MAPSEARCH: A PROTOCOL AND PROTOTYPE APPLICATION TO FIND MAPS

by

JUDITH GELERNTER

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Communication, Information and Library Studies

Written under the direction of

Professor Michael E. Lesk

and approved by

____

____

____

____

New Brunswick, New Jersey

May 2008

ABSTRACT OF THE DISSERTATION

MAPSEARCH: A Protocol and Prototype Application to Find Maps

By Judith Gelernter

Dissertation Director:
Professor Michael Lesk

Even geographers need ways to find what they need among the thousands of maps buried in map libraries and in journal articles. It is not enough to provide search by region and keyword. Studies of queries show that people often want to look for maps showing a certain location at a certain time period or with a subject theme. The difficulties in finding such maps are several. Maps in physical and digital collections often are organized by region. Multi-dimensional manual indexing is time-consuming and so many maps are not indexed. Further, maps in non-geographical publications are indexed rarely, making them essentially invisible.

In an attempt to solve actual problems, this dissertation research automatically indexes maps in published documents so that they become visible to searchers. The MapSearch prototype aggregates journal components to allow finer-grained searching of article content. MapSearch allows search by region, time, or theme as well as by keyword (http://scilsresx.rutgers.edu/~gelern/maps/).

Automatic classification of maps is a multi-step process. A sample of 150 maps and the text (that becomes metadata) describing the maps have been copied from a random assortment of journal articles. Experience taking metadata manually enabled the writing of instructions to mine data automatically; experience with manual classification allowed for writing algorithms that classify maps by region, time and theme automatically. That classification is supported by ontologies for region, time and theme that have been generated or adapted for the purpose and that allow what has been called intelligent search, or smart search. The 150 map training set was loaded into the MapSearch engine repeatedly, each time comparing automatically-assigned classification to manually-assigned classification. Analysis of computer misclassifications suggested whether the ontology or classification algorithm should be modified in order to improve classification accuracy. After repeated trials and analyses to improve the algorithms and ontologies, MapSearch was evaluated with a set of 55 previously unseen maps in a test set. Automated classification of the test set of maps was compared to the manual classification, with the assumption that the manual process provides the most accurate classification obtainable. Results showed an accuracy, or a correspondence between manual and automated classification, of 75% for region, 69% for time, and 84% for theme.

The dissertation contributes: (1) a protocol to harvest metadata from maps in published articles that could be adapted to aggregate other sorts of journal article components such as charts, diagrams, cartoons or photographs, (2) a method for ontology-supported metadata processing to allow for improved result relevance that could be applied to other sorts of data, (3) algorithms to classify maps into region, time and theme facets that could be adapted to classify other document types, and (4) a proof-of-concept MapSearch system that could be expanded with heterogeneous map types.

Acknowledgement

I am thankful for my chief advisor Michael Lesk's wisdom and kindness that have guided me throughout this research. His exceptional knowledge of data mining and word sense disambiguation allowed me to delve deeply into the heart of the problem. His translation of my algorithms into the web programming language Perl breathed life into MapSearch and created the prototype. I will miss our work on this project, but look forward to research together in future.

Thanks are due to committee members in my department Dan O'Connor and Marie Radford for advice on the rigors of form and style, and for being available when I have been in need. Thanks also to Michael Goodchild from the University of California at Santa Barbara for sharing some of his latest research and offering advice on this dissertation's geographic aspects.

I thank my colleagues in the graduate school Heather Moulaison and Sarah Legins for the careful attention they gave to indexing the map sample, which was used for the evaluation of the classification algorithms.

Nancy Kandoian of the Map Division of the New York Public Library, and Cathay Crosby of the Internet Public Library, were most helpful in providing data on how people ask for maps.

This research was supported in part by the Rutgers Academic Excellence Fund.

Table of Contents

List of tables

List of illustrations

Concluding Material

1 Introduction

1.1 Problem statement

How do we find maps in non-geographic publications? In digital publications, often we do not—maps are invisible because they are not well-indexed. Standard book records in library catalogs contain a descriptive field that tells whether a book includes a map, but does not describe the map. Articles rarely have even this level of indexing. This dissertation research is one step toward answering the map problem.

1.2 Significance

How do we know missing maps is a problem? People look for geographic data daily, as revealed in search engine logs. Sanderson and Han (2007) analyzed four weeks worth of one million queries from "a large search engine" in 2004 and found that geography words contributed the largest percentage of any query category (other categories were activity, adult, arts and humanities, shopping, computer, education, healthcare, people and science). Geography might be the largest search engine category, but how many queries might it comprise? Eighteen percent of the queries submitted to the EXCITE search engine in 2001 contained geographically-related terms (Sanderson & Kohler, 2004). Fourteen percent of queries submitted within a half-year period to the major Brazilian search engine TodoBR contained at least one geographically related term (Borges, Laender, Medeiros & Davis, 2007, p. 31).

> *So say an average of 15% of 1 million queries collected over one month are geographically-related = 150,000 queries.*

Henrich & Lüdecke (2007, p.5) found that about 12% of geographic queries are of "pure informational character" without the intent of going to a place. Many of these surely are candidates for theme maps, which are maps that overlay data on a basemap.

> *So 150,000 x 12% =18,000 queries per month [or almost a quarter million queries per year] per major search engine might be answered with a theme map.*
>
> *The numbers would be exponentially more if applied to Google which, according to searchenginewatch.com, hosted 37 billion search queries during the month of August 2007, or over 1 billion per day.*

Supposing that some proportion of these geographic queries could be answered by theme maps, this research makes published maps visible to search by keyword and classification category. The research rests upon previous work in data mining and automatic classification with ontology-supported retrieval. Data mining here is in lieu of hand-cataloging that is the traditional basis for automatic classification. Ontology-supported search aims to widen retrieval relevance in searches by keyword and classification category. Terminology underlying the information retrieval and geographic information science technicalities of this dissertation are in Appendix A., and a prototype is accessible at http://scilsresx.rutgers.edu/~gelern/maps.

1.3 Research questions

This dissertation proposes a method to make a large body of maps accessible by classification category. It outlines a method to mine maps from journal articles and then classify them without human supervision. The questions driving research are just how well can those methods work.

The first research question is: how well it is possible to classify maps automatically by region? The second is: how well is it possible to classify maps automatically by time period? The third is: how well it is possible to classify maps automatically by subject? The measure of a correct automatic classification is classification done manually. Despite the fact that two people will not choose the same categories invariably (as is shown in the section on evaluation), enough assigning of categories is unambiguous to make measuring automatic classification by a manual benchmark seem acceptable. The prototype does not need to excel at recall and precision because it has no actual competitors; it need perform only well enough to be cost-effective, and so justify its implementation.

1.4 Contributions and long term goals

Eventually the research could be broadened by applying methods contributed in this dissertation to classifying various sorts of data mined from documents, and it could be deepened by continuing to fortify the MapSearch retrieval system and enlarging its map collection.

Broaden the study

The opinion of most participants in a recent study focused on a database of journal article components (not just maps) confirmed that there is a “consistent, unmet need for systems that yield higher precision searches…[to] journal article components like figures, tables, graphs, maps and photographs” (Sandusky & Tenopir, 2008, p.977). A collection of journal article components would provide a finer-grained means to access the literature, and also a means to create reports or meta-analyses based on these components. This dissertation, therefore, is one step toward that goal in that the data mining and classification algorithms devised here for maps could be adapted to mine components of other types. The worth of the components rests largely on their accuracy, and that is of course outside the control of the system that aggregates them.

Explored also in this dissertation is ontology-supported retrieval. An ontology is a type of controlled vocabulary that groups related words, as discussed further in some of the chapters following. It improves retrieval by filling in semantic gaps such that, for example, a user’s query on “car” will retrieve items about automobiles that do not use the word “car,” and that consequently would be missed in a search using keyword matching. This is called recall. The manually-created ontologies, in the present context, have led to recall that is fairly high.

