

## NARRATIVE

### 1. INTRODUCTION AND NATIONAL IMPACT

Anyone setting out to learn a new topic would, traditionally, take a pad of paper, starting with resources available at home, visit the reference area of the local library, then move on to a steadily widening circle of resources: the library stacks, other libraries, museums, archives, local, state or federal government agencies, cultural and historical societies, newspapers, experts, and so on, depending on the topic and the purpose. The library reference collection is especially convenient for addressing the What?, When?, Where?, Who?, Why?, and How? because it has a carefully selected specialized resources. Dictionaries, encyclopedias, and the library subject headings help clarify the topic and its terminology, and the events, dates, institutions and persons involved. Biographical dictionaries describe individuals. Chronologies and newspaper indexes pinpoint pertinent dates. Gazetteers and atlases show where places are. Associating topics with persons, times, and places is especially important when learning about the arts, history, and cultural heritages. Bibliographies, catalogs, and directories lead to fuller resources in collections and resources in archives, libraries and museums.

The fruit of this searching will be notes from the diverse resources and the names, places, and topics noted are leads to more resources. The notes won't be entirely compatible, as sources differ in both what they say and how they say it, but a human can cope with inconsistencies.

The national investment in collection, preservation, conservation, and digitization in museums and libraries is huge. The program has been resource-centric, focusing on the developing, sustaining, and improving each object and each collection, a natural, sensible, and, perhaps, inevitable way to proceed. Access to collections has been built towards potential users, mostly in a "stove-pipe" style. Library and museum environments have yet to provide learners with a structured environment analogous to the reference collection. The digital environment is still weak in providing that kind of *intermediate zone* between the learner and the best resources. The web environment lacks this kind of structure and most good material is in the "deep Web" out of range of Web search engines. The approach has been primarily that of a publisher, producing one book after another, rather than of a librarian whose task it is to form a coherent array of resources for use (Buckland 2003). More assistance is needed for navigating multiple metadata, building crosswalks between different vocabularies, and integrating search results into personal computing environments.

*We propose to demonstrate how, using emerging standards and recent research, intermediate infrastructure could now provide access to the best available resources with the kind of supportive learning environment that one associates with a reference library, support learning about the what, where, when, and who of a topic, event, or cultural object, and the compiling a dossiers on the most relevant resources.*

#### **WHAT? - Topical Search across Different Media**

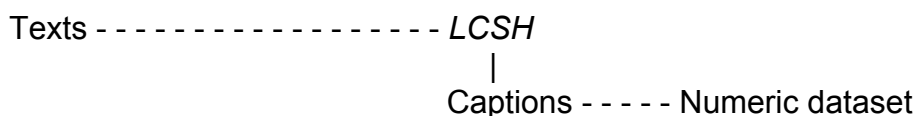
Topical metadata (classifications, indexes, subject heading lists, thesauri, etc.) are used to categorize objects in collections. But getting from the words that the learners use (the *query vocabulary*) to the terms in the metadata (*entry vocabulary*) is a fundamental problem. How is anybody looking for ALIEN LIFE FORMS to know that they should have been searching under EXTRATERRESTRIAL BEINGS and LIFE ON OTHER PLANETS? Who could guess the US Patent Classification for PEANUT BUTTER? "Relative indexes" (Dewey's name), leading from the terms and phrases of users to the most apt choices in the topical metadata, indexes from the *Query vocabulary* to the *Entry vocabulary*. Learners also need help in moving between *entry*

vocabularies. If you did know that the US Patent Classification number for peanut butter is 426/633.00, how easily would you find the corresponding International Patent Classification code: A23L 1/38? The Unified Medical Language System (UMLS) of the National Library of Medicine is the outstanding example of mapping across vocabularies, but very expensive to extend or replicate. Fortunately, tools based on Prof. Larson's "classification clustering" technique allow the rapid and inexpensive generation of indexes to (and between) entry vocabularies using statistical association and natural language processing techniques (Buckland, Chen, et al. 1999). Examples of such *entry vocabulary indexes* are available at <http://metadata.sims.berkeley.edu/prototypes1.html> include:

- English queries to *Library of Congress Subject Headings* (LCSH);
- English queries to the *Standard Industrial Classification* (SIC);
- Mapping between the SIC and its successor NAICS; etc.

These kinds of mappings are especially needed in a digital environment where schedules are not so easily arrayed for visual inspection – and *also* to search across different media.

