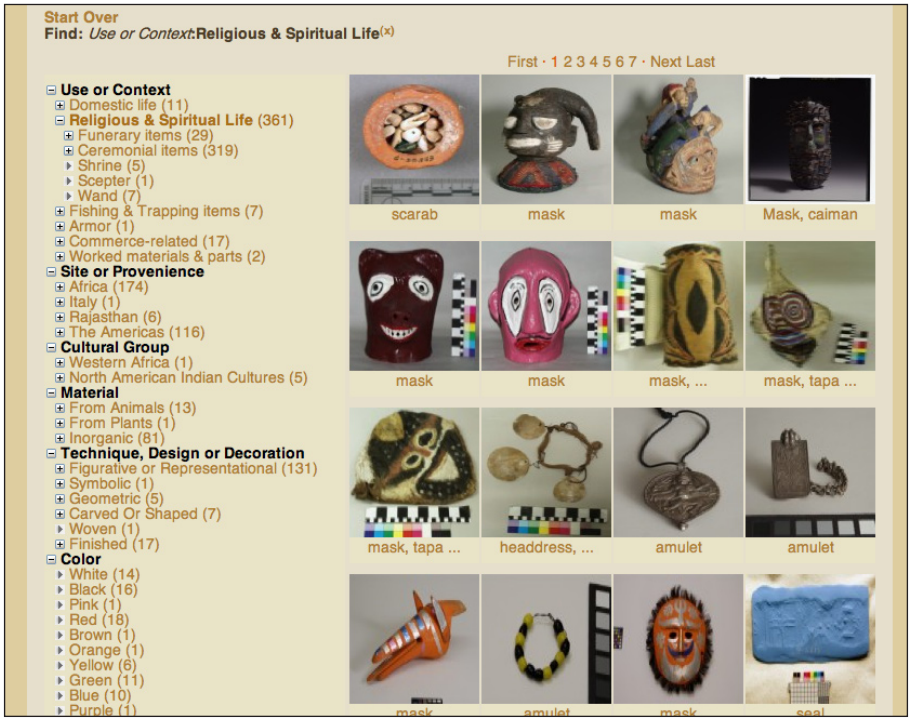


Figure 24 — Delphi's facet browser.

The facet browser allows users to both browse the collection and refine searches. The facets give the user a sense of the collection's breadth while the parenthetical counts give a sense of the collection's depth. As categories are selected on the left, the results are refined to represent the intersection of all the selected categories. All the selected categories are displayed along the top of the page. Links next to each selected category allow the user to remove that category from the current query.

6.3.3. Object Detail

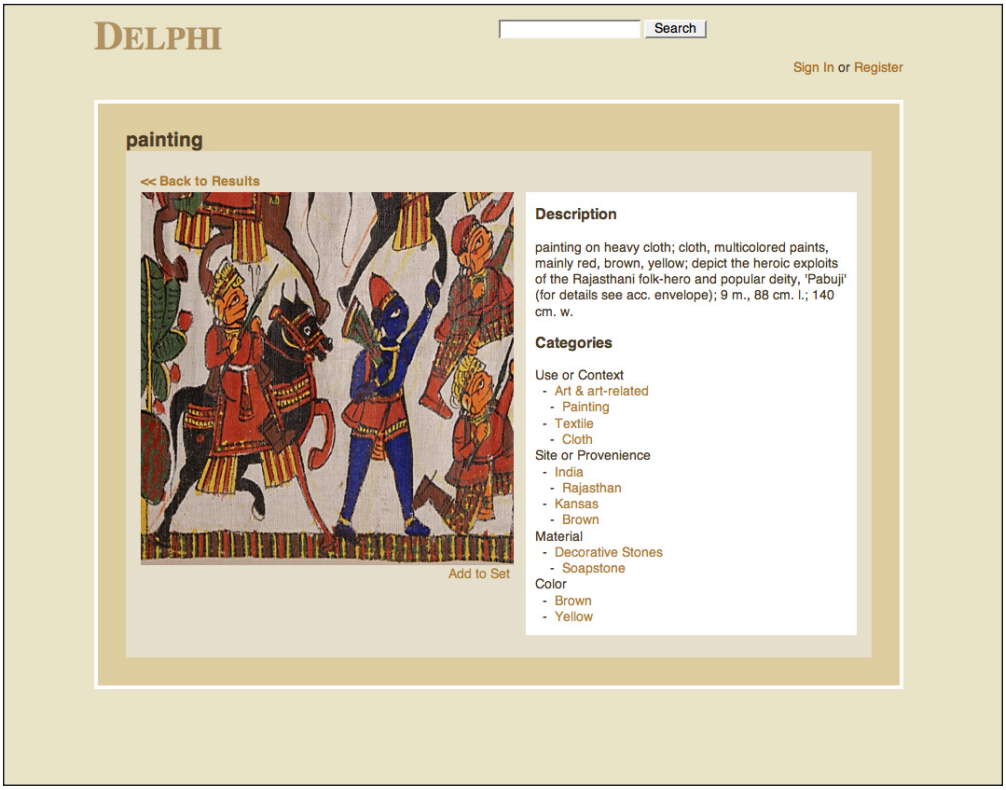


Figure 25 — Delphi's object detail page.

The object detail pages feature a high-resolution image of the object that can be zoomed and panned. The object's description and a list of categories associated with the item is shown in the right column. Users can add the object to a set of their own with the "Add to Set" link.

6.3.4. Set Viewer

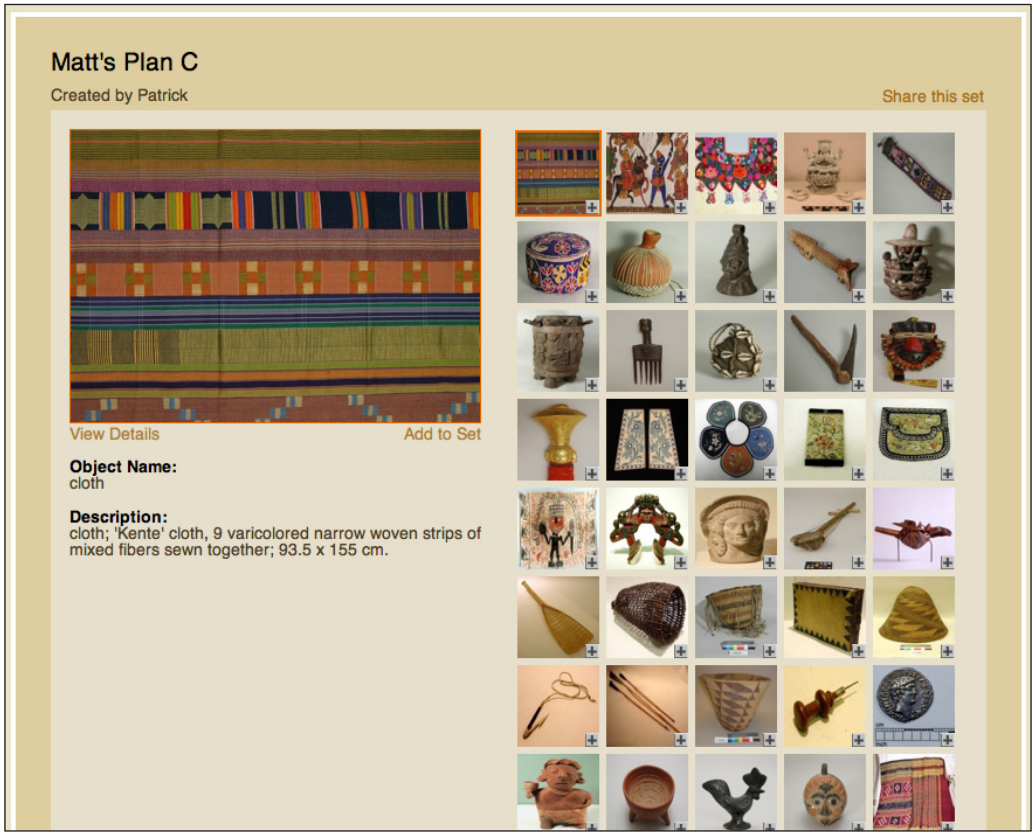


Figure 26 — Delphi's set viewing page.

The set viewer lets users explore sets of museum objects created by themselves or other users. Clicking on a thumbnail on the right brings up object details and a larger image on the left. Users can add the selected item to set of their own with the "Add to Set" link. They can also email a link to the set with the "Share this set" link.

6.3.5. My Sets

Figure 27 — Delphi's my sets page.

This page lists the user's sets with links to view, delete, and share each set.

6.4. System modules

Delphi's PHP code is divided into modules. These modules represent the business logic required to fulfill the module's function. For example, the sets modules knows what information needs to be pulled from the database to present a set to a user. The modules do not contain any HTML code, however, they do know which templates to use.

6.4.1. Authentication

Delphi's authentication system allows users to login, logout, and register for new accounts. A user can also request a new password to be emailed to them. Logged in users can change their password and email. When logging in, a user can ask to be remembered and Delphi will place a cookie in the user's browser. See the diagram below to see how the different functions of the authentication modules are presented to users.

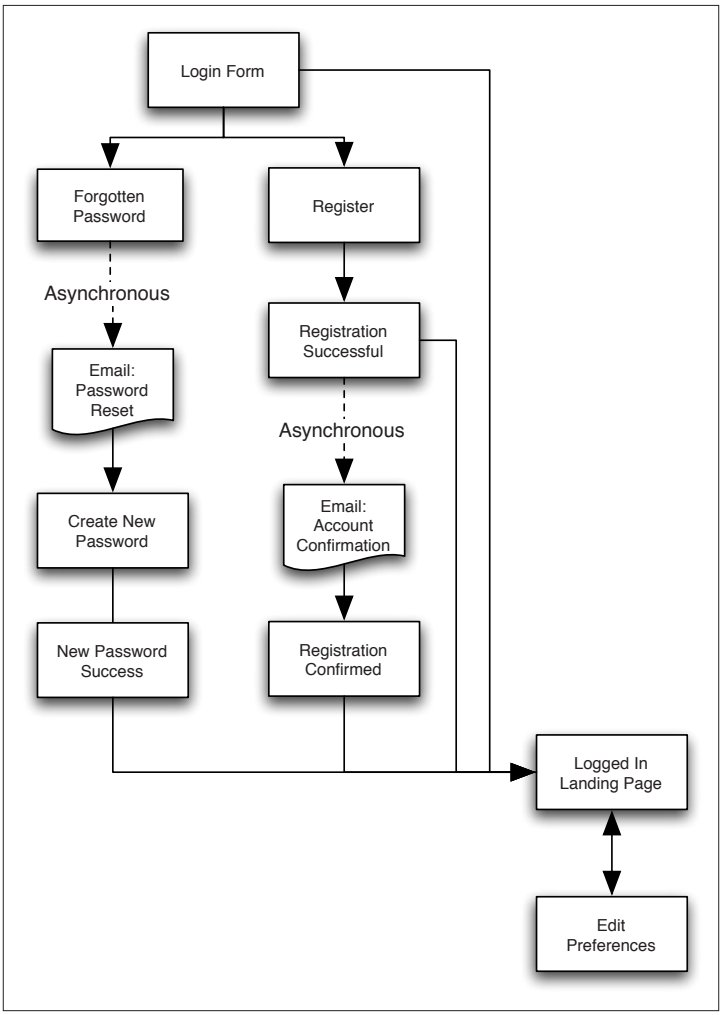


Figure 28 — Authentication flow diagram.

6.4.2. Browser

The browser module is composed of three parts: 1) a facetted browser, 2) a query results page, and 3) an object details page. The user selects facets from the facetted browser to refine a query against the Delphi database that returns a set of results with thumbnail images that link to object details pages.

The facetted browser is a core piece of the search and browse UI in the Delphi system. As shown in Figure 28, it provides both an orientation function with the matched-objects counts, as well as a query refinement function via links. The functionality is built out of the database, using a few PHP support classes and the main PHP file for the results page, together with some JavaScript that runs on the page. The facet UI logic for the browser.php page that supports browsing the entire collection is largely the same as for the results page, and so we just describe facetBrowse.php.

The module runs through a multi-step process to go from query parameters to MySQL query and then to the resulting HTML browse tree. The process is illustrated by the figure below.

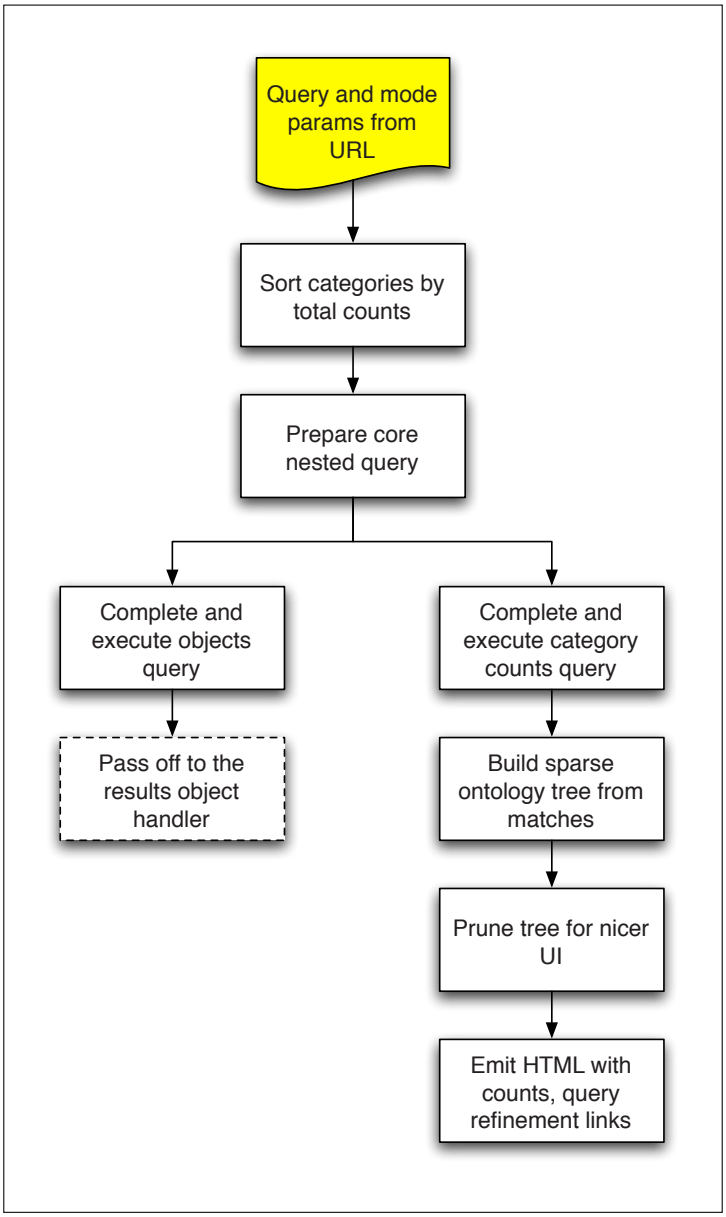


Figure 29 — Facet browser processing model.