The importance of ontologies in improving recall has been recognized widely. The World Wide Web Consortium is developing its own Web Ontology Language (OWL).[1] The ontologies created manually for time and subject in MapSearch, abbreviated for the prototype, can be expanded using methods shown here for collections that are vast. Research pursuant to this dissertation might suggest ways to expand ontologies automatically or suggest how pre-existing ontologies could be implemented more effectively.

Another way to broaden the study would be to apply the hybrid method of classification to other sorts of items. Typically, classification is wholly automated. Combining manually-

[1] OWL Web Ontology Language Overview by the World Wide Web Consortium, at http://www.w3.org/TR/owl-features/, retrieved February 22, 2008.

created ontologies with automatic classification of items into categories is a little-tried method that has been used in MapSearch with success.

Deepen the study

Those who rely upon maps might appreciate a working version of a system like MapSearch that affords easy access to maps by region, time period and theme. The need for a system that finds maps is underlined by the many sorts of difficulties in finding what is available currently: image searches for maps often are confined to search by keyword, driving direction maps are available on-the-fly but generally not in collections, and many sheet maps are undoubtedly uncataloged owing to arduous metadata specifications for manual map cataloging. Beyond this, many maps within articles are invisible to search because they are not invariably indexed.

One way to fortify MapSearch would be to enlarge its collection with maps or links to maps. The first step would be to add many more maps mined from published articles. The maps would come from heterogeneous sources. Some would be hand-cataloged and some would not. Retrieving from heterogeneous sources within a single interface represents an aspect of the information fusion problem in computer science, and one that must be considered for MapSearch to mature.

2 Background for the research

This section begins with a definition of map and delves into how people find or do not find digital maps in general, and theme maps in particular. MapSearch is presented as a possible solution to these problems. The approach taken to create the MapSearch prototype is outlined.

2.1 What is a map?

Map derives from the Latin *mappa*, signifying the cloth on which representations of the world were made. *Webster's New World Dictionary* gives a first definition of map as a "drawing or other representation, usually on a flat surface, of all or part of the earth's surface, ordinarily showing countries, bodies of water, cities, mountains, etc." The broad dictionary definition would include a simple photograph of the earth's surface, and the orthophotographs created by remote sensing, although these are excluded from the MapSearch database. The maps in the prototype database include single symbol maps, graduated size maps (in which symbols are altered in size to reflect data), graduated color maps (also called choropleth, which use color or shading to reflect data value) and combinations of these.[2]

2.2 How can one find a digital map?

Sheet maps are generally organized in physical map libraries by region, so they are difficult to find by different access points such as subject. Digital map collections, called *geolibraries*, also tend to be organized by region, sometimes with a keyword search option. Keyword search is not entirely useful even if the maps have been cataloged because users do not know what words were used in the map or its catalog record, and so relevant items are missed. Maps of businesses are available through geographic search, also called local search, implemented by major search engines such as Google, Yahoo, MSN Live, Ask, and America Online.[3] Their map tools link businesses to maps by means of map-and-hyperlink architecture (Tezuka, Kurashima & Tanaka, 2006), with maps created on the fly, according

[2] Summary discussion of the relationship of these map types and data distribution is found in Velasco and Boba (2000).

[3] Http://local.google.com, http://maps.yahoo.com, http://maps.live.com, http://local.ask.com, http://localsearch.aol.com

to the user query. The theme of most of these maps would be business. Local search provides a sort of geographic realization of the yellow pages.

> [T]his abundance of readily accessible [geographic] data presents a problem: how does the user know which collection to search for a given dataset? … Knowing where to look is still largely a matter of personal knowledge and luck (Longely, Goodchild, Maguire & Rhind, 2001, 157).

Those in geographic professions might or might not know where to look.

> Search [for geographic datasets] is then a two-stage process, requiring first an identification of an appropriate collection, and then a search within that collection for the needed data set (Goodchild & Zhou, 2003, p. 96). … The typical experienced GIS user possesses an enormous amount of information of this type, and *shares it or guards it* as the case may be (ibid: p. 103, [italics added]).

According to Goodchild and Zhou, this raw material is so valuable to those in the field that knowing where to look can be a competitive advantage so that some might even keep it secret.

2.3 Why finding a digital map can be a problem: spatial metadata

The metadata situation—lack of collection-level metadata, lack of a simple standard, and relative silence on the topic in the literature—contributes to the difficulty of finding maps.

Metadata can be defined as structured descriptions for physical or digital objects that may be used for information retrieval. Metadata describing spatial data quality can be helpful to users trying to decide whether to consult a particular data object. Goodchild (2008) describes aspects of potential spatial data inaccuracy.[4] The difficulty lies in getting inaccuracy measurements to add to the metadata. Either the cataloger must take the time to analyze data accuracy—which is inefficient—or the data provider would need to be forthcoming about data inaccuracies—which is unlikely. It would be more efficient to create a system that could analyze spatial data quality and code the results into metadata automatically. Before

[4] Some aspects of spatial data quality mentioned are decoupling (variability in how map distances correspond to real world distances), uncertainty of measurement, disparities among measurements (correct elevation with incorrect latitude, for example), disparities in measurement scale, autocorrelation (objects nearby on a plane tend to be more similar than objects that are distant), and cross-correlation (object layers may be mis-aligned).

such a system should be created, however, researchers must agree upon standard units to measure each aspect.

Widespread lack of collection-level metadata explains why knowledge of map websites can be the "guarded secret" mentioned above. Overviews of geolibrary contents are rarely available, so that a searcher must enter each site individually to determine whether it contains needed material. Metasearch in which a single query could be put to many geolibraries simultaneously is not currently available. Even metadata for a set of maps would help. Archives typically catalog documents in sets. In the same way, maps could be cataloged in sets with a single parent record, although this not how it is done currently.

Even if it were possible to search across geolibraries technically, the heterogeneity of map cataloging would present difficulties for information retrieval. This heterogeneity is surprising in that over 130 countries of the Organization have agreed upon the International Standards Organization 19115 spatial metadata standard. However, many countries use their own standards in addition (Moellering, 2005). The U.S. Federal Geographic Data Committee, for example, recommends the Content Standard for Digital Geospatial Metadata (CSDGM).

Many large private and government organizations mandate proper documentation (Albrecht, 2007, 13). But cataloging using most of these standards is time-consuming. Spatial metadata standard ISO 19115, for example, defines over 300 metadata elements. It was decided by ISO Technical Committee 211 that the minimum number of elements required is 60 (Peng & Tsou, 2003). Compare this to Dublin Core, a non-spatial metadata standard that requires only 15 elements. A 12-element Denver Core was suggested as a minimum field set for the U.S. Content Standard for Digital Geospatial Metadata. But while Denver Core and other minimal level cataloging schemes for maps have been proposed (Ercegovac, 1998), none has been widely adopted.

Ideally "each data object should be able to automatically generate its own metadata and encapsulate it into the data object in the process" (Peng & Tsou, 2003, p. 291). This is the data mining aspect of the dissertation.

What information should be in a spatial metadata record in order to enhance its retrieval value for searcher? This question is rarely considered. But it is essential to know what searchers look for in order to know what map aspects are important enough to take the time to index. How catalogers should describe maps to improve information retrieval is described by Larsgaard (2005) and Buckland et al. (2007). Little discussion of the information retrieval properties of spatial metadata is offered by Moellering (2005) or Peng and Tsou (2003). Somewhat more attention to the information retrieval properties of spatial metadata is provided by Kuhn (2005) and Nogueras-Iso, Zarazaga-Soria, and Muro-Medrano (2005). It is likely there is a connection between the relative silence in the literature on the information retrieval value of this metadata and the difficulties in retrieval of maps.