The transformation into a digital environment reduces all media (audio, numeric, text, images, and video) to bits. Yet searching across media is really hard in a digital environment. It cannot be done directly (Buckland 1991). A fragment of text can be used to search in a text file, but not among images or statistical data series. Nor can images or numeric data be searched in text files. Searching across digital media can, at best, be done *indirectly*: The headings, row stubs, and captions in numeric datasets can be used as a textual surrogate and searched by (or as) text queries, but such a search is unlikely to be reliable given the vagaries of language. But if both texts and captions can be assigned, for example, *LCSH* headings, then the *LCSH* headings can be used as a pivot to go from text to topically related numeric data, or vice-versa, thus:



This approach was demonstrated in our IMLS-funded project *Seamless Searching of Numeric and Textual Resources* (Buckland, Gey & Larson 2002). We found, however, that topical searching in socio-economic numeric datasets almost always has a geographical aspect: The data pertains to some area: country, state, county, or the like. But place names are ambiguous and unstable, and numeric data series often deal with different *kinds* of places. (Would you refer to a restaurant as being in Census tract XYZ?) Searching by place is needed, but not well supported.

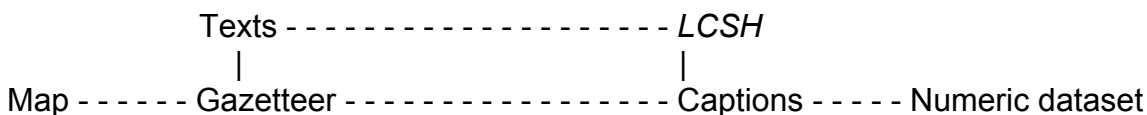
## WHERE? - Place Names and Locations

People want to learn about birds in Bolivia, the castles of Quercy, hiking in the Himalayas, where their ancestors lived, and so on, but searching by place is still hard. Catalogs, reference works, and socio-economic data series use place *names*, usually of geopolitical entities. But names are ambiguous (e.g. which Lafayette?), have variant forms (St. Petersburg, Санкт-Петербургский, Saint-Pétersbourg, etc.), and unstable (e.g. St. Petersburg 1703-1914; Petrograd 1914-1924; Leningrad 1924-1991; St. Petersburg, again, in 1991). Geopolitical entities are unstable because *boundaries* and *political structures* change. Think of local government boundaries, the Balkans, or the widely varying boundaries, over the years, of Poland.

Searches involving regions other than geopolitical entities can be difficult. Yet places, unlike topics, persons, institutions, and events, have a system for objective specification, latitude and longitude, and there is a well-established tool for linking place names with places: the *gazetteer*, most familiar as a list in back of atlases, serving as an index to the maps. Coupling online gazetteers with online catalogs would not only provide place name disambiguation, but also the

data needed for visualizing queries and retrievals in map form, and the ability to extend searches to places nearby (Buckland, Gey, Larson 2002).

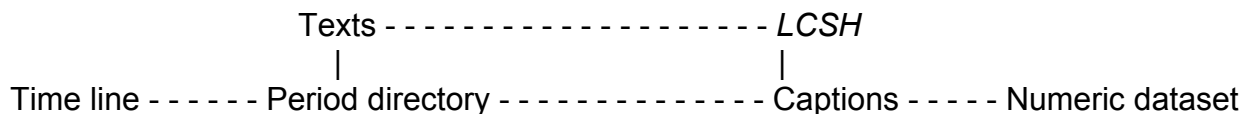
Gazetteers, in library terminology, are place name authority files. Latitude and longitude can both disambiguate the numerous Lafayettes and show the identity of different names, e.g. Beijing and Peking. (A demonstration of how bibliographical searching by place could be improved by coupling an online catalog with an online gazetteer – with values for latitude and longitude derived from the gazetteer enabling the use of a map visualization – nears completion in our IMLS-supported project “Going Places in the Catalog: Improved Geographical Search” (ecai.org/imls2002/). The structure above could and should be as follows:



But, since place names and the places themselves are unstable, each entry a gazetteer should have a time aspect: *When was that* place name in use? This leads to the importance of time.

**WHEN? - Time Period Names and Calendar Time**

Just as searches may often have a geographical aspect, there may also be a temporal aspect. Like place, time has a dual naming system: Names for periods (reigns, wars, administrations, etc.) and calendar date definitions, (e.g. 1939-1945). Directories of names of periods of time are not a developed genre, but the chronological subdivisions of LCSH (\$) are examples. And, just as gazetteers support maps, time period directories support time lines. (A nice example developed by our students is HumanSaga ([www.HumanSaga.com](http://www.HumanSaga.com))). Just as a map allows one to see what else is at or near the *same place*, time periods and chronologies (such as Wikipedia’s *List of themed timelines* and others noted in Feinberg et al. 2003) enable one to see what else was happening at, or about, the *same time*. So we also need the following:



*Gazetteers and Time period directories are related.* Just as gazetteer entries have a period aspect, time periods may have a geographical aspect. *When* the Neolithic period was depends on the culture and, therefore, the region. MARC records reveal multiple “civil war” periods: 49-48 BC (Rome); 1642-1649 (Great Britain); 1861-1865 (United States); 1936-1939 (Spain); 1945-1949 (China); 1967-1970 (Nigeria); 1970-1975 (Cambodia); and more. These points in place and time could be shown in a time line and/or on a map. (See Appendix A).