Once the categories and keywords have been extracted from the URL parameters, an initial query is performed to get the global counts associated with each category. This is used to re-order the categories before the code assembles the core MySQL query. We build a nested query to get the logical AND of all the categories (and/or keywords), using named sub-queries. For a keyword and two example categories, the core query looks like:

```
SELECT oc.obj_id from obj_cats oc,
  (SELECT oc.obj_id from obj_cats oc,
    (SELECT id as obj_id FROM objects o
     WHERE MATCH(name, description) AGAINST('mask') AND NOT o.img_path IS NULL) subK
   WHERE oc.obj_id=subK.obj_id and oc.cat_id=50006) sub1
WHERE oc.obj_id=sub1.obj_id and oc.cat_id=20000
```

For such nested queries, MySQL cannot always rely upon indexes as it does the joins among the sub-queries, and so it may resort to a scan of the temporary tables. To minimize the associated cost, we place the categories with the least matches on the inside (most deeply nested sub-query), and move outward in order. This minimizes the cost of the temporary table scans and greatly improves query performance. One the core query is assembled, it is wrapped once to get the objects that match:

```
SELECT SQL_CALC_FOUND_ROWS o.id, o.objnum, o.name, o.description, o.img_path from objects o,
  (SELECT oc.obj_id from obj_cats oc,
    (SELECT oc.obj_id from obj_cats oc,
      (SELECT id as obj_id FROM objects o
       WHERE MATCH(name, description) AGAINST('mask') AND NOT o.img_path IS NULL) subK
     WHERE oc.obj_id=subK.obj_id and oc.cat_id=50006) sub1
   WHERE oc.obj_id=sub1.obj_id and oc.cat_id=20000) tqMain
WHERE o.id=tqMain.obj_id limit 40
```

Using the SQL_CALC_FOUND_ROWS mode allows us to then query with "SELECT ROWS_FOUND()" to get the total number of objects, even though we only retrieved the first page of 40. This precludes a second expensive query to get the total count.

The core query is wrapped again to get the category counts for the result set:

```
SELECT c.id, c.parent_id, c.facet_id, c.display_name, count(*) from categories c,
  (SELECT oc.obj_id, oc.cat_id from obj_cats oc,
    (SELECT oc.obj_id from obj_cats oc,
      (SELECT oc.obj_id from obj_cats oc,
        (SELECT id as obj_id FROM objects o
         WHERE MATCH(name, description) AGAINST('mask') AND NOT o.img_path IS NULL) subK
       WHERE oc.obj_id=subK.obj_id and oc.cat_id=50006) sub1
     WHERE oc.obj_id=sub1.obj_id and oc.cat_id=20000) tqMain
   WHERE oc.obj_id=tqMain.obj_id) tqTop
WHERE c.id=tqTop.cat_id group by c.id order by c.id
```

We have considered creating a view for the inner query to improve performance, but it is not clear whether the MySQL query cache already takes care of this for us. In any case, once we have the results of the category counts, we use the Facet and TaxoNode classes to handle the results and construct a tree from all the categories with counts. Ordering by category id ensures that parent nodes will be seen before their child nodes, and so this is fairly straightforward. The resulting tree is then traversed and we remove nodes that are not useful to the user. In particular, we mark any node that has a single child and where the counts for the parent and child are the same. In this case, we prune the parent as it does not help the user understand the results any better, nor will it be a useful query refinement. We also perform various other heuristics to balance the resulting tree, and then we traverse it once more to produce the HTML for the facetted browser UI. In the output pass, we just keep track of nesting for indents, and build a link into each category that adds the

associated category to the current query string and defines an <A> element around the category name to effect the refined query.

Although the SQL query syntax can get rather complex and requires careful handling to optimize performance, the process is really fairly simple. Similarly, the logic to prune the tree and emit HTML is not particularly complex, although again care must be taken with details of the current query string. We are still experimenting with pruning strategies to produce the nicest UI for users. One area of future work will decorate the ontology further to help optimize the pruning strategy - this is discussed in the Future Work section, below.

The query results are presented on a webpage as a four-column grid of image thumbnails and object name labels. The thumbnails and object name labels are clickable links to the object details page of the represented object.

The object details page provides a higher-resolution of the image of the object as well as descriptive text. Implementation considerations for the pages mainly focused on sourcing the image and text.

The image, if it exists, is drawn from the Delphi database; if it doesn't, a placeholder is used instead. On details pages that have images, if Flash is installed on the user's browser, then the Zoomify image zoom-and-pan Flash applet is embedded in the place of the higher-resolution image.

The description is sourced from a concatenation of several columns in the database that have been determined to contain richer descriptive text. In addition to the concatenated text, the description includes output from a call to the facet browser, which returns the categories in each facet below a "generic" level of the facet hierarchy as determined by the facet pruning algorithm. These categories are presented as navigational links that returns a browser results page with the clicked-through category selected as the query.

The details page also includes integration with the sets functionality, providing a link that makes a JQuery call that allows users to add the object on the details page to one of their sets. We are also prototyping tagging functionality on the object details page.

6.4.3. Front Page

Implementation considerations for the front page included inclusion of a search box and the addition of the Featured Sets and Featured Objects functionalities, which highlighted areas of interest deeper in the Delphi site.

The search box is a PHP driven form that provides a front-end to the category and keyword search functionality of the facetted browser. Text in this box is submitted as an HTTP GET request and returns a browser results page with results from a latched category or keyword match.

The Featured Sets and Objects functionality is implemented through Smarty tags. The PHP behind the template retrieves the IDs of pre-created sets and the associated image for display on the front page in the case of Featured Sets and object ID numbers and the associated thumbnails and names in the case of Featured Objects. The Featured Sets are set by the site administrator from existing sets, while the Feature Objects are randomly selected on each page refresh from a handpicked group of objects with high-quality images and rich description text.

6.4.4. Sets

The sets module implements adding objects to sets, creating new sets, sharing sets, and of course viewing sets. The sets module integrates with the authentication module to bind sets to individual users. The database implementation of sets consists of just two tables. One for set metadata and one for joining set objects with the main objects table.

Any user can access any existing set, however, only logged in users can create new sets. Users can only add items to sets they created themselves, but users can "share" any set. When sharing a set, Delphi asks the user for the email address of the recipient and then emails a link to that recipient. In the future we would like to expand Delphi's community features by enabling sharing of sets between registered users and increasing the discoverability of sets created by other users.

6.4.5. Tags

Delphi's tagging system allows users to tag individual objects within the collection, search for objects that are tagged with a specific tag, and delete and rename individual tags. Tagging is designed to require user authentication so that each tagging operation can be recorded by the system for future analysis. The database tables are implemented in MySQL and the functionality is implemented in PHP. The following database tables support tagging:

» `tags` is the main table that contains tags, tag id numbers, and the total number of times each tag is currently applied to all the different objects within the system.

» `tag_user_object` table is used to support both tag management and searching functionality because it allows a search for the triplet tag-user-object relations in a fast, uncomplicated way. It contains the triplets of tag_id, tag_user_id, and tag_object_id. Each triplet uniquely identifies a specific tag, a user who is currently using that tag, and an object, for which this user uses this tag.

» `tag_usage` table is going to be used for future research of issues such as tagging use, connections amongst users, and areas of tagging activity. This table is used to store information about each tag: when was it created, added, or deleted, by whom, for what collection object and when.

6.5. Source code management and licensing

To avoid ownership disputes, we decided early on in the project to release Delphi under an open-source license. The Museum was very supportive of this decision. We initially chose to use the GNU Public License (GPL), but later switched to the Berkeley Standard License (BSD). The BSD license is less restrictive than the GPL and allows commercial developers more freedom to incorporate our code into their own projects.

Delphi's source code resides in a Subversion repository provided by SourceForge.net. Hosting our code repository on SourceForge allows anonymous public access to our code as well as a public record of our code changes. We would like to see Delphi live on beyond this current project. Public hosting of our code is the first step towards building an active development community. Of course assets belonging to the Hearst Museum such as object images and metadata, were never checked into the repository. Instructions for accessing the code repository are available on our SourceForge project page, http://sourceforge.net/projects/delphi.

7. Conclusion

7.1. Usability assessment key findings

Though response variables were important to measure, we mainly were interested in capturing qualitative measures from our post-test survey and by allowing free-form questions and comments during the actual testing process, when screen situations were "fresh". The following examples are important moments we tried to catch 'on the fly.'

» How did users navigate through the site (to the home page from other pages, to view object details, to add items to a set, etc.)

» whether the flow was natural to the users;

» whether they would want annotation capabilities at certain points

» whether they understood where they were in the site

» whether they could clearly describe what had just happened

What we looked for when each scenario was performed was 1) participant's ability to and 2) participant's method for carrying out each task. Would participants use links and/or buttons we had provided (would tools be noticeable?) or would they opt for keyword search? Back button? Were they satisfied by the choices we gave them, and on the pages where we gave them certain abilities? (See Appendix Pilot Test Tasks for more details). Additionally, we administered a post-test survey, with user satisfaction / ease-of-use questions based on a 1-5 Likert scale (See Appendix post-test survey results for complete data).

From our pilot test results, and also from our participant's answers on the post-test survey, we were able to make the following observations about our user interface design:

7.1.1. Key Findings

» Participants were able to complete all tasks quickly and with few errors

» Over reliance on Back button as opposed to "Back to Results" link.

» Users did not seem particularly interested in "View Details" page, feeling the information on View Sets page to be sufficient

» Participants expressed the need for a confirmation message when they had performed certain actions but also expressed strong negative feelings for having to click confirmation messages

» Plus signs are a plus! They immediately telegraphed to our participants that something could be done with the thumbnail image, and tempted the users to explore. Add them to larger images for a more unified experience throughout the site.

» Participants had an over reliance on keyword search box

» Participants liked the ability to browse the collection via "featured sets"

» Implement View Sets layout changes: move Set Description 'above the fold' to eliminate confusion with Object Description

» Private or public option: our participants wanted to add "notes to myself" about an image, or sets, as opposed to notes for the community at large

» Provide more info about object images on mouse-over.

» Participants rated their experience using the Delphi Museum Browser as very good (4.6 out of a possible 5) regardless of their computer competency level

7.2. Future work

There are a host of tasks we hope to address in the near future, as well as some more ambitious ideas we would like to pursue longer term. We have mentioned some issues in the sections above, to which we add the following:

1. While we have a vocabulary of sorts for the time periods in the original metadata, most of the terms are meaningful only to trained anthropologists, and so do not lend themselves to the public UI for search and browse. In addition, named time periods like "Early Roman" do not allow for comparison of objects across cultures. We have proposed a model that defines time ranges for periods and so converts the existing metadata to a common scale that can be used for all the objects in the collection (in the model of [7]). There are also some explicit dates in the original metadata, but these require additional processing to be used reliably. One the one hand, there are a number of syntax patterns in use ("ad870", "870 a.d.", "dated: 870") that must be parsed. On the other hand, many dates after about 1850 are (according to museum staff) just as likely to be accession dates as reliable dates for the object origin, and so should be used in the text mining. Because of these factors and general time constraints, we have deferred this work.
2. We are working on improving the UI that is generated from the ontology and the web database. As part of this, we want to be able to elide parent categories that are not really needed to help the visitor understand the context of a given concept. For example, if the location ***Sonoma, California*** is matched for an object, we do not need to show the ascendants of ***California*** (***Western U.S., United States, North America***) for most visitors. However, if we also elided ***California***, many users might not be clear that we were referring to the city in California. The points in each facet at which the concept can stand on its own without ascendant context is what we refer to as the generic level. We want to decorate the facets with this indicator to allow the UI to better adapt the presentation of category titles. Similarly, there is a level of detail beyond which non-expert users will generally be confused or simply uninterested, or that as a policy the museum does not want to present to the general public. Another example from the Location facet is the indication of sites where objects are found. These have designa-

tions like “CA-Nev-17” that are meaningful only to trained anthropologists familiar with a given area (in this case, a particular site in Nevada county, California). Further details on the location are sometimes available, but the museum staff do not want these made public for fear that “arrowhead hunters” will seek out the site and begin digging. We are considering decorating the public detail level so that the UI can prune out details for most users, but show the narrower (hidden) concepts for visitors designated as “researchers”.