2.4 Theme maps and those who use them

"Thematic map" (Burrough & McDonnell, 2000, p. 2) is widely applied to maps showing a general purpose theme such as soil depth or composition or aridity, as well as to maps showing the distribution of a non-spatial attribute. The basemap is a vehicle to display theme data. Some property combined with location or the geo-atom has been considered the smallest discrete unit to which geographic data can be reduced (Goodchild, Yuan & Cova, 2007, p. 243). Search for map by theme has been called conceptual search (Kammersell & Dean, 2007).

Different kinds of digital theme maps are found in

(1) Web analogs of geographic and non-geographic print publications
(2) Websites
(3) Mash-ups, the original term for maps made by combining data from different sources
(4) Scans of printed and aerial sheet maps in geolibraries

These different kinds of theme maps are found in different kinds of collections. Articles with maps may sometimes be discovered by searching for "map" in the physical description field of a proprietary database such as Wilson OmniFile or Humanities Full Text. Maps associated with websites are often retrievable by including the query term "map" in addition to the theme and location in Google Image search. Google Earth maps can be uncovered by filtering for the Google Earth proprietary file format, .kml or its compressed version, .kmz, in the Advanced section of the Google interface. Mashups are featured in dedicated websites such as Geocommons. Each geolibrary with digital maps such as the Alexandria Digital Library must be entered individually as no metasearch across sites is currently possible.

Professionals in many fields generate or employ theme maps in research. Epidemiologists use maps to track diseases; historians track political boundaries; economists, demographics; businessmen follow franchises; archaeologists plot tells; and linguists look for dialect boundaries.

A profile of a typical map user by Marley describes a person looking for themes, not only location (Marley, 2001, pp. 12-15). Marley's solutions as to how the user should find a relevant map centers on the map librarian, or better organization of physical map collections (ibid: 24-25), or else the user should make the map himself (ibid: 15). Granted this text was published in 2001, but even so, there is no vision of a do-it-yourself search engine for maps.

2.5 MapSearch in response to present problems

Hundreds of thousands of geography-related questions are put to major search engines annually, as pointed out in chapter 1. Many of these questions might well be answered by theme maps. Looking for a theme map in a geolibrary is complicated by the fact that some geolibraries do not index maps by theme, and further, by the fact that there is no central map index. This research aims to answer geography professor Peterson's suggestion that what is needed is a search engine designed specifically for maps that would categorize maps by content (Peterson, 2007, p. 132).

2.6 Approach

The dissertation revolves around the question of how is it possible to make it easier for people to find maps. Related research shows others' answers to the problem. The beginning of the research considers how people ask for maps, which in turn, suggests what metadata should be indexed. Next it considers how that metadata may be obtained automatically, and outlines the method for how that metadata may be indexed automatically. Results of automatic classification on the training set are discussed, along with the limitations of the indexing algorithms. A web model, or prototype, was created to demonstrate visually how the indexing works, and the next section presents decisions behind the layout of controls on the model—the interface design. The final evaluation confirms statistically the findings demonstrated visually with the training set through the interface, i.e., that the system is able to index maps automatically, and it includes evaluations of the interface design and of other system aspects as well.

This dissertation proposes a method for the unsupervised, or automatic classification of maps in published articles or documents in .pdf format based on region, time and subject. The following will be studied:

- User queries
    - How do people ask for maps? (chapter 4)
- Data discovery
    - Which words in and around the map should be harvested for purposes of indexing? (chapter 5 and Appendix B)
- Clustering
    - What indexing facets will reflect how people ask for maps? (chapter 6)
    - Is it possible to create algorithms for the unsupervised classification of maps by region, time and subject? (chapter 6)
- Result ranking
    - What measures should be used for ranking all the maps that match a query (chapter 6)?
- Categories for classification
    - What categories should be created to subdivide the facets? (chapter 6)

Interface

How should an interface be designed to allow retrieval by keyword and category facet? (chapter 7)

Evaluation

How well do the classification algorithms work in comparison to manual classification? (chapter 8)

3 Related research

This dissertation rests on almost five decades of work in information retrieval as well as on more recent work in geographic information systems. In this chapter, others' research related to the dissertation is described under the headings of information retrieval, automated classification and clustering, natural language processing and lexical search, determining relevance, indexing attributes for maps, and systems combining some or all of these principles to retrieve maps.

3.1 Information retrieval

An overview of the history of information retrieval is provided by Van Rijsbergen (1979), and an introduction to the state of the art by Manning, Raghavan, and Schütz (2007). Milestones include papers by Cleverdon along with Mills and Keen (1966), Salton (1968), Salton and McGill (1983) and Lesk (1986). International Text Retrieval Evaluation Conferences (TREC), Conferences of the Special Interest Group on Information Retrieval (SIGIR), and the Joint Conference on Digital Libraries (JCDL) are held regularly to exchange research advances.

Let a few basics about information retrieval suffice in light of the solid overviews mentioned above. Digital information storage and retrieval was envisioned after the Second World War by Vannevar Bush (1945). An example of such a system was realized just a few decades later in databases in which a user entered a word, and the system returned all documents containing that word. A *keyword* or *query term* may be requested by itself or in a set. Query terms may be combined using set theory (A and B), (A or B), (A and not B) after *Boolean* mathematics. To this day, the "and, or, not" seems still to challenge searchers. The Google search box disposes of it entirely, and the Google Advanced screen clarifies "and" as "with **all** of the words," "or" as "with **at least one** of the words," and "not" as "**without** the words."

Search within the full text of a document is a form of *uncontrolled* indexing. The problem with uncontrolled indexing is that one word can have many meanings and retrieve a document that is irrelevant to the query. This lowers *recall*. Instead, terms in either the query or the document can be *controlled* to eliminate synonymy. The debate about the relative merits of controlled versus uncontrolled terms for information retrieval is not resolved (Peters & Kurth, 1991; Shaw, 1993). Justification for the use of automated indexing rests to some degree on the assumption that the occurrence of particular words is related to their importance in the document.[5]

Indexing control can be implemented manually by indexers using a controlled vocabulary list, or automatically via thesaurus or ontology. It is not only the consistency of the application of terms that influences results; the nature of the controlled vocabulary matters too. The extent of index vocabulary is *exhaustivity*. The degree of precision of the vocabulary is *specificity*. Raising exhaustivity and specificity tends to improve retrieval (Salton & McGill, 1983, p.55). It has been suggested that a combination of automatic and manual indexing results in optimum representation of a document's contents (Sensuse, 2004). Manual indexing, of course, is almost impossible with web-scale data sets.

The user decides whether the items retrieved by the query are relevant. The set of the items retrieved is judged on the basis of recall and precision with respect to the total set in the database. Recall is defined as number of relevant items retrieved divided by the number of items in the collection. Precision is defined as the number of relevant items retrieved divided by the total number of items retrieved. Recall and precision have an inverse relationship, so that improving one jeopardizes the other.

When information retrieval systems scale from test size to full size, recall at least for full text systems has been shown to suffer. It suffers because users to do know exactly what words and phrases appear in the items (Blair & Maron, 1985, p. 295). Full text retrieval for this reason has been supplemented with manually-assigned indexing terms (ibid: 298). Manually created ontologies, even though not manually assigned, help retrieval as well.

[5] This occurrence characteristic of words in documents has been described by Zipf's law, in which Frequency x Rank is approximately equal to a constant (Salton & McGill, 1983, p. 60).

3.2 Data mining and automatic classification

Data mining is the process of extracting particular sorts of data. Required reading is considered to be the 1993 paper by Agrawal, Imielinski, and Swami. The first application is considered to be the mining of a customer database from a British department store to find trends that could help the store increase its sales.

The dissertation task of classifying maps begins with crawling, that is examining automatically, a large data set to extract the maps and harvest the associated metadata. Harvested metadata is aggregated in advance of database querying. Simeoni, Yakici, Neely and Crestiani review the benefits and costs of metadata harvesting (2008, p. 14). The main benefits are that it speeds query processing and supports scalability, as well as improving the potential for interoperability. Some disadvantages are that it might not necessarily be cost-effective to aggregate resources, aggregating in a computer domain can be less fault tolerant that searching peer to peer, and the metadata copied from originals might have become inaccurate if the originals were updated.