Understanding historical context requires the ability to focus on both place *and* time. For an example of a search interface combining a map *and* a time line see our *Hindi Surprise Language Project - Document Display in Java TimeMap*: <http://ecai.org/imls2002/hindi/HindiDocRetrieve.html> (Reproduced as Appendix B)

We have been working on the design of a directory of time period names, and hope to have an online prototype by May 2004 (Feinberg 2003). It closely resembles, in structure, a gazetteer:

<u>Place name</u>	<u>Kind of place</u>	<u>Where (lat. &amp; long. values)</u>	<u>When</u>
<u>Period name</u>	<u>Kind of period</u>	<u>When (calendar values)</u>	<u>Where</u>

**WHO? - Biographical Dictionaries**

Just as gazetteers disambiguate place names, personal name authority lists disambiguate persons. A biographical dictionary (aka Who's Who) is a name authority file with biographical information attached, with an underlying structure: a sequence of triplets of date, place, and topic/action, often with proper names (person or institution) also, e.g.

Emanuel Goldberg. Born Moscow 1881. PhD under Wilhelm Ostwald, University of Leipzig 1906. Director Zeiss Ikon, Dresden, 1926-1933. Moved to Palestine 1937. Died Tel Aviv 1970.

Almost every element could be expanded: Moscow, Dresden, etc., in a gazetteer; Ostwald in a biographical dictionary; Zeiss Ikon through a business directory; etc. The dates can be used to learn what else was happening at that time in that place or topic. A search on Leipzig (aka Leipsic!) qualified by publication date 1906 (or soon after) yields contemporary descriptions.

## Learning

To learn about something is to place it into a comprehensible (culturally relevant) context. Addressing multiple avenues for learning of materials is at the heart of new teacher education and forms the basis of what all new student teachers need to show they can do in order to get their preliminary teaching credential. (In California, this is the focus of the California Commission on Teacher Education assessment tool--The Teacher Performance Assessments). The crux of all that is new in teacher credentialing and testing is the use of multiple avenues into materials based on educational theories about learners and learning, multiple intelligences, and the need for differentiated instruction (Gardner 1993, Goleman 1995). In this context, giving teachers the ability to search by who, when and where--person, time and place—to select diverse resources is important: Learning about *other persons* aids interpersonal understanding; working with *time* is important for developing analytical thought; and notions of *place* underlie any sense of movement (kinesthetic/naturalistic understanding).

Traditional advice for anyone undertaking to learn more about some topic is to use the six facets Who, What, When, Where, Why, and How. The first four are each supported by specialized tools in a traditional reference collection. But web search engines don't support this systematic, analytical approach and library catalogs and online bibliographies support it only weakly. We request support to demonstrate how the digital environment for learning could be transformed by constructing *intermediate infrastructure* analogous to that provided in a reference library.

## Metadata and the Architecture of “Intermediate infrastructure”

Records for individual objects in collections cannot carry much contextual content. The metadata for every record would become large and obsolescent - like an encyclopedia! Nor can the user's workstation hold all the contextualizing resources. The only workable and cost-effective solution is to have access to an *intermediate environment* containing access to basic reference tools and the relationships between them: gazetteers; time period directories; biographical dictionaries; as well as bibliographies and collections and also the relationships between them. We need an environment that makes it easy to move between specialized resources, following up clues and links – and to compile a selective dossier of excerpts from them. The resources would be network-accessible reference resources, e.g. gazetteer servers such as the Alexandria Digital Library gazetteer, the NIMA GEOnet Names Server (GNS), and the *Getty Thesaurus of Geographic Names On Line*, with suitable “service protocols.”

Harmonizing the excerpts from heterogeneous resources into a coherent dossier is a significant challenge (cf Lagoze and Hunter's work with CIMI, 2001), but METS (the Metadata Encoding and Transmission Standard) offers a solution not available when we addressed this

challenge a decade ago (Buckland, Butler, et al. 1994). We envisage using METS to enable learners to create dossiers in the form of an ad hoc web portal.

### Standards and Protocols

Searching disparate sources requires interoperability which has three aspects:

Search across collections, which comes in three forms:

- Crawling: Web crawlers search networks and accumulate anything matching the query;
- Harvesting: Collecting index data to create a “union index.” The Open Archive Index Protocol for Metadata Harvesting (OAI-PMH) is a popular protocol (Cole 2003).
- Federating: Search and retrieve protocols, primarily ISO 23590 (aka Z39.50), enables any single client to search any remote server.