3. We look forward to integrating tag annotations into the ontology once we get sufficient usage of this feature. If we get very many tags, we will need a means of filtering noise and vetting useful additions. We are separately developing a model for authority and reputation management and tools to support review of submitted annotations (including tags). The museum staff claims to have a host of retired curators, professors and graduate students who are capable of reviewing such annotations, and would be willing to do so. These online curatorial assistants will be identified and vetted by museum staff to ensure a baseline of reputation and expertise. The online reviewers will declare their area of interest within the collections, and so filter the annotations that they will review - in this way, they can remain within their domain of expertise. The support tools will enforce certain guidelines (e.g., that two online reviewers must vet and not reject an annotation before it is presented to museum staff for review). The tools will also track statistics of how often each online reviewer makes a decision that is accepted or rejected by museum staff, and as well how useful the resulting concept is (measured by the use in the application UI). The tools will track these same statistics for the original annotator, to identify and recognize expertise among the visitors. In addition to providing tools to help maintain the ontology, we think that this kind of system can further motivate contributions from the community.
4. We hope to deploy the system for other museums as well, at which point we will be able to evaluate how portable the ontology facets are, and how usable the text-mining tools are.
5. In order to make the site compelling enough to build a core of enthusiasts who regularly return to work with the site, we hope to leverage some best practices from commercial sites for social media. These include the formation of interest groups (a.k.a. “tribes”) around areas of the collection or themes in anthropology. By recognizing the areas of interest and expertise (from annotations, patterns of set-building, etc.), we can link people to one another to foster “tribe” formation. Leveraging the work of sociologists and successful entrepreneurs, and recognizing the connection between a museum site and a social media community, we think we can do something very exciting for this and (eventually) other museums.
6. People will return to and spend more time on a site where they feel a sense of community. To support this as well as general discovery, we would like to analyze and preset groupings of similar sorts of sets that different users created (e.g. 1)= “coins”: 1a) spanish dubloons, 1b) square viking coins 1c) beads used as money; 2)=”mummies:

2a) animal mummies 2b) egyptian mummies, etc.). Various metrics are possible, but it would be interesting both to connect sets in this manner as well as to connect the authors. Teachers working on a similar theme might well share and benefit from one another's work.

7. My Sets Page: We would like to give users more set-managing capabilities (e.g., removing objects, batch-processing, etc.)

8. References

[1] Buckland, M., Gey, F., and Larson, R., *Access to Heritage Resources Using What, Where, When, and Who.* (Presentation at Museum and the Web, San Francisco, April 11-14, 2007.) (HTML)

[2] Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., and Robbins, D. C.. *Stuff i've seen: A system for personal information retrieval and re-use.* In SIGIR, 2003.

[3] J. Paul Getty Trust, Art & Architecture Thesaurus (AAT). Available online (as of 5/1/2007) at: http://www.getty.edu/research/conducting_research/vocabularies/aat/about.html.

[4] Mitchell, Joan S., (Ed.), Dewey Decimal Classification and Relative Index, Edition 21, Forest Press, Albany NY, 1996

[5] Office for Subject Cataloging Policy, Collection Services, Library of Congress, LC Classification Outline, Sixth Edition, Washington, 1990

[6] Quan, D., Bakshi, K., Huynh, D., and Karger, D. R., *User interfaces for supporting multiple categorization.* In INTERACT, 2003.

[7] Petras, V., Larson, R., and Buckland, M.,. *Time Period Directories: A Metadata Infrastructure for placing Events in Temporal and Geographic Context.* In: Opening Information Horizons; 6th ACM/IEEE-CS Joint Conference on Digital Libraries 2006. pp.151-160.

[8] Ranganathan, S. R., *Elements of library classification*, Poona; N. K. Publishing House 1945

[9] Schmitz, P., *Personal Photo Search and Browse Tool.* (Unpublished) final project report for IS 290-2 Search Engines Fall 2005.

[10] Schmitz, P., *Evaluating the Quality of Ontology in Semantic Applications.* (Unpublished) final project report for IS 218 Quality of Information, Fall 2006.

[11] Soergel, D. (1974). *Indexing languages and thesauri: Construction and maintenance.* New York: Wiley.

[12] Soergel, D. (2004). *Human-Computer Interaction Encyclopedia: Information Retrieval.* Berkshire Encyclopedia of Human-Computer Interaction July 2004. Retrieved on 5/1/2007 from http://www.dsoergel.com/NewPublications/HCIEncyclopediaIRShortEForDS.pdf.

[13] Yee, P., Swearingen, K., Li, K., and Hearst, M., *Flamenco Faceted Metadata for Image Search and Browsing*,in proceedings of ACM CHI 2003.

A. Appendix: Project Charter

The charter is available online at the following url:
http://jonlesser.net/wiki/index.php?title=Project_Charter

2. Appendix: Recommendations from Needs Evaluation

- Target user population is the interested/directed general public (eg amateur researcher)
  - need to give this population a name
  - any further need finding has to relate to this population; either more of the public or museum staff that directly contact the public
- Site content
  - Images are a must - concentrate on areas of collections with images.
  - Enable seeing an image from different angles; being able to rotate and see enlargements of the portions of images to show texture are great.
  - A balance of info/image must be achieved for a positive and satisfying experience, return visits, and avoidance of unpleasant user experiences.
  - If we include a DB dump, don't put it on the main object page; link it. While the "data Dumps" are not useful for general public users, it makes sense to provide these metadata either in a separate link, or only for "registered users."
  - Watermarks will be tolerated as long as they do not destroy the visual appeal.
  - Showcase collections created by others. For example: as a good introduction to browsing or viewing the most popular objects or blogs about objects.
- Other types of multimedia objects need to be introduced/explained. While they might be considered useful by some users, most users do not know they can search for other multimedia on the museum's web site.
- Presentation of Content (and search results)
  - If possible, include context beyond just the object. Content that puts object in context is how general public "experiences" an object in an anthropology museum (different than an Art museum).
  - "If you liked this, you might want to look into this" – provide object relationships guidance.
  - Present object's description if an object's image is not found.
  - For objects with minimal information (only a name and a catalog#), provide a message such as "The object is in the museum's collection, but the additional information is not currently available through the online browser. If you have a strong interest in learning more, please contact the staff by filling out this application"
  - Filling the above form should require the users to write a "research justification" - a process that is currently used to weed out the frivolous applicants who want to find more info or see the real objects.

- » Collapse the visual presentation of results with minimal information, or provide them only for “registered users.”

» User Activities

- » Search
  - » Keyword search a must! Many users would come to the website with a specific goal of what they need to find.Most important fields: material, technique, artist, culture, peoples, timeframe context of now.
  - » Minimize search time: the general public is time conscious. it’s not about idly browsing -- it’s about efficiently searching and finding relevant objects that apply to their area of research, and then “bookmarking” them, so they are easily findable, easily shareable...
  - » The users would benefit from “guidance” for searching.
- » Creating sets and sharing them.
  - » Many would want their sets to be private for some of their goals.
  - » Some would want their sets to be shared among a group of specified users if they are working on a group project.
  - » Some would want their sets to be viewed by the general public.
- » Annotations
  - » Include features for authoring annotated collections. Whether they are going to do it depends on whether it will serve their current goals when using the browser.
- » Tagging
  - » 1st: Ability to comment on individual objects.
  - » 2nd, ability to comment on sets
  - » Ability to send URL, with comments
- » Privacy
  - » default user activity to private because users tended to want to comment on things for their own personal use. Most public have expressed more interest in creating content for themselves than for public consumption. Some have explicit privacy concerns, eg for children.
- » Profile
  - » Not interested: the museum browser does not lend to interest in chatting with others about museum objects: at least not for the general public (ie not social networking)
- » Registration
  - » Is acceptable for the users who are really interested in additional features provided by the system.

3. Appendix: Description of target user

Target User: Amateur Researchers / “Seekers”

Target User Qualities:

- Off-and-On Museum Goers
  - When there is an exhibit that has personal appeal, they’ll go and peruse. Might not read all the label text, but will wander around looking for eye catching things.
  - When an item jumps out at them, they’ll want to learn more about it.
  - May even be spurred on to do their own research about the item online.
- Internet Savvy, avid online searchers
  - In conjunction with information, they’re searching for relevant images
- Task Oriented / Time Sensitive
  - They are looking for something in particular. They have a goal in mind. If they do not succeed in finding their item, they will abandon search and look elsewhere.
  - These people described themselves primarily as searchers (not browsers)
- Interest in seeing online pre-made “Collections”
  - Want to see exhibits created by authoritative museum stuff/researchers (NEW)
  - Want to “learn something new” or “to hear a story” created by authoritative museum stuff/researchers (NEW)
  - Want to see the objects that have already been organized along some dimension
  - Want to see the objects probably have a written narrative
  - Want to see the object relationships. Many said that what has meaning to them “is when they can compare and contrast the same object across cultures.”
  - Would be interested in seeing the objects that others have blogged about.
- Low tolerance for “pain”
  - The story around an object that puts it in context was sited often as something user wants to see
  - If content from TMS is not organized consistently, it will be tedious to look through
  - They have little interest in seeing data from database which is “dumped” into the record. For most of our participants, this is a negative, exhausting experience, which would make them go elsewhere

- Images are a must
  - They want to see images that are relevant, and not have to wade through a lot of content that doesn't have an associated image
  - Pictures are THE most important item to have. Several said if there aren't pictures, they'd give up early
- "Authoring"
  - They would most likely leave comments for themselves – free associations, reminders, "field notes," etc.
  - They do not see it as their mission to leave information behind for others
- Public vs. Private
  - They would like to have a privacy option
  - Research is of a very personal nature to them
  - Not ready to be seen until they have completed it (if ever)

4. Appendix: Usability testing materials

Our main persona for whom we designed our user interface is Theresa Conant – a student/ researcher who approaches a museum collection with specific search goals in mind, and who has at least mid-level search skills. We created scenarios to approximate what her possible tasks might be regarding creating, adding to, editing and sharing sets within an online museum browsing experience. We felt we could draw upon the iSchool student resources as sharing very similar characteristics with our main persona, insofar as they are competent online searchers, without having specialized knowledge within the museum domain.

4.1. Participants

From our previous prototype tests (lo-fi and 1st interactive), many iSchool students had already heard about the Delphi Project, and expressed their interest in being part of any future usability tests. We sent out email to a varied group of both First and Second Years. Because of the abundance of volunteers, we were able to select our participants mainly on a first-come, first-serve basis. We had one male and two females, ranging in age from mid twenties to mid thirties. Participants rated their computer competency as "excellent," "competent" or "good." Regarding their museum experience, all three had visited a museum in the last three months, and had visited museum websites mainly to look-up hours, directions, prices and current exhibits. All three cited Flickr, Amazon and Yelp as websites they were familiar with. Two participants had previous experience making sets or lists on a website, and both cited del.icio.us as the place where they had done so. (We ran a total of five pilot tests, but for the purposes of this assignment, we're summarizing the results of the first three...)

4.2. Apparatus

For the pilot usability study, the only equipment we needed was a laptop that had the url for our demo page launched on the Mozilla Firefox browser. We also used iShowU, an application that makes BOTH an audio and visual recording of the participant's voice + mousing behavior, so that it can be watched and listened to after the testing session is over. This was a really great asset, as all three of our team members could watch each video, and verify that we had all seen and heard the same thing. It was also very beneficial to pause the recording, to see exactly what the participant had just done – which is sometimes difficult to ascertain during live events. We conducted each session in private South Hall rooms (107 and 110) which we had reserved in advance for these purposes.

4.3. Procedure

» Each test took approximately 60 minutes.

» The facilitator was the sole spokesperson. Note-takers and observers greeted participant, and then were silent through the rest of the session.

» Consent form(s) and pre-test questionnaire were administered in the beginning by the facilitator.

» The facilitator read the pre-test material that explained the Delphi Museum Browser, and the "think aloud" testing process, etc. (see Task Scenario in Appendix for more details).

» The participant sat in front of a laptop with a browser open to the site.

» The facilitator sat to the left of the participant while note-taker and observer sat in background, with clear view of the screen.

» The facilitator did not provide a demo of the system because we wanted to test whether or not it was self-explanatory.

» The facilitator started the iShowU recording application, and then gave the participant the first task.

» Participants were encouraged to talk aloud regarding their thought process, as well to comment about their questions/preferences etc. while executing tasks.

» Once the final task was completed, the facilitator asked the participant to fill-out our 11 question post-test survey.

» Most participants wanted to stay and give more feedback / suggestions for improving the site.

4.4. Test Measures

As already stated, we looked for a 1) participant's ability to and 2) participant's method for carrying out each task. Would the tools we provided be easily observed and utilized? What scenic route did they take through the browsing experience? And did participants express satisfaction with the choices we gave them, and on the pages where we gave them certain abilities?

4.5. Timing

We weren't really concerned about how long it took each participant to complete a task. In fact, it was instructive to us to simply "watch them go." Participants seemed to enjoy 'playing around' with certain functions, and we did not discourage them from doing so. Also, during the 'talk aloud' process, some participants had much to tell us, while others were more reticent. This difference in talkativeness obviously influenced the time it took to

complete a task. It was therefore necessary to go back and watch the recordings to get more accurate timing results, subtracting out the talk time.

4.6. Errors

Although we attempted to log errors, what we discovered from 1) observing the pilot-testers in action and 2) watching the recordings afterward was that no one actually committed any "errors" per se. (Perhaps we should have tried to agree upon what constituted an 'error' in advance.) In the end, we felt each participant was able to achieve their desired goal, and pretty easily at that. They simply had different ways and thought-processes of going about it. If there were any "errors," it was from a few tiny bugs left in our program, that gave our pilot-testers an unexpected result (that we had to explain).