Knowledge discovery uses algorithms to transform data mining results into understandable information (Wright, 1998). Large amounts of data are required. Approaches may sort data statistically, visually, or semantically into classification categories, for example, in order to turn data into knowledge.

The unsupervised or automatic classification of items into groups is also called *clustering.* Jain, Murty and Flynn (1999) and Oberhauser (2005) review this literature. Automatic classification is generally done in one of two ways. In machine learning, a system is given a large body of already-classified items and via “supervised learning” induces how a not-yet-seen item would be classified. Alternatively, the researcher will look at a large number of items, induce general rules governing how those items are classified, and then feed those rules into a system. The items on which the rules are made are called the *training set*. The items on which those rules are then tested are the *test set*, or evaluation set, or validation set.

It is possible to classify from metadata to categories directly, without a controlled vocabulary or ontology. An ontology is a classification that includes interrelationships among its terms. Oberhauser does not even mention ontologies in his discussion of automatic classification methods (2005, chapter 2). Even so, ontologies have been found to improve information retrieval, and techniques are discussed in section 3.3.

Yao, Etzkorn and Virani created a system to retrieve software components using a mediating onotology (2008). They call the ontology-fortified system alternately smart search, semantic search, or knowledge-based search (p. 614, 613). They use OWL (Web Ontology Language) and add weights from which, using an algorithm borrowed from other researchers, they compute similarity between nodes. The method they use to validate classification results is similar that used in this dissertation. Experts classify items, and then the system classifies the same items, with the system results compared to the manual results that are assumed to be the most accurate.

3.3 Lexical search using an ontology and weighted indexing

Natural language processing concerns how a system finds meaning in words. It has been considered at several different levels (Salton & McGill, 1983, p. 259-260). Most systems retrieve on a lower level basis of morphological traits by removing prefixes and suffixes and generating word stems for matching. This is the basis of the dominant vector space model of information retrieval. More refinement in retrieval results (that is, a better match for the query) is obtainable with lexical matching, in which metadata words are enhanced by dictionary or thesaurus classes.

Any controlled vocabulary, whether classification, thesaurus, taxonomy or ontology, may be used to improve information retrieval. It can be used at the back end of a system without user awareness either for classification of the metadata (Salton & McGill, 1983), or in what has been called blind feedback (also called pseudo-relevance feedback) (Leveling, 2007), or to expand the query term (Shiri & Crawford, 2006).

The breadth of subject coverage distinguishes an entire classification or ontology from a limited domain thesaurus, taxonomy or domain ontology. Because a thesaurus is limited in scope, the boundaries of what to include become important. Its terms can be collected from the top down by building upon an existing taxonomy or from the bottom up by extracting terms from a document set, or by some hybrid approach that combines top down and bottom up. Terms comprising the thesaurus or ontology usually are unequal in their role as category indicators. For this reason, weights are assigned to emphasize better indicators. The literature is reviewed and the value of different weighting schemes for information retrieval is discussed in Wolfram and Zhang (2008).

An example of how a thesaurus works is found in (Salton & McGill, 1983, p. 76). Thesauri may be constructed manually, semi-automatically or fully-automatically. Spärck Jones (2005) discusses building a thesaurus automatically. Ideally, a thesaurus is a living document that is maintained along with the data set with the addition of terms, and increasing granularity along with data set growth as needed, with the removal of terms that have become outdated.

A commonly used general-purpose controlled vocabulary for information retrieval is the ontology called WordNet. It was started in the mid 1980s by a Princeton psychologist, Dr. Miller, and is presently up to version 3.0. Its word groupings in synonym sets, or synsets, can be illogical. “Kitten” does not share a synset with “cat,” for instance. WordNet’s drawbacks for information retrieval are outlined by Gabrilovich and Markovitch (2007, 2329-2330).

Automated indexing by subject has been based on the Library of Congress Classification, which is a system to classify books. Its controlled vocabulary includes relationships among categories that seem logical, and some have used it as a tool for information retrieval. Automatic classification using the Library of Congress Classification has been studied by Larson (1992) for bibliographic records; and by Wang, Hodges and Tang (2003) and by Prabowo, Jackson, Burden and Knoell (2002) for web pages. Larson (1992, p. 146) using the Library of Congress Classification and Wang and Lee using the Dewey Decimal

Classification (2007, p.138) found independently that it may not be possible to use either system fully for classification automatically. Description of each experiment is not detailed enough to assess fully the source of the failing. Kim and Lee (2002, p. 494) using a related classification system for books called the Colon Classification were able to classify 81% of over 2500 book titles automatically.

Automated indexing by region has relied on Library of Congress Subject Headings and Geographic Name Authorities for books, as well as certain gazetteers. A gazetteer is a directory of place names, coordinates, and how places relate to one another hierarchically. The task of determining which place is meant by an address or a region is known as *grounding* or *localization*, or, to use a term coined by Leidner in 2004, *toponym resolution* (Leidner, 2007, p.34). The difficulty in indexing by place is that many places throughout the world have the same name. Rather than confront the disambiguation problem, many researchers prefer to create their own non-ambiguous thesaurus, as did Leveling (2007). How to build a sound geo-ontology is the research of Abdelmoty, Smart and Jones (2007). The ontology must expand in order to stay current. Smart, Abdelmoty, El Geresy, and Jones (2007) discuss how to combine a geo-ontology with rules for self-maintenance.

3.4 Determining relevance: semantic similarity

The goal of the ranked list is to return to the user the most relevant documents at the list top. The way to accomplish this is still unclear. “Comparing two objects relatively is still one of the biggest challenges and it now concerns a wide variety of areas in computer science…” (El Sayed, Hacid, Zighed, 2007b, p. 49). It requires determining not only what matches exactly, and what matches approximately, but measuring the degree of approximation. Possibly because what makes a document appear relevant itself is not wholly understood (Saracevic, 2006), translating relevance into algorithms is necessarily complex. Roddick, Hornsby and de Vries (2003, p.112) offer ways in which distance is understood in terms of data sets that meet or overlap.

Yager and Rybalov (1998) review the literature as to various similarity measures, and El Sayed, Hacid, Zighed offer a state-of-the-art review (2007). Varelas et al. (2005, p. 11) outline four methods for using an ontology to determine similarity between terms: *Edge counting,* where similarity is measured as function of length of path between terms in the ontology; *Information content* where similarity is measured in the probability of the occurrence of the term in the database; *Feature-based* where similarity is measured as a co-occurrence of terms with similar meaning in the same synset (Chua & Kulathuramaiyer, 2004) and *Hybrid,* some combination of these.[6]

Rules for defining similarity for geographic information retrieval could be based on Euclidean distance between places, as can be calculated based on their coordinates, or a words only approach known as geosemantic proximity (Brodeur, Bédard & Moulin, 2005). If topological measures are used for similarity, measures include region that are the same or not the same, but also regions that might intersect, touch, overlap, cross, lie within one another, or contain one another, as well as lie at minimum distance (Hill, 2006, p.191). The advantage of semantic relevance is that it would make spatial information retrieval interoperable with retrieval of other document types (Kuhn, 2005). Heuer and Dupke designed their own means of calculating spatial relevance for geotags (2007, p.202). Implementation of a GEO server—GEO profile of the Z39.50 standard—adds elements from the FGDC Content Standard for Digital Geospatial Metadata: overlaps, fully enclosed within, encloses, fully outside of, near (Hill, 2006, p.196-7). Results could be ranked according to geographic relevance (so that places physically near the location of the query location would list first) or semantic relevance.

3.5 Indexing attributes for maps

Metadata fields are said to be *indexed* when that data can be sorted independently. For example, a subject field is indexed to allow subjects to be sorted alphabetically and cross-referenced. A page number field is not indexed, so that a user could do a keyword search for "250 pp." and retrieve all the books in the database with 250 pages, but there is no way in this non-indexed field to list books in order of page length. Usually fields are indexed on essential aspects of the item, or the aspects that people use to search.