Once the searching is completed the retrieved set can be used for multiple purposes, including searches of what has been collected, e.g. the University of Illinois Champaign-Urbana Digital Gateway to Cultural Heritage Materials portal. Whereas crawling and harvesting produce inherently obsolescent access, federated searches guarantee current results.

Format interoperability: Conversion mechanisms (“crosswalks”) are needed where different metadata formats are in use, e.g. between Z39:2 MARC to Dublin Core and back, nowadays done via XML.

Content interoperability: Interoperable mechanics are in vain if the indexing terms and their meanings are inconsistent. Solutions are the shared adoption of a single controlled vocabulary (e.g. *Library of Congress Subject Headings*) or detailed mappings of relationships across vocabularies, either handcrafted by experts (e.g. UMLS) or computer-generated (our Entry Vocabulary Indexes).

Our concern is to show, through working examples, how the basic amenities associated with a reference library could be developed in the emerging museum and digital library environments. We resolutely use existing standards to the extent feasible, and will make recommendations for changes and additions sparingly. But it is clear that there is much to be done. There is, surprisingly, no national or international standard for the content or format of gazetteer entries. (We were encouraged by NISO to propose a draft gazetteer standard, based on our NSF-funded collaboration with Academia Sinica and the Alexandria Digital Library, whose Gazetteer Standard is now substantially what we would recommend, but we lacked the resources to follow through. (Mostern & Brose 2004a, b) Although one can point to excellent developments (e.g. in the Perseus Digital Library; and Informedia II (Wachtlar 2003)) *standards* for biographical dictionaries and time period directories also appear lacking.

The tools to do what we propose are now at hand, especially three:

- XML for structured records and for conversion between different formats;
- ISO 23950 (aka Z39.50) for federated search;
- METS (Metadata Encoding and Transmission Standard) for encoding descriptive, administrative, and structural metadata describing digital objects.

## 2. ADAPTABILITY

The project addresses a general and shared problem. If successful, the techniques embodied in the prototypes will be available for others to adopt or improve on. Our experience is that it is the *demonstration* of effective methods that leads to innovation, rather than the software itself. Nevertheless, the software will be documented and made freely available as source code. Our resolute adherence to the best available standards maximizes the ease of adoption.

## 3. DESIGN

The overall project objective is to design, develop, demonstrate, and evaluate a client search interface with a working prototype of an intermediate infrastructure available as an aid to searching network accessible collections, with special attention to What, Where, When, and Who, and compiled specialized, highly selective dossiers of the most relevant resources. Note the *Table of Relationships and Objectives* at the end of this Narrative.

Year One is concerned with building the underlying structure: an interface with tools to access at least one instance of the four types of specialized resources for What, When, Where and Who respectively.

Year Two will be used to add additional resources in each of these four types, to develop dossier (portal) creation software, and for Testing and Evaluation.

### **Direct Support for the Learner**

Objective 1: Build a client interface. A convenient interface with the tools needed to find and to use the specialized resources discussed below. (Year 1).

Objective 2: Dossier software. A tool to enable the Learner to manage a dossier of multimedia selections and to keep them in the form of an ad hoc web portal. (Year 2).

### **Constructing Infrastructure**

Objective 3: Support for WHAT: Convenient access to topical ontologies. At least one in Year 1 (we assume LCSH), with an added index (“entry vocabulary index”) leading from the *Query vocabulary* of the learner to the stylized terminology of LCSH entries (“entry vocabulary”). Also the software for displaying syndetic structure: Preferred and Non-preferred terms, Broader terms, Narrower terms, and Related terms. In Year 2, we add others.

Objective 4: Support for WHERE: Convenient access to gazetteers. At least one in Year 1, we assume the NIMA *GEOnet Names Server*. Also the map visualization software both to display selected records and also as a basis for specifying locations as queries. In Year 2, we add others.

Objective 5: Support for WHEN: Convenient access to time period directories. In Year 1, we plan to start with an experimental directory derived largely from harvesting LCSH chronological subdivisions. Also software for creating and displaying timelines. In Year 2, we add others.

Objective 6: Support for WHO: Convenient access to biographical dictionaries. In Year 1 we plan to start with a large library Name Authority File and a biographical dictionary to be selected. Also software for displaying selected records. In Year 2, we add others.

**Testing and Evaluation.** (Explained below.) – [Revised May 3, 2005: The original objectives 7 & 8 were replaced by a single objective 7.]

Objective 7: Evaluation by university faculty of the effectiveness of this intermediate infrastructure for improving undergraduate instruction (Year 2).

**Adaptability, Documentation and Dissemination.** (Explained below.)

Objective 9: Make recommendations for improved standards and protocols in the light of our experience. (Both years.)

Objective 10: Document and disseminate findings, difficulties and achievements: Progress reports (Year 1); Final documentation and dissemination (Year 2).

See [TABLE OF RELATIONSHIPS AND OBJECTIVES](#).