4.7. Results for quantitative measures

Tasks	001	002	003
1: Browsing thru collection looking for object that matches 2 attributes	2 mins	5 mins	7 mins
2.1: Click on item to see more details	3 secs	10 secs	7 secs
2.2: Add an item to default set	35 secs	40 secs	30 secs
3.1: Search for Baskets	2 mins	5 mins	5 mins
3.2: Save a result to a new set	5 secs	10 secs	6 secs
4.1: Find your sets	1 sec	1 sec	1 secs
4.2: Change name and description of one of your sets	7 mins	9 mins	6 mins
5: Share a link of one of your sets to a friend	15 secs	17 secs	25 secs

5. Appendix: Interview Materials

5.1. Interview protocol

5.1.1. Before session

- make sure to have working recording devices (audio, camera, video), batteries/mics as needed
- in the pair, decide who will interview and who will take notes
- be early to the session, to have time to set up and clear up any issues

5.1.2. Interview

- Introduction and orientation
  - introduce yourself, say you're from our final project group
  - briefly describe the purpose of the study and the procedure for the session
  - make sure to get informed, voluntary consent
    - ensure that participant is comfortable with the study and ask if they are OK with being recorded (ask for each type of recording, eg audio/video/image)
    - clear up confidentiality issues; assure them that we will keep their responses in strict confidence and that we are using their responses only for the project; assure them that their identities will be anonymized in any presentation of the study
    - make sure they understand that they can refuse to answer any question; also, they can stop the recording or stop the session at any time
  - tell them there is no right or wrong answer; we really want to hear what they have to say
  - ask for and answer any questions they may have
- Preamble
  - these are standard easy questions; demographics, background
  - for example
    - What's your job at the Hearst Museum?
    - How long have you been working at the Hearst Museum?
  - Main [see representative questions]
- Closing
  - Would you like to add anything? Is there anything I haven't asked you?

- Ask notetaker if they have anything else they want to ask about
- Can we contact you again if we have further questions? What would be the best way?
- Is there anyone you know that we should talk to?
- Thank you!

5.1.3. After session

- transfer recordings onto a computer, upload to FTP
- within 24 hours
  - notetaker writes up notes, using recording (if it exists) to enrich them
  - interviewer reviews and adds to the notes

5.2. Representative interview questions

5.2.1. General task analysis questions

Based on the slides from Prof. Marti Hearst I-213 class.

- Working with the current system (which we are NOT trying to replace)
  - What kinds of people (clients, users) are you working with?
  - Why do they come to you and what tasks do they want you to perform or assist in performing?
  - Where are the tasks performed? How often? Are there any constraints on the tasks, such as time/location?
- What is the relationship between the user and the data?
  - How do they come about the information which they use to search for the objects in your collection? Does learning these kinds of initial information require special education (experience, connections)?
  - Do your clients interact with each other? How do they do it?
  - What do they do with the data, once they completed their search?
  - What happens when things go wrong?
  - Try to identify people with different needs and preferences, with respect to their attitudes about using the current systems: Is there a way to describe general categories of users? If yes, what makes one group different from another? What makes the users in one category similar to each other?
- Demonstration
  - Ask for a demo of how they handle the necessary tasks using the current TMS

(The Museum System).

- Looking into the future system
  - What would you/your users would like to have been different about your work/tasks/system?
  - Ask what, if anything, must be in the system in order for them to be interested in using it together with (or over) the current system.
  - Try to identify potential future users of the future system, their needs, preferences, and attitudes. Request contact information if possible.

These are more for the analysis stage, but are useful for figuring out what kinds of things we should ask. These are from Prof. Maneesh Agrawala's CS 160 slides.

1. Who is going to use the system?
2. What tasks do they now perform?
3. What tasks are desired?
4. How are the tasks performed?
5. Where are the tasks performed?
6. What's the relationship between the user and the data?
7. What other tools do the users have?
8. How do users communicate with each other?
9. How often are the tasks performed?
10. What are the time constraints on the tasks?
11. What happens when things go wrong?

5.2.2. Collections managers

- How often do you use TMS in a typical week?
- Tell me about the last time you used TMS.
  - What did you need to accomplish?
  - Is that a common task?
- What is the most common reason you use TMS?
- What are the most useful fields when describinig an object?
- When you go about searching for an object, what fields do you search on?
- How much time everyday are you involved performing searches? (are you able to complete your regularly assigned work as well as the searches?)
- How many searches do you do in a given day?
- Do you interact with data in the collection besides perfoming searches? (Data input,

corrections, etc.)

» What other responsibilities belong to a Collections Manager?

» How does information get into the system?

» How accurate are your searches in locating objects? Do you usually succeed on the first try? Does it take several refinements?

» What do you do if nothing comes up in the search, yet you know it should?

» Could you show us a demonstration of a recent search you performed?

» Do you ever receive additional information about an object from a researcher that you'd like to add to your system? Are you able to? How?

» Do you ever want to put your users in touch with other users who seem to be involved in the same type of research? Are you able to do this? How?

» About workflow

  » Check workflow diagram - where do you fit in this diagram?

  » What is your role?

  » What are you trying to acheive through this system?

  » With whom do you communicate?

5.2.3. Public Relations Staff

» Goals of the PR Staff interview

  » identify the parts of the collection which draw the interest from the general public.

  » analyse the needs, practices, and preferences of the different groups of general public.

» Questions Museum PR Person

  » What kind of access do normal everyday people have to the collections?

  » What are your goals for public interaction with the collection?

  » What kinds of outreach programs are they running now.

  » Job duties. Who comes to visit. How you learn about the visit, what arrangements you make.

  » What kinds of information do you provide/request from(a)visitors and (b) the museum stuff?

  » What interest groups are there amongst museum visitors?

  » How often do the general public visitors have special education in arch./ethn./hist. that relates to their interest?

» Can you put us in touch with some of these special interest/clearly defined general

public representatives/groups?

5.2.4. General Public

- Do you have a special interest in:
  - Anthropology, archeology, art studies?
  - other (please specify)
- What kinds of activities do you do in these areas of interest? How much time do you spend on these activities?
- What kinds of online communities/groups/mailing lists do you participate in?
- If yes, are any of these groups relate to your interest for anthropology/archeology?
  - Also, if yes, then ask what groups, why they participate, how interact with others, what are the likes/dislikes.
  - If no, then ask why not?
  - Also, if no, then what would need to be changed/added so that he/she would like to interact with others archeologists online?
- Depending on the answer to the question above: why yes or why not? If yes, what would like to do?
- Do you use any online search systems/browsers to find information for your area of interest?
- If you had access to the browser of the museum's online collection, what would you like to be able to do?
  - what kinds of information would want to find by using such a browser?
  - how much is interested in different types of information media (text, images, recordings, etc)?
  - would you like to create your own collections?
  - how would you feel about the need to log in or the requirement to log as a way to get additional permissions?
  - would you be interested in creating a profile?
  - would you be interested in seeing the profiles of other people?

6. Appendix: Competitive analysis notes

» Arago: People, Postage, and the Post. Strengths of this site include: a beautiful visual design, zoom/pan tools for viewing high-resolution images allow close inspection of objects, it supports personal collections of objects with user accounts, and it has advanced search that includes location and date range visualizations. The site maintains browser back button functionality with a flash implementation. Virtual exhibits combine narrative with objects in the collection. Items in the exhibitions can be explored with all the same tools as objects outside the exhibits. Identified weaknesses were: breadcrumbs are implemented in a confusing, non-standard way and the Flash implementation breaks some standard browser affordances such as right-click contextual menus and scrolling with a scroll wheel. Also, it can be difficult to understand the full breadth of the collection. Objects are not well connected with other objects in the collection.

» New York Public Library Digital Gallery. Strengths of this site include: it provides a nice entrance to the collection with many jumping in points ("Today's Gallery Pick", "Curators' Choice", etc.), it tracks search history without requiring a user to login, allows users to mark an object as "selected" and then see a list of all selected objects, and has lots of browsing features that engage the user and enable exploration. Identified weaknesses were: large images are not very large, integration with the third-party printing fulfillment service is awkward, and the series of icons along the top are non-functional and apparently part of the logo.

» Victoria and Albert Museum. Strengths of this site include: it is visually very appealing, has many different types of collections (furniture, glass, architecture) just like the challenge we're undertaking, and all collections share the same search frameworks: Object, Date, Technique, Artist/designer, Place, Dimensions, Museum number. As you go from collection to collection (ceramic, sculpture, childhood,) everything stays the same, except for the color, so there is good visual consistency. Also, because of the color, you know 'where you are' at any given moment. Identified weaknesses were: searches must use exact text - for example, "doll houses" vs "doll's house" vs "dolls house", without the exact term "dolls' house", did not bring up any results; searches must specify the exact fields - for example "Ceramic toys" brought up zero hits, and yet there are plenty of them; "Buy image" brings up an email editor, instead of a screen that tells you ordering information, etc. If you get taken to another museum, or other collections, there's no links to get back to Search Collections V&A museum.

» Thinker ImageBase. Strengths of this site include: It contains tens of thousands of images indexed by descriptive keywords; people can create their own virtual galleries from image databases and each work of art can be re-ordered or deleted as required. The text field next to each work allows you to input additional commentary, galleries can be either shared with the public or kept private, and all the images in the ImageBase are zoomable. Identified weaknesses were: keyword searching doesn't support

controlled vocabulary: if “gardening” doesn’t return much, you must try “gardens,” or “garden,” or “gardener,” or “gard@.” The ImageBase does not contain a wealth of art-historical information, terminology and/or analysis. The flow of using the website is very confusing - for example, when you find an image through the search textbox, or using advanced search, you cannot directly add images to a gallery. Instead, you have to go into the My Gallery >> Search, and then type keyword, and then add images. There is no efficient way of browsing other’s galleries except for the fixed category and long list .

» eMuseum from Gallery Systems. Gallery Systems is a commercial CMS vendor, selling both Embark and TMS. Their web portal module is called eMuseum, and has been used in some nicely designed, if not particularly useful sites for searching museum collections. The eMuseum support provides keyword search and media browse capability (i.e,. viewing of thumbnails for results and slightly larger images for individual items). There is no support for personal galleries or sets. There is no support for facetted browsing, although they do allow allow search within specific portions of the metadata (e.g., artist name). The keyword search appears to have been done poorly, with no relevance ranking (e.g., partial string matches may come up first) and no apparent ontology or vocabulary support. Two examples of nice collections nicely presented are:

    » Brooklyn Children’s Museum (http://www.brooklynkids.org/emuseum/code/emuseum.asp) Site has a very nice visual design. Search has facets of a form, but cannot be used in a facetted browse mode to refine queries. Hard to understand how facets interact. Search has functionality useful to collections managers on main public page (e.g., search by object number), which seems to indicate that they chose a researcher as the key persona for their audience, or that they skipped this sort of analysis altogether. Keyword search lacks good relevance ranking (Top result for “mask: is a painting of a mask, 2nd to top result is an adze, used among other things to carve masks).

    » Giza Archives (advanced search for photos). This site is really useful only for archivists and professional curators, as the search model is based upon METS categories; even enthusiastic web visitors will likely be very confused by this site. They did an interesting Flash application that allows users to find objects by location at the site, but unless one already understands a good deal about the Giza site, this is also a bit confusing.

7. Appendix: Unused scenarios

7.2.1. Persona 1: Alan Prewitt, Professor

Alan Prewitt has been working on a book, "The Last Days of the Sioux Nation" for over a year now. The book will have about three dozen color images by Alan's estimation. The book covers a lot of different aspects of the Sioux people and Alan has been diligently collecting images for potential use for some time now. A colleague told Alan about the Hearst Museum's new collection browser called Delphi. Alan has a rough idea of what is in the Hearst collection and wants to look deeper to find out if there are any objects he can use for his book.

Back in his office, Alan fires up his computer and navigates to the Hearst Museum's website where he finds a link to Delphi. Alan is not sure exactly what the Museum has related to Sioux people, so he simply types "Sioux" into the search box on the front page of Delphi. The systems brings back a few thousand results and Alan can clearly see the number of results and how many pages it will take to display them all. Alan scans the thumbnails next to each search result on the first page of results but nothing catches his eye. At the top of the Results page, Alan can see a breakdown of results by category along with the number of results in each category.

Looking to the categories of results Alan sees that 59 of his search results fall into the category "weapons". Alan only has one image lined up for the chapter on weapons of the Sioux people, so he clicks on the "weapons" link to see if the Museum has anything to flush out the chapter. Of the weapons results, only about a third have thumbnails. Alan clicks on a result titled, "Dakota long bow". The details of the bow, along with a large photo, load quickly in Alan's browser. It's interesting, but not right for Alan's book and he hits the back button on his browser. Alan spends the next 15 minutes exploring the results for Sioux. He is able to navigate mostly by clicking on taxonomy terms and rarely finds it necessary to type a new query into the ever-present search box.

Alan finds a few objects he thinks he would like photographs of for his book, but feels he needs to see them in person first. Alan contacts the Museum by email to make the necessary arrangements. The Museum responds by asking Alan to organize the objects he would like to see into a set on Delphi and provides him with some basic instructions for doing so. Alan logs back onto Delphi and makes an account as instructed by the museum staff.

Once Alan logs in with his new account he finds himself presented with a number of options, one of which is creating a new set. Clicking that option, Alan is prompted to name the set and instructed on how to add items to the set. He then searches for the objects he would like to see at the Museum. Alan uses the search box with very specific queries that include details he remembers about the objects of interest. He also notices some of the items he's interested in are in a "recently viewed" list on the side of the screen. Each object detail

prominently features an “Add to Set” button.