[6] Programs automatically implementing similarity measures for the ontology WordNet are available freely on the web (http://search.cpan.org/dist/WordNet-Similarity).

How should maps be indexed? Geographic information reduced to a primitive form has been termed the geo-atom (Goodchild, Yuan & Cova, 2007, p. 243). The geo-atom pair includes location and some attribute. Theoretically, that attribute could represent time or subject. Or perhaps it could represent both time and subject.

A time attribute seems critical. The topography of a place changes over eras; political boundaries change over centuries. The Alexandria digital library mentioned above includes a search by time period. The very aspects of topography or political boundaries suggest also a search by subject. Indexing by subject is one of the most basic in the history of library science.[7] The MARCO system described in section 3.6 indexes maps by subject (Samet & Soffer, 1996).

Perry, Hakimpour and Sheth propose that space, time and theme should be considered as retrieval elements for a basic web search system (2006). Kemp, Tan and Whalley call these three the "space-time-theme composite" (2007, p.84). If these three aspects could be used for referencing all types of information, as is argued by Hill (2006, p.183), they seem a solid foundation for indexing maps.

3.6 Systems to retrieve maps

Some systems use a map for an interface but do not index maps. The TOXMAP system for environmental health resources has a map interface (Hochstein, 2006), for example, as does the Perseus digital library specializing in Greek and Latin classical literature. MetaCarta sells a product called Geographic Text Search that identifies implied or explicit references to geographic locations within documents, assigns latitude/longitude coordinates to the references, indexes the document, and then enables a search for indexed documents through a map interface.

[7] The other two are author and title. These three were suggested in Charles Cutter's "Rules for a dictionary catalog," in M. Carpenter and E. Svenonius, Eds. *Foundations of Cataloging: A Sourcebook*, Littleton, CO: Libraries Unlimited, 1985 pp.65-71.

An information retrieval system notable for its map interface and ontology-supported retrieval was created by Kemp, Tan, and Whalley (2007). Their system retrieves marine environmental data. Their facets are based on region, time, and theme. Instead of the ontology being hidden from the user, it is set in a browse interface to particularize the query and identify the level of hierarchy required. The documents in their system are accompanied by metadata. Ontologies selected are specific to the documents in the collection which, the authors admit, will hinder collection expansion.

A map librarian is the most "expert system". Map librarians are employed, among other things, to help patrons find what they are looking for. Finding maps without help can be challenging, since many map libraries are organized by region, but patrons often seek maps by time and theme as well, as discussed in chapter 4.

The online access systems for physical map collections often reflect the physical organization by region. Such is the case, for example, in the map retrieval system of Princeton University (Shawa, 2006). The system was built using off-the-shelf software and the maps were hand-cataloged. Searching by region may be a semantic or drop-down menu option to name the state or city sought, or a visual option to size a box to the footprint sought on a map, or a precision option of entering coordinates as in the Alexandria geolibrary, the 3D Grifinor project for geographic visualization, or the STEWARD system to query documents for location (Lieberman, Samet, Sankaranarayanan, & Sperling, 2007).[8]

Conceptual, or theme search option for maps is rare. MARCO (Map Retrieval by Content) was devised by Samet and Soffer (1996) to determine map themes. It separates maps into layers and automatically indexes both geographic and content layers.[9] To do this, each map is interpreted upon entry into the database. The approach is entirely graphic. "The user identifies those symbols in the legend that are relevant for their <sic> application, and the acquisition process extracts this information from the raster scanned maps" (p. 784). As

[8] Alexandria is at http://www.alexandria.ucsb.edu and Grifinor is at http://www.grifinor.net

[9] Scans of printed maps show attributes as part of the picture, but maps in native digital form may store the attribute data in a layer separate from the geographic base.

shown by a search on the authors' names in the ACM Digital Library, the paper has been cited for its information retrieval content, and specifically for its indexing of symbolic images, but apparently not in the context of this particular application to find maps on the basis of theme.

In contrast to the MARCO system that employs *graphics* to retrieve maps by subject, another information system employs *metadata* to retrieve maps by subject. A test collection was created from 1000 records describing geospatial data from Geoscience Data Catalog at the U.S. Geological Survey (Nogureras-Iso, Zarazaga-Soria, & Muro-Medrano, 2005, section 4.4). The authors maintain that their results should generalize to digital library retrieval experiments on subject attribute data, similar to what is being considered here. In preparation for the information retrieval, the data was disambiguated against a general-purpose thesaurus (WordNet) to add synonyms. The search system implements standard library catalog Z39.50 information retrieval protocol, with relevance ranking on the classic vector-space retrieval model that has proven reliable in many contexts. The metadata records were hand-tagged to create broader/narrower term hierarchies. Thesaurus-smoothing of the index allowed queries to contain words not found in the metadata. Results showed a precision—recall curve comparable to what would be expected in a typical text retrieval system. However, the authors found that the thesaurus was of limited use because it was not domain-specific. By contrast, MapSearch created in this dissertation extracts metadata automatically and uses this metadata to retrieve maps by subject and time period as well as region.

The National Geologic Map database allows for retrieval of catalog records for maps, and some digital maps, by geologic theme as well as area (using a visual map interface), scale, and format.[10] Additional options are whether the map includes GIS data and is available for download or purchase. The precision of digital maps in this system is uncharacteristic of maps in published articles. Yet, the National Geologic Map database represents the sort of system that could be a partner for MapSearch, were the different map collections to be searchable at once, and the information fusion problem surmounted.

[10] http://ngmdb.usgs.gov/ngmdb/ngm_compsearch.html

4 User defined facets

"IR does not have well-developed methodologies for working with [user] data or producing reproducible research from it, so this type of research tends to be neglected." (from *An Information Retrieval Research Agenda,* Callan, Allan, Clarke, Dumais, Evans, Sanderson, Zhai, December 2007, p.30) This chapter attempts to end this neglect by a performing a qualitative analysis of user data in two case studies. For the purpose of this dissertation, preliminary data was collected from the Map Division of the New York Public Library and from the Internet Public Library, and then analyzed to find patterns in how people ask for maps. The data is used as a further guide in the indexing of maps and in adding features to the interface to facilitate customary searches.

4.1 How do people ask for maps?

The literature review in the previous chapter established the attributes of location, time and theme that are important for map indexing. This section presents two studies with data on how people actually ask for maps by way of confirmation of those attributes. Rather than attempt a large-scale ethnographic study, the opinion of experts in library reference services was sought. The resulting two studies balance each other. The questions recurred over decades for one study, and for the other, they were collected over a several month period.

4.2 Location, time and theme queries in a map context

Location, time, and theme are used as codes to show patterns in queries for maps. But what do these codes mean in the language generally, and in the context of how people ask for maps?

<u>Location</u>

A place word denotes some bounded region. The person who asks for a map of some location may intend the map to show that location only or that location surrounded by – but probably not dwarfed by – neighboring locations. For example, most maps of New Zealand probably will include a point for the capital Wellington, but a person who asks for a map of Wellington almost certainly wants a large scale map of the city rather than a map of the country. *In essence, a query for a location is key to the desired map scale.*

Time

Maps usually show land. Because in human time, land is stationary, what is implied by a map date it the date of the land in the [cartographer's] present. The person who asks for a map of a village in Eastern Europe in 1910 probably wants a map that was made around 1910. How should a map be dated that shows Neanderthal burial sites? Should the map be dated to today because it shows sites of modern excavations, or should it be dated as prehistory because it concerns Neanderthals? *In our context, the time period of the map is implied by the map theme,* and our Neanderthal map would be categorized as prehistoric.

Theme

All maps are tools of geography, but geography is not the theme of every map. We can tell the theme either from the map itself or from its context. Two maps that look the same might have different themes if they were created to illustrate different ideas. For example, the same map of New Zealand map with a label on Wellington could have a theme of government in a report on international relations, or a theme of homeland security in military brochure. *The person who looks for a map on a particular theme, therefore, might be satisfied either with a layer of data above the geographic base that pertains to that theme or with a map illustrating a discursus on that theme.*

4.1.1 Case study at the New York Public Library Map Division

Method

The Map Division of the New York Public Library has an international collection dating from the 16th century to the present. It holds almost half a million maps, together with thousands of atlases and books on cartography.

A librarian who was known to the author was asked to provide a list of the questions she encounters most frequently. Nancy Kandoian drew upon her almost 30 years of experience at the Map Division to furnish a list of patron she has encountered most frequently (personal communication, December 12, 2007). The questions appear below in no particular order.

Codes were established on the basis of the established indexing categories of region, time and theme. To assign codes to questions, questions were read for outstanding characteristics. The coded questions appear below.

Data: recurring questions	Code
I need a Sanborn map [current, for NYC address].	{theme+region}
I'd like to see a map of ... [some area or place, e.g. a region in Pakistan, France, Tierra del Fuego, Nevada].	{region}
I need a Sanborn map from before 1961 [usually for a Queens or Brooklyn address, relating to a grandfathered zoning issue].	{theme+time+region}
I'd like to see a series of fire insurance/real estate maps through time for a site or neighborhood in NYC.	{theme+time+region}
I'd like to see a series of fire insurance/real estate maps through time for a site or neighborhood [outside NYC, most often NJ].	{theme+time+region}
I'd like to see a series of maps through time that show the development of [such and such] an area. (May be city plans or other region maps; may relate to wetlands or other particular kinds of natural or man-made changes such as shoreline or roads.)	
I'm trying to locate a place/village in [Central/Eastern Europe, Ireland, Italy] from [some time usually before 1920].	{time+region}
Can you help me find some information about an old map that I have?	{region}
Can you help me locate this [place/address/cemetery/church]?	{theme+region}
I'd like to find out what was previously on [such and such a site].	{theme+time+region}
I need a site plan [or large scale contour map] for a site in [a particular city].	{theme+region}
I need to see some maps to get some design ideas for [such and such a project; may be related to cartography, publishing, textiles, theater, whatever].	{not coded}

Findings

One of the twelve questions was not coded because it does not relate to any specific region, time or theme. Of the eleven remaining: three concern region + theme and either assume the present day or are not time specific, five concern region + time + theme, one concerns region + time, and two region only.

Nine out of the coded eleven, or 82%, concern theme and time as well as region and could not be answered easily by existing geo-based search systems. Therefore, theme and time and region facets are recommended for a search interface.

4.1.2 Actual questions from the Internet Public Library

Method

The Internet Public Library provides resources and reference services. It serves children largely, although not exclusively. The following questions were drawn via keyword search on "map" and "geography" from over 60,000 of those which patrons emailed to staff at the Internet Public Library during 2006 and 2007. They were reported courtesy of Ms. Crosby, the Assistant Director for User Services of the Internet Public Library (personal communication, 2007).[11]

The same codes were used for this study: region, time and theme. Outstanding characteristics of questions merited a code. Mark how detailed are the requests for dates and themes as well as places. The data appear below.

Data

The questions below, like those in the Map Division study, are listed in no particular order and are coded for region, time and theme.

[11] Our formal request to examine query logs from the Microsoft live search engine was rejected in December 2007.

Where can I find a detailed, yet easy to read world map of mountains, rivers, lakes seas, deserts – that I can either find online or at the library that I can check out?
{theme+region}

I am looking for images of maps of Detroit, MI. I specifically need three maps, one from before 1943, one from between 1943 - 1967 and one from after 1967. The closer they are to those dates, the better they will be (for example, a map from 1886 won't work, but one from 1930 will work). {time+region}

I really don't know anything about … the city is called Ramnicu Valcea (or Rimnicu Vilcea) which is located in southern Romania in Wallachia. I've searched on the web and I can't find any information that I want. I've already tried the local libraries near and at my community but there is no information at all. Mainly what I've been looking for are detailed maps of the city in the year 1845, 1988 and recently, showing the name streets and showing a little outside the city as well. I would also like more knowledge on the history, and culture of the people. Pictures of the city has [sic] also been a difficult thing to find, I would very much like to see what the city looks like: streets, shops, etc. Please, any information at all and that I've requested would help.
{theme+time+region}

I am a fifth grader, and am in the middle of a state report. I am doing New Jersey and I need a good place to find printouts of a topographcal [sic] map, a political map, a points of interest map, and the New Jersey state license plate.
{theme+region}

How to find maps of Maine, MA and Halifax, Nova Scotia Canada dating back to the 1800's {theme+region}

How might I get access to walking maps of Paros, Naxos and Crete, islands of Greece, electronically or in print? {theme+region}

Hi. I want to write a triology [sic] story based on fantasy. So I was wondering whether you can find me a map of a place, a historic place, where I can base my story in? And also, could you email th [sic] map to me? By the way, I'm 13 and I want to take a carr as a childrens' book writer. {time+region}

What is the elevation of Crowley's Ridge, Ar, where it is intersected by the pioneer (circa 1865) Southwest Trail (aka Military Road) between Memphis, Tn and Fulton,

Ar? Any description of terrain, flora and fauna would also be helpful. I'm planning to describe how a Conestega wagon would cross at this point.
{theme+region}

What river in Massachusetts is at the highest elevation?
{theme+region}

I need to know the Physical setting of Visalia, CA elevation, physical features how the physical setting affects the weather, climat [sic], Native flora, Native fauma [sic]?
{theme+region}

Findings

All 10 questions fit within the three codes of region, time and theme. Seven represent region + theme, two represent region + time, and one represents region + time + theme.

4.2 Interpretation

The case studies of New York Public Library and Internet Public Library map queries suggest that map indexing should include region, time and theme. That map indexing can be reduced to a few basic points of indexing seems to contradict the complex metadata schemes that have been set up for map cataloging. While it is undisputedly useful to know the source of each map, its scale, date of creation, and so on, such as are provided in many map cataloging schemes, these are not necessarily critical to "the user" in defining a search. More research is called for in defining how particular user communities look for maps, with the proviso that some quite useful information is not readily available through indexing that is automatic.

What do the patterns of coding among recurring Map Division and actual map-related queries show for MapSearch? (1) How best to index the maps based on facets employed to search, and (2) how to set up an interface to facilitate actual search patterns.   Later sections of the dissertation therefore concern how to index maps (chapter 6 on Classification algorithms), and how to design an interface to facilitate searching (chapter 7 on Interface design), with the following chapter concerning how to carry out the indexing based on products of data mining so that it is unnecessary to catalog by hand.

5 Overview of automatic cataloging and classification of maps

5.1 System architecture

The introduction suggests that many people would probably find useful a system that gives one-stop access for many types of maps—oversize sheet maps in library or antiquarian collections cataloged manually, driving maps that are generated on demand, maps that are mashed up by enthusiasts, maps that illustrate educational or travel websites, and maps inside books and articles. This chapter describes a system that potentially could accommodate all these map types—once the fusing search across different maps is resolved. The system described here initiates research by focusing access to maps in published articles.

Fig. 2 below depicts the basic functions of the MapSearch engine as it was designed for this project. The system will extract maps along with words that are associated with the map from published articles. These words are the metadata that are used to assign maps to categories with the help of an ontology. When the user enters a browse category, the system compares the category to target terms in the metadata. Results are displayed according to how similar the target terms are to the query. Each stage of this process is described below.