With all the objects Alan is interested in seeing contained in a single set, he clicks on the set’s detail page. Remembering the instructions sent to him by the Museum staff earlier, Alan locates the “Make an appointment to see these objects” button on the set detail page. Alan clicks on this button and fills out a simple form. Soon after, a Museum staff member confirms the appointment by email. The next week Alan arrives at the Museum and is welcomed by the staff who show him to a room where all the objects he requested are out and waiting for inspection.

7.2.2. Persona 2: Joyce Reisner, Collections Manager

It’s peaceful down here in the climate-controlled environment and a little bit unworldly, surrounded by aisles brimming with artifacts of bygone eras and faraway lands. Joyce is in the basement of the Museum’s storage facility on San Pablo Ave., far away from the Hearst Museum on campus. That means she won’t to be able to personally consult the card catalogs or other paper records which she used to rely upon for object and location information. But she’s not worried.

Today she’s fulfilling a researcher’s proposal to photograph festival clothing and ritual materials of Korea. In the past, a request like this easily would have taken the better part of a week, mainly just to cull through records that might be relevant, pull items from the shelves, and then make a notation that they were viewed. But now scholars can research the collection with Delphi on their own time, thank goodness! The researchers have come all the way from New York, so time matters to them as well. They already know exactly what they wish to examine, and have brought along a print-out of the specific object numbers, complete descriptions and thumbnail photos – all that’s missing is the location number, which Joyce can easily look-up.

As she gathers the items from the shelves, she notices a few vibrant textiles that are not on the researcher’s list. The researchers are blown away when they see them! They excitedly tell her that these garments were worn by one extended family group in the northern region of Korea -- known for their hand-dyed fabrics in maroon and saffron colors which represented their clan. The garment’s provenience had been speculated upon when they arrived at the Museum, but until now, no one really knew their exact age or origin for certain. These serendipitous discoveries are why Joyce loves her job! The puzzle pieces keep falling into place, and one day the complex tapestry of the past will be revealed.

She encourages the researchers to add their theory about these textiles on Delphi, uploading their comments to the artifact’s records. Everyone hopes this will stimulate a welcome discussion in the academic community.

Joyce decides that these colorful items that the researchers selected today would make great visuals for the Museum’s website. She searches for these specific items online with the

Delphi, then highlights them as one of her "favorites," so that others near and far can also enjoy them, just as she does.

7.2.3. Persona 3: Sally Grant, Museum Education and Public Relations

Sally starts her day with checking her email at the office. This morning, she sees that an elementary school teacher is asking for her help setting up a field trip to the Museum. This time, the theme of the 5th grader's field trip is "Native American ceramics."

Due to the newly developed Delphi tool, teachers can now search for related items on any theme, and come up with a set of objects that they'd like to see. Sally sometimes gives them back additional materials that she thinks are relevant. A virtual gallery can be created based on Sally and the teacher's collaboration about a specific theme. It's a valuable resource for other teachers who might be planning a field trip to the Hearst Museum later. Sally also arranges a docent who can guide field trip tours during their visit, and who often fetches relevant objects that are not in the exhibit. There are so many wonderful objects in the collection that have never been exhibited! She also consults with collections managers to get their assistance finding objects that may not get viewed as often, but which have just as much value.

The exhibit gallery is small, and only changes once a year. So along with creating a PR webpage of the current exhibition, Sally also creates virtual galleries highlighting specific artifacts and special collections like "Native American Baskets" or "Indian textiles with mirrors," so that a wide range of items will be seen. Sally wants the public to know that what's displayed in the exhibit gallery is just the tip of the iceberg compared to what the Museum preserves in its vast collection. She likes to get input from collection managers, volunteer docents and even teachers to help her choose each monthly theme.

After the field trip, students and teachers usually leave comments, share ideas and discuss the objects. Students become very interested in the Hearst Museum collections and create their own galleries on the Museum's website, and email their favorite images to friends. Sally's walls are adorned with beautiful arts and crafts "Thank You" note projects from previous field trips, that have been inspired by their visit to the Museum. It makes Sally really happy that the Museum has enriched their lives, and makes her hopeful that they'll be passionate visitors from now on.

7.2.4. Persona 4: Theresa Conant, Academy of Art Student

The semester is midway through, and preliminary fashion sketches and samples of proposed materials must be turned in tomorrow for a project midterm grade. As a first year student at the prestigious Academy of Art, Theresa has a lot to prove. The competition is intense, and if she wants to make a name for herself, she has to come up with something really memorable, something that has meanings on several levels. Theresa has been looking for a concept that could even become her "signature."

She leafs through her sketchpad again for some inspiration. The Native American patterns she sketched from a few weeks back are simple, yet very strong. This could work, but she needs to see a few more designs and expand upon them in her own style.

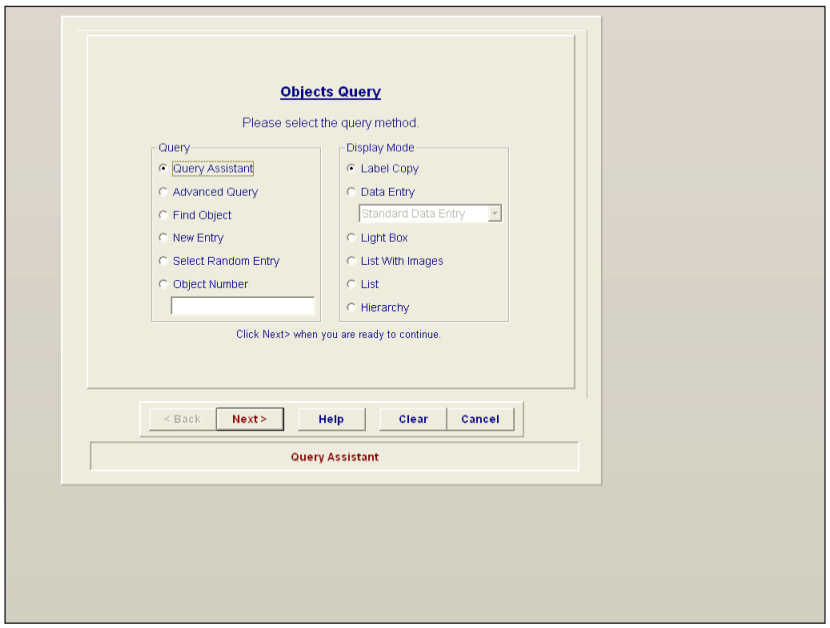
Unfortunately, she's supposed to fill-in tonight for a sick co-worker at her job, and it's already 2pm in the afternoon. Theresa remembers that the Hearst Museum over in Berkeley has one of the largest Native American collections in the country. Well, between work and rush hour traffic, she doesn't have time to visit the exhibit now! She goes online to the Museum's website and goes to the Delphi home page . She quickly browses through their North American collection, first looking at Native American textiles, then moving onto Native American baskets. She zooms in to get a closer look at the details. She takes particular note of the materials used, and the unique stitches and weaving that she may be able to incorporate in her own designs. Every time she sees something she likes, she adds it to her set. It hasn't taken her long at all, and she's already got more than 50 images in her set. Confident that she's got ample to work with, she emails the set to herself.

Thanks to Delphi, and all the time and effort it's saved her, Theresa is able to finish the majority of her project before leaving for work, and even gets some sleep after an exhausting day.

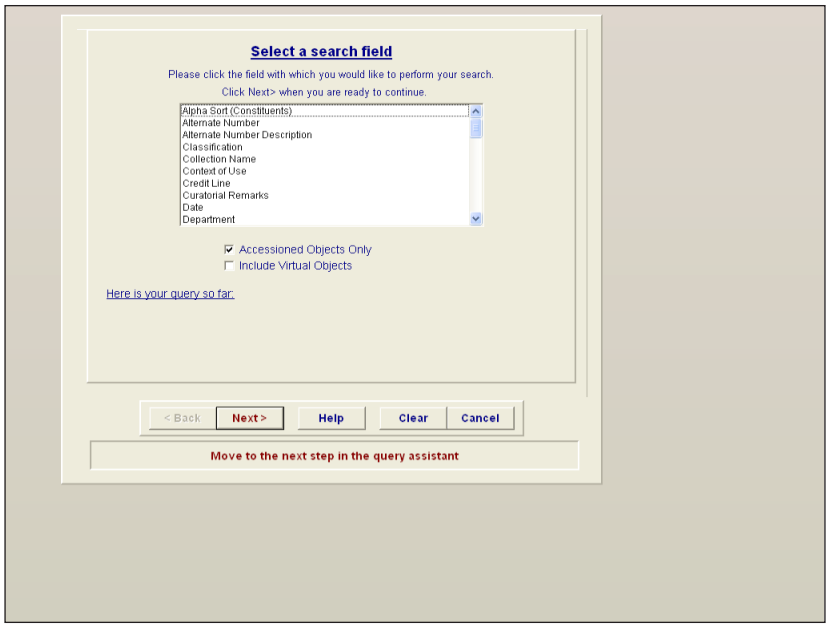
8. Appendix: TMS research

8.1. Screenshots

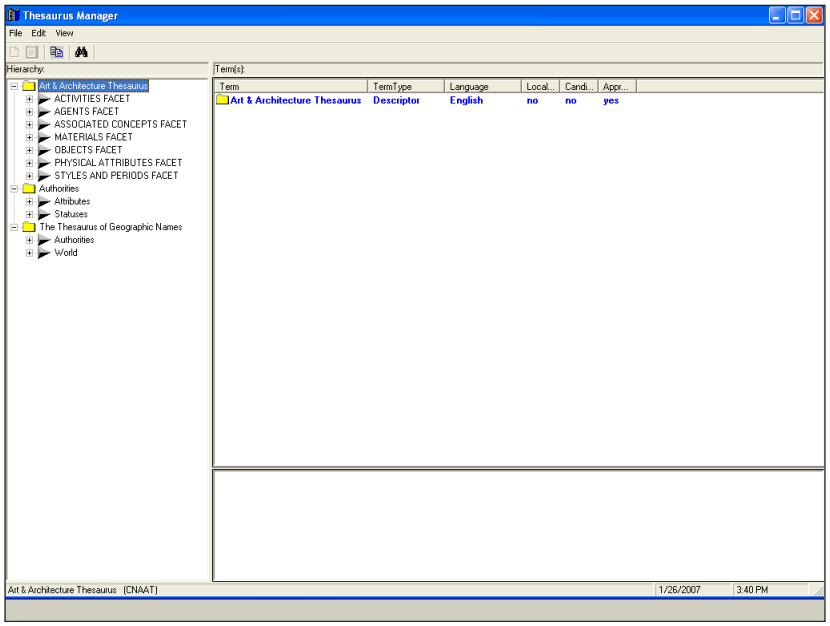
The followng screen shots are from The Museum System (TMS) by GallerySystems. The Hearst Museum uses TMS to manage its collection.



TMS Screenshot 1

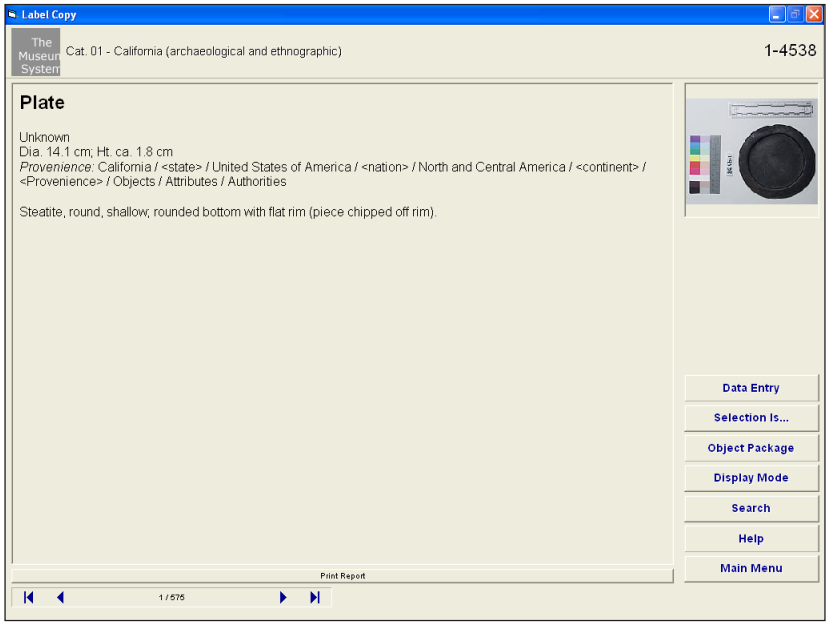


TMS Screenshot 2



TMS Screenshot 3

TMS Screenshot 4



TMS Screenshot 5

TMS Screenshot 6

8.2. Commonly-used search fields in TMS (expert perspective)

» Catalogue number
» Alternate number
» Object name
» Provenience
» Site
» Object related constituents
» Acquisition related constituents
» Date collected
» Date accessioned
» Accession number
» Culture

9. Appendix: Recommended websites and resources

SFMOMA
http://sfmoma.org/msoma/index.html
Podcasts: http://sfmoma.org/education/edu_podcasts.html

Metropolitan Museum of Art
http://www.metmuseum.org/Works_of_Art/collection.asp?HomePageLink=permanentcollection_l
https://www.metmuseum.org/mymetmuseum/index.asp

The Museum of Modern Art, NY
http://www.moma.org/collection/browse_results.php?criteria=O%3ADE%3AI%3A5&page_number=1&template_id=6&sort_order=1

Victoria and Albert Museum
http://images.vam.ac.uk/ixbin/hixclient.exe?_IXSESSION_=&submit-button=search&search-form=main/index.html http://www.vam.ac.uk/collections/index.html