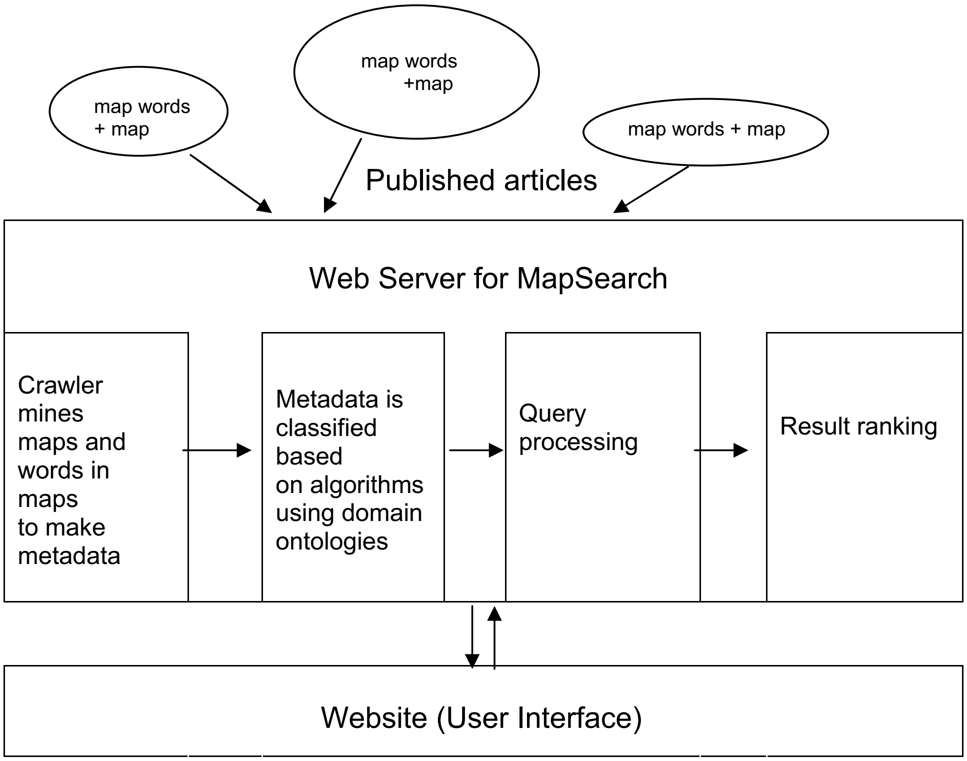


Fig. 1 The architecture of the proposed map retrieval system.

5.2 Harvesting metadata

5.2.1 Collecting the sample

The 150-map training set was collected article by article. Journals from a wide range of disciplines were viewed to see whether their articles included maps. The training set was collected with an aim to diversify subjects. Yet, the subjects were not balanced among the disciplines before the articles were selected, so the method cannot be considered stratified sampling. Most articles were in Adobe's native .pdf format. A few were from web pages and were converted to .pdf.

Articles came from databases including Asia Studies Full Text, Humanities Full Text, Wilson OmniFile, JSTOR, the Homeland Security Database, the New York Times full text, as well as a web articles and open access journals such as *Ecology and Society*. Articles in some of the proprietary databases had been cataloged. The MARC 300 field for physical description notes whether an article includes a map, but cataloging rules do not require that any details about the map be provided in the physical description field. Even in articles that were cataloged, therefore, it was necessary to inspect each map individually. Most articles were not cataloged at all and those containing maps were discovered by serendipity or by a keyword search on the word "map".

5.2.2 Creating and storing

Maps were clipped manually from articles using the Adobe snapshot tool and, using Microsoft Paint, stored in .jpg compression format. Later the differentiation between text and map, and map extraction, will be done automatically. It is necessary to know which articles are only text, which have photographs, charts and other illustrations, and which have maps. The procedure is straightforward. An Optical Character Recognition (OCR) program opens every document and scans each page. Preliminary comparison of the open source OCR program Tesseract with ABBYY's FineReader and LizardTech's Document Express showed that Document Express was best for this purpose.

Programs to distinguish maps from other graphics are being developed independently by Prof. Lesk at Rutgers and Prof. Giles at Penn State University. The program would decide it had found a map based on border line irregularity and color variety, as opposed to the straight lines and color blocks that are found in tables, charts, cartoon, pictures and other sorts of diagrams. A method has been developed to separate a map into constituent layers with the intention of analyzing the symbols (Dhar & Chanda, 2006). Professors Lesk and Giles also are developing independently layer extraction programs that will separate map into basemap and text. A layer extraction program would allow the automatic generation of a very large map sample along with “words in map” metadata to re-test the algorithms.

Metadata was taken manually for maps in the training set, such that principles could be derived for the automatic generation of metadata as set out in Appendix B. It was found that such information is found in words that can be scanned in the map (please see Figure 2a) as well as the map caption, the article or book title, and references to the map that appear in the text (Figure 2b). Principles underlying automatic production of metadata as set out in Appendix B were derived from generation of metadata manually for the 150 map sample.

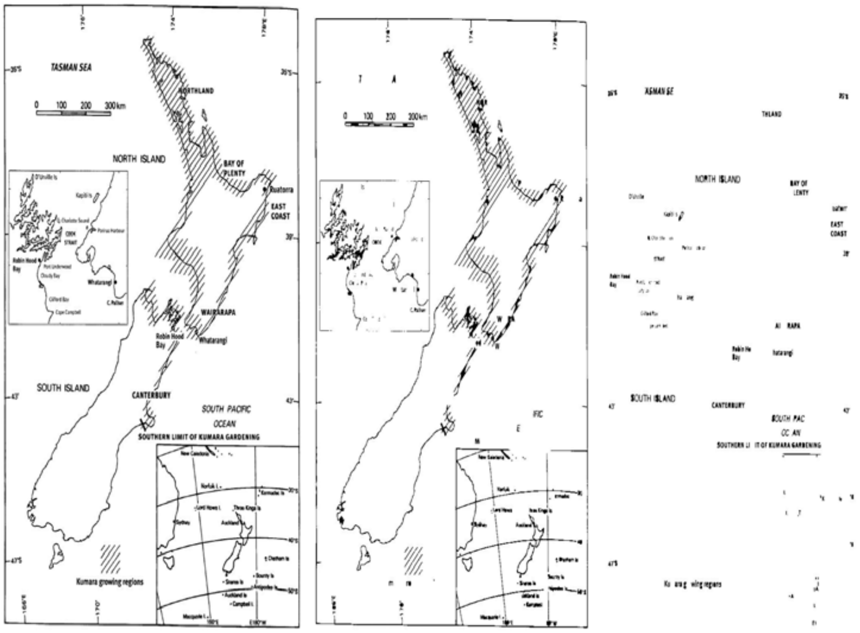


Fig. 2a Left: Map as it appears in the article; Center and Right: Demonstration of the text layer extraction program that produces the words in map used as metadata. The only words-in-map that

are clear enough for system use are *North Island, South Island, Canterbury and Southern L[im]it of Kumara Gardening*.

Caption: Fig. 2. Map of New Zealand

Title: Experimental Archaeology Gardens Assessing the Productivity of Ancient Maori Cultivars of Sweet Potato, *Ipomoea batatas* [L.] Lam. in New Zealand

Referring sent: The four pre-European kumara cultivars "considered by Maori informants to be of pre-European origin or introduction (Yen 1963:33), 'Rekamaroa' was collected from Ruatoria, the East Coast, and the Bay of Planty; 'Hutihuti' from the East Coast, Bay of Pleny, and Northland; 'Taputini' from Northland; and 'Houhere' from two locations in Northland (Yen 1963) (see Fig. 2).

Fig. 2b Words from the article that can be taken using the data mining algorithm outlined in Appendix B. The words-in-map from 2a, and these mined words from 2b, together constitute the bag of words used as map metadata.

Perhaps the most pressing data mining question remains one of extent: how much is enough? A balance must be struck between mining enough metadata to classify each item precisely, and too much metadata that will let in noise and cause recall to suffer. In the case of MapSearch, examining a great many instances has allowed the deduction of rules as to the location and amount of text that should describe a map best.