Legion of Honor
http://thinker.org/gallery/index.asp
http://search2.famsf.org:8080/mygallery/view.shtml?record=306037

BAM/PFA
http://www.bampfa.berkeley.edu/index.html
http://www.bampfa.berkeley.edu/exhibits/index.html

Bancroft Library Foundations of Anthropology site
http://bancroft.berkeley.edu/Exhibits/anthro/index.html
http://bancroft.berkeley.edu/Exhibits/anthro/10videos1.html

Jonathan Goodrich's Museum Studies MA thesis
http://library.jfku.edu/search/ffinal+project+museum/ffinal+project+museum/61%2C225%2C225%2CE/frameset&FF=ffinal+project+museum+goodrich&1%2C1%2C

Lisa Granger's Museum Studies MA thesis
http://library.jfku.edu/search/ffinal+project+museum/ffinal+project+museum/61%2C225%2C225%2CE/frameset&FF=ffinal+project+museum+granger&1%2C1%2C

Minneapolis Institute of Arts
http://artsmia.org/art-of-asia/buddhism/
http://artsmia.org/viewer/detail.php?id=72862&i=27&v=2&dept=1&cc=China&class=ceramic t

10. Appendix: Ontology

The complete ontology we generated can be accessed online at:
http://jonlesser.net/wiki/index.php?title=Ontology_Dump

11. Appendix: Dump Column Configuration File

```
<?xml version="1.0" encoding="UTF-8"?>
<dumpColConfig>
  <colSep value="|"/>
  <colInfo name="ObjectID">
    <comment>The ObjectID column is essential for metadata association,
             but it is useless for mining.</comment>
  </colInfo>
  <colInfo name="ObjectNumber">
    <comment>Not clear how the ObjectNumber is used, but it is useless for mining.</comment>
  </colInfo>
  <colInfo name="ObjectName">
    <comment>The ObjectName column has much object name info, some materials, some colors,
        some activity ("cooking stone"), "spinning top". "flaking hammer".
        Note "crustacean shell " as exclusion and other cat for "shell"
        Vast majority of tokens are strongly indicative of a useful facet -
        consider all non-matches carefully.
        Some instances of "decorated" - does this belong in colors? Where?
          decorated, carved, engraved, painted, worked,
          lacquered, inlaid, incised, writing, "in relief", etc..
        Some instances of '?' - need to remove for matching, but use to reduce
        relevancy. Perhaps should treat most of the noise tokens the same way...
        Some embedded newlines - need to convert to spaces. Especially for coin names???
    </comment>
    <tokenSep value="[;,:]"/>
    <tokenSep value="w/"/>
    <tokenSep value="with "/>
    <!-- Note space after to preclude "without" match -->
    <noiseToken value="fragments"/>
    <noiseToken value="fragment"/>
    <noiseToken value="fragements"/>
    <noiseToken value="fragement"/>
    <noiseToken value="frgmnt"/>
    <noiseToken value="frags."/>
    <noiseToken value="frags"/>
    <noiseToken value="frag."/>
    <noiseToken value="frag"/>
    <noiseToken value="chips"/>
    <noiseToken value="sherds"/>
    <reduce from="[\n]" to="#"/>
    <!-- Reduce all numbers to '#' Fix from value to be regex for number -->
    <reduce from="[\s]" to=" "/>
    <reduce from="arabesque:" to="statue of "/>
    <!--  Reduce all white space to single space - especially for newlines -->
    <facetInfo name="UseorContext" relevancy="0.8"/>
    <facetInfo name="Material" relevancy="1"/>
    <facetInfo name="CulturalGroup" relevancy="1"/>
    <facetInfo name="Color" relevancy="1"/>
    <facetInfo name="TechniqueDesignorDecoration" relevancy="1"/>
    <facetInfo name="SiteorProvenience" relevancy="0.8"/>
    <!-- Few, but seem to be okay -->
  </colInfo>
  <colInfo name="Dated">
    <comment>The Dated column has interesting date info for a timeline,
        but it is useless for mining.
        To use it for dates, a custom parser will have to be built that accepts
        a very wide variety of syntax, both for modern dates as well as for ancient
        dates (e.g., "81 b.c.", "107 ad")
    </comment>
  </colInfo>
  <colInfo name="Title">
    <comment>The Title column has a few object name tokens and some location info.
```

```
        There are few entries.
        This should probably be lower reliability than other columns, but the data
        there seems useful.
    </comment>
    <tokenSep value="[;,/\[\]\(\)]"/>
    <!-- Need to figure out how to elide [] brackets and parens-->
    <facetInfo name="UseorContext" relevancy="0.8"/>
    <facetInfo name="Material" relevancy="0.7"/>
    <facetInfo name="CulturalGroup" relevancy="1"/>
    <facetInfo name="Color" relevancy="0.7"/>
    <facetInfo name="TechniqueDesignorDecoration" relevancy="1"/>
    <facetInfo name="SiteorProvenience" relevancy="0.8"/>
    <!-- only a few, but might as well use them -->
  </colInfo>
  <colInfo name="Medium">
    <comment>The Medium column has rich materials info and object name info, some colors,
        Note many compounds with a colon (metal:copper, ceramic:terracotta). Not clear
        if we should preserve  with synset matching, or split on colon and then just
        filter the hypernyms. Note e.g., "ledger: stone" argues for splitting.
        What about "technique: painting"? Counter example: "plant material: cane".
        Pro example: "ceramic: b" (?). Also "paper: gilt", and "shell: mother of pearl"
        Either way?: "wood: stick".
        Vast majority of tokens are strongly indicative of a useful facet -
        consider all non-matches carefully.
        See comments for ObjectName as well.
        Syn: for porcelain: porcelin.
        Sub-type: for red: corsal-red, cinnabar-red
        Sub-type: for blue: green-blue, greenish-blue
        Sub-type: for black: reddish-black, brownish-black, greyish-black, greenish-black
        Sub-type: for white: off-white, gray-white
        Sub-type: for brown: red-brown (syn reddish-brown), iron-brown, blue-brown
        Colors: blue-green, grey-green.
        Excl for yellow: "yellow satinwood tree"
        Excl for grey: "grey parrot"
    </comment>
    <!-- Need to figure out how to elide [] brackets and parens-->
    <tokenSep value=" - "/>
    <tokenSep value="\(a\)"/>
    <tokenSep value="\(b\)"/>
    <tokenSep value="\(c\)"/>
    <tokenSep value="[;,/\[\]\(\)]"/>
    <noiseToken value="\(?\)"/>
    <noiseToken value="?"/>
    <noiseToken value="."/>
    <!--  Reduce all white space to single space - especially for newlines -->
    <reduce from="[\s]" to=" "/>
    <noiseToken value="fragments"/>
    <noiseToken value="fragment"/>
    <noiseToken value="fragements"/>
    <noiseToken value="fragement"/>
    <noiseToken value="frgmnt"/>
    <noiseToken value="frags."/>
    <noiseToken value="frags"/>
    <noiseToken value="frag."/>
    <noiseToken value="frag"/>
    <noiseToken value="sherds"/>
    <!--  Reduce all white space to single space - especially for newlines -->
    <reduce from="[\s]" to=" "/>
    <!--  Remove trailing periods -->
    <reduce from="[\.]$" to=""/>
    <facetInfo name="UseorContext" relevancy="0.7"/>
    <facetInfo name="Material" relevancy="1"/>
    <facetInfo name="CulturalGroup" relevancy="0.5"/>
    <facetInfo name="Color" relevancy="1"/>
    <facetInfo name="TechniqueDesignorDecoration" relevancy="1"/>
```

```
  <facetInfo name="SiteorProvenience" relevancy="0.8"/>
</colInfo>
<colInfo name="Dimensions">
  <comment>The Dimensions column may be useful for some metadata, but it is
    useless for mining.</comment>
</colInfo>
<colInfo name="Markings">
  <comment>The Markings column has few useful entries. There is some utility
       (used for, for), info and object name info,
       Lots of garbage chars. Elide all numbers (lots of nonsense).
  </comment>
  <!-- Need to figure out how to elide [] brackets and parens-->
  <tokenSep value="[;,/\[\]\(\)]"/>
  <!-- & is a separator -->
  <tokenSep value="&amp;"/>
  <!--  Reduce all white space to single space - especially for newlines -->
  <reduce from="[\s]" to=" "/>
  <!--  Remove trailing periods -->
  <reduce from="[\.]$" to=""/>
  <noiseToken value="fragments"/>
  <noiseToken value="fragment"/>
  <noiseToken value="fragements"/>
  <noiseToken value="fragement"/>
  <noiseToken value="frgmnt"/>
  <noiseToken value="frags."/>
  <noiseToken value="frags"/>
  <noiseToken value="frag."/>
  <noiseToken value="frag"/>
  <noiseToken value="sherds"/>
  <!-- not a complete token, so elide notation tokens -->
  <reduce value="in black ink" to=""/>
  <facetInfo name="UseorContext" relevancy="0.8"/>
  <facetInfo name="Material" relevancy="1"/>
  <facetInfo name="Color" relevancy="1"/>
</colInfo>
<colInfo name="CuratorialRemarks">
  <comment>The CuratorialRemarks column has some useful entries, and lots of oddities.
  Need to understand "cataloguer category:" tokens - should we believe these or toss them?
  Also the "previously catalogued as/assigned to" entries
  There are a few location items, and a few object names.
  There is some culture info
  Lots of garbage chars. Elide all numbers (lots of nonsense).
  In general this does not look like a column we should expose to the public.
  There are a few very interesting notes, but a great deal of administrative
  details that should nto be exposed.
  </comment>
  <tokenSep value="[;,]"/>
  <!--  Reduce all white space to single space - especially for newlines -->
  <reduce from="[\s]" to=" "/>
  <!--  Remove trailing periods -->
  <reduce from="[\.]$" to=""/>
  <noiseToken value="fragments"/>
  <noiseToken value="fragment"/>
  <noiseToken value="fragements"/>
  <noiseToken value="fragement"/>
  <noiseToken value="frgmnt"/>
  <noiseToken value="frags."/>
  <noiseToken value="frags"/>
  <noiseToken value="frag."/>
  <noiseToken value="frag"/>
  <noiseToken value="sherds"/>
  <facetInfo name="UseorContext" relevancy="0.8"/>
  <facetInfo name="Material" relevancy="1"/>
  <facetInfo name="CulturalGroup" relevancy="1"/>
  <facetInfo name="Color" relevancy="1"/>
```

```
  <facetInfo name="TechniqueDesignorDecoration" relevancy="1"/>
  <facetInfo name="SiteorProvenience" relevancy="0.8"/>
</colInfo>
<colInfo name="Description">
  <comment>The Description column has much object name info, much materials, colors,
   some activity ("gambling song"), "abrading stone".
   Lots of "nagpra description" instances. Prefix on some other comment - elide?
  </comment>
  <!-- Need to consider expanding these more widely -->
  <genExclusion value="lacking"/>
  <genExclusion value="missing"/>
  <!-- Need to figure out how to elide [] brackets and parens-->
  <!-- & is a separator -->
  <tokenSep value="&amp;"/>
  <tokenSep value="\(a\)"/>
  <tokenSep value="\(b\)"/>
  <tokenSep value="\(c\)"/>
  <tokenSep value="[;,/\[\]\(\)]"/>
  <!--  Reduce all white space to single space - especially for newlines -->
  <reduce from="[\s]" to=" "/>
  <!--  Remove trailing periods -->
  <reduce from="[\.]$" to=""/>
  <noiseToken value="fragments"/>
  <noiseToken value="fragment"/>
  <noiseToken value="fragements"/>
  <noiseToken value="fragement"/>
  <noiseToken value="frgmnt"/>
  <noiseToken value="frags."/>
  <noiseToken value="frags"/>
  <noiseToken value="frag."/>
  <noiseToken value="frag"/>
  <noiseToken value="sherds"/>
  <facetInfo name="UseorContext" relevancy="0.8"/>
  <facetInfo name="Material" relevancy="1"/>
  <facetInfo name="CulturalGroup" relevancy="1"/>
  <facetInfo name="Color" relevancy="1"/>
  <facetInfo name="TechniqueDesignorDecoration" relevancy="1"/>
  <facetInfo name="SiteorProvenience" relevancy="0.8"/>
  <!-- Few, but seem to be okay -->
</colInfo>
<colInfo name="Provenance">
  <comment>The Provenance column has mostly location info.</comment>
  <tokenSep value="[;,\(\)]"/>
  <reduce from="site [a-z]" to=""/>
  <facetInfo name="SiteorProvenience" relevancy="1"/>
</colInfo>
<colInfo name="Notes">
  <tokenSep value="[;,:-]"/>
  <noiseToken value="each"/>
  <reduce from="[\n+]" to=""/>
  <facetInfo name="Material" relevancy="0.7"/>
</colInfo>
<colInfo name="Edition">
  <tokenSep value="[;,]"/>
  <reduce from="[\.\n]" to=""/>
  <facetInfo name="UseorContext" relevancy="0.8"/>
  <facetInfo name="Material" relevancy="1"/>
</colInfo>
<colInfo name="Remarks">
  <tokenSep value="[;,]"/>
  <facetInfo name="CulturalGroup" relevancy="0.6"/>
  <facetInfo name="SiteorProvenience" relevancy="0.8"/>
</colInfo>
<colInfo name="SiteName">
  <comment>The SiteName column has mostly location info.</comment>
```

```
  <tokenSep value="[;,]"/>
  <facetInfo name="SiteorProvenience" relevancy="1"/>
</colInfo>
<colInfo name="SiteNumber">
  <comment>The SiteNumber column has nothing useful.</comment>
</colInfo>
<colInfo name="ObjNm_ObjName">
  <tokenSep value="[;,\(\)]"/>
  <tokenSep value=" - "/>
  <tokenSep from="w/"/>
  <!--  Remove trailing periods -->
  <reduce from="[\.]$" to=""/>
  <reduce from="[/]" to=""/>
  <reduce from="acc db:" to=""/>
  <noiseToken value="fragments"/>
  <noiseToken value="fragment"/>
  <noiseToken value="fragements"/>
  <noiseToken value="fragement"/>
  <noiseToken value="frgmnt"/>
  <noiseToken value="frags."/>
  <noiseToken value="frags"/>
  <noiseToken value="frag."/>
  <noiseToken value="frag"/>
  <noiseToken value="sherds"/>
  <facetInfo name="UseorContext" relevancy="0.8"/>
  <facetInfo name="Material" relevancy="1"/>
  <facetInfo name="CulturalGroup" relevancy="1"/>
  <facetInfo name="Color" relevancy="1"/>
  <facetInfo name="TechniqueDesignorDecoration" relevancy="1"/>
</colInfo>
<colInfo name="ObjNm_Remarks">
  <comment>There is not much here. Some techniques</comment>
  <tokenSep value="[;,]"/>
  <noiseToken value="each"/>
  <facetInfo name="UseorContext" relevancy="0.8"/>
  <facetInfo name="Material" relevancy="1"/>
  <facetInfo name="TechniqueDesignorDecoration" relevancy="1"/>
</colInfo>
<colInfo name="Culture">
  <tokenSep value="[;,\(\)]"/>
  <tokenSep value=" and "/>
  <facetInfo name="CulturalGroup" relevancy="1"/>
  <facetInfo name="color" relevancy="1"/>
</colInfo>
<colInfo name="Style">
  <comment>There is nothing here</comment>
</colInfo>
<colInfo name="Dynasty">
  <comment>There is nothing here</comment>
</colInfo>
<colInfo name="Period">
  <comment>There is nothing here</comment>
</colInfo>
<colInfo name="Reign">
  <comment>Great stuff, but no idea what to do with it. Find a way to create
    dates from these?</comment>
</colInfo>
<colInfo name="Nationality">
  <comment>Empty</comment>
</colInfo>
<colInfo name="N_Name">
  <comment>Empty</comment>
</colInfo>
<colInfo name="N_Title">
  <comment>Empty</comment>
```

```
</colInfo>
<colInfo name="N_Description">
  <comment>Empty</comment>
</colInfo>
<colInfo name="N_Signed">
  <comment>Empty</comment>
</colInfo>
<colInfo name="N_Markings">
  <comment>Empty</comment>
</colInfo>
<colInfo name="N_Inscription">
  <comment>Empty</comment>
</colInfo>
<colInfo name="N_Notes">
  <comment>Empty</comment>
</colInfo>
<colInfo name="N_CuratorRemarks">
  <comment>Empty</comment>
</colInfo>
<colInfo name="Movement">
  <comment>Empty</comment>
</colInfo>
<colInfo name="School">
  <comment>Empty</comment>
</colInfo>
<colInfo name="AltNum">
  <comment>Tons of useless numbers</comment>
</colInfo>
<colInfo name="altNums_Description">
  <comment>Tons of useless numbers</comment>
</colInfo>
<colInfo name="ObjectType">
  <comment>Useful distinction perhaps, but nothing to mine</comment>
</colInfo>
<colInfo name="TextEntry">
  <comment>Some provenience, some material, a few design</comment>
  <tokenSep value="[;:,\(\)]"/>
  <reduce from="\[ditto\]" to=""/>
  <reduce from="\[oakland crossed out\]" to=""/>
  <facetInfo name="Material" relevancy="0.8"/>
  <facetInfo name="CulturalGroup" relevancy="0.7"/>
  <facetInfo name="TechniqueDesignorDecoration" relevancy="0.7"/>
  <facetInfo name="SiteorProvenience" relevancy="0.8"/>
</colInfo>
<colInfo name="Purpose">
  <comment>Nothing useful, and probably in appropriate</comment>
</colInfo>
<colInfo name="ComponentName">
  <comment>Materials, decoration, color</comment>
  <tokenSep value="[;,:]"/>
  <tokenSep value="\([a-z]\)"/>
  <tokenSep value="w/"/>
  <noiseToken value="\([a-z]\)"/>
  <facetInfo name="color" relevancy="1"/>
</colInfo>
<colInfo name="ComponentNumber">
  <comment>Unused</comment>
</colInfo>
<colInfo name="ComponentType">
  <comment>Empty</comment>
</colInfo>
<colInfo name="PhysDesc">
  <comment>Materials, some provenenience</comment>
  <tokenSep value="[;,]"/>
  <noiseToken value="each"/>
```

```
    <noiseToken value="fragments"/>
    <noiseToken value="fragment"/>
    <noiseToken value="fragements"/>
    <noiseToken value="fragement"/>
    <noiseToken value="frgmnt"/>
    <noiseToken value="frags."/>
    <noiseToken value="frags"/>
    <noiseToken value="frag."/>
    <noiseToken value="frag"/>
    <reduce from="[\n+]" to=""/>
    <facetInfo name="Material" relevancy="0.8"/>
    <facetInfo name="SiteorProvenience" relevancy="0.8"/>
  </colInfo>
  <colInfo name="ObjTitlesTitle">
    <comment>The ObjTitlesTitle column has much placename and object name info,
        some activity, some materials, one or two colors.
        Exclusions for color: "campus green", "green valley". Will we have "clear"?
        If so, then excl: "clear lake". "black and white"???
        Probably also useful as part of notes for object - seems to have
        interesting comments on objects.
        A number of individuals mentioned, e.g., bill reid. Might be an
        interesting facet, but would take much work...
    </comment>
    <tokenSep value="[;,]"/>
    <facetInfo name="Material" relevancy="0.7"/>
    <facetInfo name="SiteorProvenience" relevancy="0.9"/>
    <facetInfo name="TechniqueDesignorDecoration" relevancy="0.8"/>
  </colInfo>
  <colInfo name="TableID">
    <comment>The TableID column may have some linking function,
      but it is useless for mining.
    </comment>
  </colInfo>

</dumpColConfig>
```

12. Appendix: MySQL Schemas

The MySQL schemas are fairly straightforward. We group them into the main objects schema and a schema for identity/authorization. The schemas are presented here and should be largely self-explanatory. Nevertheless, we provide a few notes of explanation.

We define a general `DBInfo` table for system constants, although some of these may be better maintained in PHP config files. The objects table contains the primary image to use, for speed and simplicity (the table is quite large, and we want to avoid another join when possible). We will likely add an `obj_imgs` table for additional images when we add that support in the next version. The sets functionality is fairly straightforward, and the `sets` and `set_objs` tables provides the basic functionality.

The `facets` and `categories` tables have information to build a tree from the DB entries, to support the facetted browser code. Each category row also caches counts for the number of object matches and matches for objects with images. This greatly speeds up generation of the main collection browse page. The synonyms are stored in the `hooks` table (this is a term from NLP that indicates strings that "hook" a category in the text). Exclusions have the more obviously named table. The `obj_cats` association table mapping mined categories to objects rounds out the schema.

In general, we include information about when user-created objects are first created and then updated, so that we can perform analysis of activity in the system at a later date. We have also defined a model of roles and permissions to easily manage access control of content in the system. At this point, we have very few roles defined - general users, administrators and researchers. Each user can be assigned one or more roles to describe how they can work in the system. For each role, we define permissions that are associated to that role. These include things like being able to modify certain tables, delete user-created objects, etc. We have not yet taken this very far, but will build upon this for the online review and reputation tracking work planned for a future release.

12.1. Main objects schema

```
-- ----------------------------------------------------------------------
-- SQL create script for delphi main object info tables
-- ----------------------------------------------------------------------

USE delphi;

-- The DBInfo table has a single row and is just used to hold system-wide
-- parameters such as the sizes of alternate image sizes, the version of this
-- DB schema, etc.
DROP TABLE IF EXISTS DBInfo \p;
CREATE TABLE DBInfo (
  `version`           VARCHAR(16) NOT NULL,
  -- TFacetMaskWidth must agree with the FacetMaskCache definition
  `facetMaskWidth`    TINYINT(2) NULL DEFAULT 32,
  -- cache the number of objects and those with images
```

```
  `n_objs_total`      INT(8) NOT NULL DEFAULT 0,
  `n_objs_w_imgs`     INT(8) NOT NULL DEFAULT 0,
  -- Be general about orientation, and store long and short side sizes.
  `thumb_long_side`   SMALLINT(4) UNSIGNED DEFAULT 80,
  `thumb_short_side`  SMALLINT(4) UNSIGNED DEFAULT 60,
  `medium_long_side`  SMALLINT(4) UNSIGNED DEFAULT 640,
  `medium_short_side` SMALLINT(4) UNSIGNED DEFAULT 480,
  `thumb_basepath`    VARCHAR(255) NULL,
  `medium_basepath`   VARCHAR(255) NULL,
  `large_basepath`    VARCHAR(255) NULL,
  `creation_time`     timestamp NOT NULL default '0000-00-00 00:00:00',
  `mod_time`          timestamp NOT NULL default CURRENT_TIMESTAMP
        on update CURRENT_TIMESTAMP
);
SHOW WARNINGS;

INSERT INTO DBInfo( version, creation_time )
  VALUES( '0.2 alpha', now() ) \p;
SHOW WARNINGS;

-- Define the main object table
-- The image info is denormalized for simplicity and performance,
-- based upon the initial system requirements. Must also
-- provide additional images with a proper normalized table.
DROP TABLE IF EXISTS `objects` \p;
CREATE TABLE `objects` (
  `id`            INT(10) UNSIGNED PRIMARY KEY NOT NULL,
  `objnum`        VARCHAR(80) NOT NULL,
  `name`          VARCHAR(255) NOT NULL,
  `description`   text NULL,
  -- All paths are relative to the configured image roots`
  `img_path`      VARCHAR(255) NULL,
  `creation_time` timestamp NOT NULL default '0000-00-00 00:00:00',
  `mod_time`      timestamp NOT NULL default CURRENT_TIMESTAMP
        on update CURRENT_TIMESTAMP,
  INDEX `obj_id_index` (`id`)
)ENGINE=MyIsam;
SHOW WARNINGS;

--Sets allow users (or the system) to group together items for
-- a particular reason (or simply as favorites).
DROP TABLE IF EXISTS `sets` \p;
CREATE TABLE `sets` (
  `id`             int(10) unsigned PRIMARY KEY NOT NULL auto_increment,
  `name`           VARCHAR(255) NOT NULL,
  `description`    text NULL,
  `policy`         enum('public', 'private'), -- later: friends, etc.
  `owner_id`       int(10) unsigned NOT NULL,
  `creation_time`  timestamp NOT NULL default '0000-00-00 00:00:00',
  `mod_time`       timestamp NOT NULL default CURRENT_TIMESTAMP
        on update CURRENT_TIMESTAMP,
  INDEX `sets_id_index` (`id`),
  INDEX `sets_id_owner` (`owner_id`),
  CONSTRAINT `se_ibfk_1` FOREIGN KEY (`owner_id`)
      REFERENCES `user` (`id`)
)ENGINE=MyIsam;
SHOW WARNINGS;

DROP TABLE IF EXISTS `set_objs` \p;
CREATE TABLE `set_objs` (
  `set_id`   int(10) unsigned NOT NULL,
  `obj_id`   int(10) unsigned NOT NULL,
  INDEX `so_set_index` (`set_id`),
  INDEX `so_obj_index` (`obj_id`),
    -- Ensure we have no dupes in a given set
```

```
  UNIQUE INDEX `so_so_index` (`set_id`,`obj_id`),
  CONSTRAINT `seo_ibfk_1` FOREIGN KEY (`set_id`)
      REFERENCES `sets` (`id`),
  CONSTRAINT `seo_ibfk_2` FOREIGN KEY (`obj_id`)
      REFERENCES `objects` (`id`)
)ENGINE=MyIsam;
SHOW WARNINGS;

/*
 * Categories are the nodes in the taxonomy.
 * Note that we tie each category to a facet explicitly.
 * We also specify how it is represented in the caches.
 * Note that categories that are tied to ranges, etc. do not have
 * the value in the cache - the category is assigned IFF the range
 * condition is met, and then the category acts as a boolean attribute.
 * TODO add creation and modification times.
 */
DROP TABLE IF EXISTS `categories` \p;
CREATE TABLE `categories` (
  `id`              INT(10) UNSIGNED PRIMARY KEY NOT NULL auto_increment,
  -- May not have a parent if facet root ; otherwise parent is non-null and >0.
  `parent_id`       INT(10) UNSIGNED NULL,
  -- Note that the names may not be unique, since they can appear in multiple
  -- facets and even in multiple places in one facet.
  `name`            VARCHAR(255) NOT NULL,
  `display_name`    VARCHAR(255) NOT NULL,
  `facet_id`        INT(10) UNSIGNED NULL,
  -- Whether children are exclusive
  `select_mode`     ENUM ('single', 'multiple' ) NOT NULL DEFAULT 'multiple',
  `always_inferred` TINYINT(1) NOT NULL default 0,
  `n_matches`       int(10) NOT NULL default 0,
  `n_matches_w_img` int(10) NOT NULL default 0,
  INDEX `cat_id_index` (`id`),
  INDEX `cat_name_index` (`name`),    -- Removed until need clarified
  CONSTRAINT `cat_ibfk_1` FOREIGN KEY (`parent_id`)
    REFERENCES `categories` (`id`),
  CONSTRAINT `cat_ibfk_2` FOREIGN KEY (`facet_id`)
      REFERENCES `facets` (`id`)
)ENGINE=MyIsam;
SHOW WARNINGS;

/*
 * Provides all the hooks (matching tokens) for Categories.
 * TODO add creation and modification times.
 */
DROP TABLE IF EXISTS `hooks` \p;
CREATE TABLE `hooks` (
  `id`       INT(10) UNSIGNED PRIMARY KEY NOT NULL auto_increment,
  `cat_id`   INT(10) UNSIGNED NULL,
  `token`    VARCHAR(255) NOT NULL,
  INDEX `hk_cat_index` (`cat_id`),
  INDEX `hk_token_index` (`token`),
  CONSTRAINT `hk_ibfk_1` FOREIGN KEY (`cat_id`)
      REFERENCES `categories` (`id`)
)ENGINE=MyIsam;
SHOW WARNINGS;

/*
 * Provides all the exclusions (countraindications) for Categories.
 * TODO add creation and modification times.
 */
DROP TABLE IF EXISTS `exclusions` \p;
CREATE TABLE `exclusions` (
  `id`       INT(10) UNSIGNED PRIMARY KEY NOT NULL auto_increment,
  `cat_id`   INT(10) UNSIGNED NULL,
```

```
  `token`    VARCHAR(255) NOT NULL,
  INDEX `ex_cat_index` (`cat_id`),
  INDEX `ex_token_index` (`token`),
  CONSTRAINT `ex_ibfk_1` FOREIGN KEY (`cat_id`)
      REFERENCES `categories` (`id`)
)ENGINE=MyIsam;
SHOW WARNINGS;

/*
 * Facets define how we set up the caches and the annotations.
 * A facet includes the facet name as well as policies for adding
 *  annotations using the facet categories.
 * Certain features are key in the UI binding, including the description of:
 *  - geographic placenames and relations (e.g., regional containment)
 *  - people names and relations (e.g., familial containment)
 * We mark these in the facets so we can pick the right place to hunt for
 *  the home towns and family placenames of Creators.
 * This script is here for documentation only; it is created from the facetmap
 * imported at index time. The column sizes for the maskIndex and maskSize
 * must match the optimized definitions in the FacetCache table.
 */
DROP TABLE IF EXISTS `facets` \p;
CREATE TABLE `facets` (
  `id`             INT(10) UNSIGNED PRIMARY KEY NOT NULL auto_increment,
  `name`           VARCHAR(255) NOT NULL,
-- In UI, if not filtering. E.g.: “All People”
  `display_name`   VARCHAR(255) NOT NULL,
  `num_categories` INT(10) UNSIGNED UNSIGNED NULL,
  `num_masks`      TINYINT(3) UNSIGNED NULL
--  `root_cat_id`  INT(10) UNSIGNED NULL,
--  CONSTRAINT `fcat_ibfk_1` FOREIGN KEY (`root_cat_id`)
--      REFERENCES `categories` (`id`)
)ENGINE=MyIsam;
SHOW WARNINGS;


/*
 * Provides associations of categorizations of objects.
 * TODO add creation and modification times.
 */
DROP TABLE IF EXISTS `obj_cats` \p;
CREATE TABLE `obj_cats` (
  `obj_id`      int(10) unsigned NOT NULL,
  `cat_id`      int(10) unsigned NOT NULL,
  `inferred`    TINYINT(1) NOT NULL DEFAULT 0,
  `reliability` TINYINT(1) NOT NULL DEFAULT 9,
  INDEX `oc_obj_index` (`obj_id`),
  INDEX `oc_cat_index` (`cat_id`),
  UNIQUE INDEX `oc_obj_cat_index` (`obj_id`,`cat_id`),
  CONSTRAINT `obc_ibfk_1` FOREIGN KEY (`obj_id`)
      REFERENCES `objects` (`id`),
  CONSTRAINT `obc_ibfk_2` FOREIGN KEY (`cat_id`)
      REFERENCES `categories` (`id`)
)ENGINE=MyIsam;
SHOW WARNINGS;


CREATE TABLE `tags` (
`tag_id` int(11) NOT NULL auto_increment,
`tag_name` varchar(50) NOT NULL,
`tag_count` int(11) NOT NULL default ‘1’,
PRIMARY KEY (`tag_id`),
UNIQUE KEY `tag_name` (`tag_name`),
KEY `tags_idx` (`tag_id`)
) ENGINE=MyISAM
```

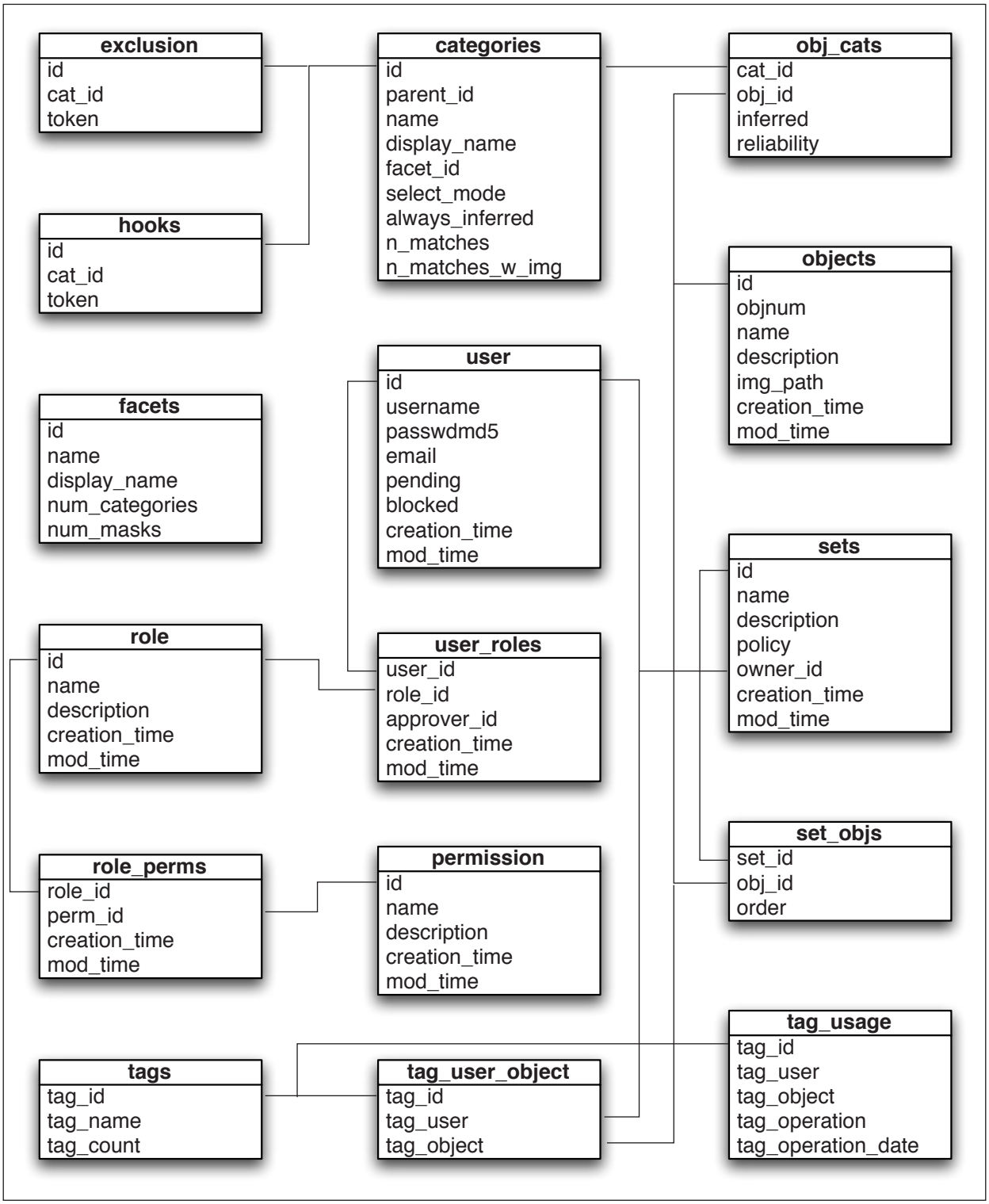
```
CREATE TABLE `tag_user_object` (
`tag_id` int(11) NOT NULL,
`tag_user_id` int(11) NOT NULL,
`tag_object_id` int(11) NOT NULL,
PRIMARY KEY (`tag_id`,`tag_user_id`,`tag_object_id`),
KEY `tag_id_idx` (`tag_id`),
KEY `tag_user_idx` (`tag_user_id`),
KEY `tag_object_idx` (`tag_object_id`)
) ENGINE=MyISAM

CREATE TABLE `tag_usage` (
`tag_id` int(11) NOT NULL,
`tag_user_id` int(11) NOT NULL,
`tag_object_id` int(11) NOT NULL,
`tag_operation` enum('create','add','delete') default NULL,
`tag_operation_date` date NOT NULL
) ENGINE=MyISAM
```

12.2. Identity/Authorization schema

```
--
-- Table structure for table `user`
-- Should the user table have the core role to denormalize and save a join?
-- How often will users have multiple roles?
--
DROP TABLE IF EXISTS `user`;
CREATE TABLE `user` (
  `id`            int(10) unsigned NOT NULL auto_increment,
  `username`      varchar(40) NOT NULL UNIQUE,
  `passwdmd5`     varchar(32) NOT NULL,        -- MD5 of the pw
  `email`         varchar(80) NOT NULL,           -- allow for very long email addresses
  `pending`       boolean NOT NULL default true, -- on creation, is unverified
  `blocked`       boolean NOT NULL default false,
  `creation_time` timestamp NOT NULL default '0000-00-00 00:00:00',
  `mod_time`      timestamp NOT NULL default CURRENT_TIMESTAMP
        on update CURRENT_TIMESTAMP,
  PRIMARY KEY  (`id`),
  INDEX `user_i_username` (`username`)
) ENGINE=MyIsam DEFAULT CHARSET=latin1;
SHOW WARNINGS;

--
-- Table structure for table `role`
--
DROP TABLE IF EXISTS `role`;
CREATE TABLE `role` (
  `id`            int(10) unsigned NOT NULL auto_increment,
  `name`          varchar(40) NOT NULL UNIQUE,
  `description`   text,
  `creation_time` timestamp NOT NULL default '0000-00-00 00:00:00',
  `mod_time`      timestamp NOT NULL default CURRENT_TIMESTAMP
        on update CURRENT_TIMESTAMP,
  PRIMARY KEY  (`id`),
  INDEX `role_i_name` (`name`)
) ENGINE=MyIsam DEFAULT CHARSET=latin1;
SHOW WARNINGS;

--
-- Table structure for table `permission`
--
DROP TABLE IF EXISTS `permission`;
CREATE TABLE `permission` (
```

```
  `id`            int(10) unsigned NOT NULL auto_increment,
  `name`          varchar(40) NOT NULL UNIQUE,
  `description`   text,
  `creation_time` timestamp NOT NULL default '0000-00-00 00:00:00',
  `mod_time`      timestamp NOT NULL default CURRENT_TIMESTAMP
        on update CURRENT_TIMESTAMP,
  PRIMARY KEY  (`id`),
  INDEX `perm_i_name` (`name`)
) ENGINE=MyIsam DEFAULT CHARSET=latin1;
SHOW WARNINGS;

--
-- Table structure for table `user_roles`
-- Associates roles to users
--
DROP TABLE IF EXISTS `user_roles`;
CREATE TABLE `user_roles` (
  `user_id`       int(10) unsigned NOT NULL,
  `role_id`       int(10) unsigned NOT NULL,
  `approver_id`   int(10) unsigned default NULL,  -- Do we want to track this?
  `creation_time` timestamp NOT NULL default '0000-00-00 00:00:00',
  `mod_time`      timestamp NOT NULL default CURRENT_TIMESTAMP
        on update CURRENT_TIMESTAMP,
  PRIMARY KEY  (`user_id`, `role_id`),
  CONSTRAINT `ur_ibfk_1` FOREIGN KEY (`user_id`)
        REFERENCES `user` (`id`),
  CONSTRAINT `ur_ibfk_2` FOREIGN KEY (`role_id`)
        REFERENCES `role` (`id`),
  CONSTRAINT `ur_ibfk_3` FOREIGN KEY (`approver_id`)
        REFERENCES `user` (`id`)
  -- Do we need to index the roles? Will we ask "who all has role X"?
) ENGINE=MyIsam DEFAULT CHARSET=latin1;
SHOW WARNINGS;

--
-- Table structure for table `role_perms`
-- Associates permissions to roles
-- Another possibility would be to have a mask or two of perms
-- that are stored in the role table, again to save joins.
--
DROP TABLE IF EXISTS `role_perms`;
CREATE TABLE `role_perms` (
  `role_id`       int(10) unsigned NOT NULL,
  `perm_id`       int(10) unsigned NOT NULL,
  `creation_time` timestamp NOT NULL default '0000-00-00 00:00:00',
  `mod_time`      timestamp NOT NULL default CURRENT_TIMESTAMP on update CURRENT_TIMESTAMP,
  UNIQUE KEY  (`role_id`,`perm_id`),
  CONSTRAINT `rp_ibfk_1` FOREIGN KEY (`role_id`) REFERENCES `role` (`id`),
  CONSTRAINT `rp_ibfk_2` FOREIGN KEY (`perm_id`) REFERENCES `permission` (`id`)
  -- Do we need to index the roles? Will we ask "who all has role X"?
) ENGINE=MyIsam DEFAULT CHARSET=latin1;
SHOW WARNINGS;
```

13. Appendix: Schema diagram



Schema diagram with key relations.