These instructions for the data mining algorithm are in Appendix B. It was decided to risk taking too little metadata rather too much, and in only specific areas of the article, rather than follow the usual method of having the computer scan the whole article and weight judgment on those words that appear most frequently. It is rarely necessary to read the whole article to determine what a map is about, so neither should the computer need to go through the entire text.

5.2.3 Data cleaning

The main test of quality, whether the map is accurate, is beyond the bounds of this program. It is possible, however, to create full maps from many files, and weed out those maps which are somewhat illegible. It is a separate question whether an out-of-focus map is better than none and whether these should be included anyway.

Seven maps in the original articles were full-page spreads, requiring clipping and saving as two separate .jpg files. These double maps for housing, precipitation, farm yield, the Islamic world, and three of the Mediterranean, were knit together for the prototype. This could be done automatically by a panorama program such as comes with today's average digital camera.

A number of maps originally absorbed into the collection were removed due to their illegibility, owing to the fact that the .pdf article available was scanned in low resolution. It is possible to judge the graininess, or pixilation of the image, and the sharpness of its edges. MapSearch does not presently have this capability, although such adaptations could be added.

5.2.4 Weighting and indexing

The process of weighting and indexing described in Salton and McGill (1983, p. 71-75) remains standard procedure. The first step is to remove high frequency function words, called a stop list. MapSearch uses a pre-existing list. The next step might be to remove prefixes and suffixes to pare down to the stem. This is important in the present application in which the text might refer to a people rather than a place (Bahamian rather than Bahamas). Algorithms are available freely for stemming such as the Porter Stemmer.[12] This algorithm has a longer list of suffixes than are employed in MapSearch. A few lines of code in MapSearch cover aspects of stemming. The suffixes –s, –ed, –ing, –ist, and –es are removed if it leaves a valid word, or if removing the suffix and adding an –e would leave a valid word. The same is true for adjectival endings –an, –ean, –ician, –ern, –ian which are removed if a

[12] On the Porter Stemming algorithm, http://tartarus.org/martin/PorterStemmer/, and a perl script for the algorithm at http://tartarus.org/martin/PorterStemmer/perl.txt.

valid word remains. Future research will entail comparing freely available stemming algorithms or adding to what is in use for MapSearch to see whether better results may be obtained.

The words remaining, called a "*bag of words,*" are for potential matching with the domain ontologies. The next step for potential matching is to assign weights based on where the word is located in the article, how often it occurs, and for subject, whether it matches with a single or double-starred word in the ontology.

5.2.5 Information extraction

This is the task of identifying and classifying all recognized names and relations in the metadata. Information extraction (Leidner, 2007) is also termed data discovery (Tan, Steinbach, and Kumar, 2005). Details of the process as it pertains to MapSearch follow.

5.3 Metadata Classification

5.3.1 Classification rules

We need to decide what a map is and how to assign it to categories as a preliminary to setting up the retrieval algorithms. Some basis for decision is offered here.

Maps are considered for the database if they are at a minimum resolution of about 200 by 300 pixels; otherwise they will not be clear enough for on-screen viewing. The resolution of a map depends upon its original scan. Aerial photographs will be considered maps only if they include place labels. Two-dimensional and three-dimensional representations will be included if they show areas larger than a single building, in order to exclude architectural drawings. Some maps found in non-strictly geographic articles show areas with labeled cities. These maps seem to depict no theme, just region. However in the context of the article, the cities selected for labeling are generally significant for the article subject. Therefore, such a map also may be considered a theme map.

Each item should be assigned to *at least* one category within each facet, and two categories within each facet if the two categories score within perhaps 25% of each other. The generosity of percentage is due to the principle that it is better to class an item where it does not belong than to miss a classification. For items that get similar point values for three or more categories, assign to the first two categories based on metadata location (giving preference to the caption first, etc.). Assigning an item to more than one category within a facet acknowledges the work of Phelps and Wilensky that a single item has attributes of many categories and may be relevant in more than one context (2000).

Items are assigned categories upon entry into the database. The system performs post-coordinate or post-combination indexing in which each category may be used independently to retrieve the item.[13]

5.3.2 Classification categories

Analysis of map-related questions dictates the facets of region, time and subject (theme). More in-depth analysis of a large number of map-related questions would suggest facet subdivisions. However, in that the potential of MapSearch outstrips the potential of finding maps covering a wide range of time periods and subjects, actual questions might be an inadequate guide to subdivisions. It was decided instead to create categories consistent with what the wide-ranging map collection MapSearch aims to comprise.

The number of categories per facet was selected in part based on expediencies of how categories could be displayed on screen. Too many categories per level would make browsing by category excessively complicated, so the number of categories per facet was limited to between ten and fifteen. Each category can itself be expanded as the database expands.

5.3.3 An ontology for each facet

An ontology is a tool to improve the relevance of items retrieved. It works behind-the-scenes or at the back-end so that the user does not even know that it is there. The ontology is divided into domains of terms on the same level in an array that correspond to MapSearch

[13] From R. Pearce-Moses, a Glossary of Archival and Records Terminology, Retrieved December 26, 2007 from http://www.archivists.org/glossary/term_details.asp?DefinitionKey=989

categories. Any hierarchical elements in the ontology can be used for ranking of results by relevance, as discussed further below. Different ontologies include different words, so logically the choice of ontology affects retrieval.

MapSearch uses domain ontologies for region, time period, and subject to aid in retrieval of maps. General-purpose classifications for region (Geonames and the World Gazetteer) and subject (Library of Congress Classification supplemented with Library of Congress Subject Headings) were adapted to correspond to the MapSearch categories.[14] No ready-made classification or ontology for time was found, so an abbreviated set of time words was compiled. Each domain ontology was refined for exhaustivity (extent of the vocabulary) and specificity (precision of the vocabulary) during the phase of iterative testing with the training set of maps.

5.4 Query processing: algorithms and assessment

"[Q]uery processing is a major bottleneck in standard web search engines, and the main reason for the thousands of machines used by the major engines" (Chen, Suel & Markowetz, 2006, p. 277). Asking a program to inspect an entire ontology for matches or near matches slows operations. System designers cut processing time by using an ontology subset, or domain ontology, that corresponds to the topic area of the data, and by using the edge processing technique that compares similarities and judges relevance relatively fast. The MapSearch system expedites query processing by indexing items as they enter the database.

5.4.1 Algorithms

Each of the 150 items in the training set was inspected carefully and classified by region, time and theme categories. General principles were extracted to construct an algorithm that can then predict how to classify items as yet unseen. In terms of region, observed principles were combined with principles found by others as reported by Leidner (2007, p. 116).

[14] Much work has been done on how to establish a domain taxonomy with WordNet, that amounts to deciding which subset of WordNet to use for a domain (see Kim, Seu & Rim, 2004 or Chang, Huang, Ker & Yang, 2002, for example). Khan, McLeod and Hovy (2004) describe a mechanism for selecting these concepts automatically. The Protégé ontology developed at Stanford has solved this problem by presenting the user with a Protégé frames editor that allows users to develop WordNet classes and a hierarchy.

In writing the instructions list, or algorithms, that give general rules to classify, a concern is whether to cluster items based on rules or classes. Tan, Steinbach and Kumar (2005, p. 211-212) define *rule-based ordering* to be a method in which each rule is implemented, one by one, in decision tree fashion. An advantage of the method is that rules can be ordered so that every item is classified by the rule that takes highest priority. A disadvantage is that lower-ranked rules are harder to interpret. Tan, Steinbach and Kumar (ibid: p. 212) define *class-based ordering* as a method in which rules are sorted on the basis of classes, and all rules fire simultaneously. This method makes rule interpretation easier, although a higher priority rule may be overlooked in favor of rule with lesser value in predicting a classification, possibly making the classification less accurate. Rule-based ordering was elected primarily because the metadata location is so important with respect to reliability of prediction, and therefore rules must be applied in location order.

5.4.2 How to use an ontology to classify metadata or query

Each item manifests attributes of more than one class. Weighting allows clustering of attributes for classification. There are standard means to assign weights. In binary indexing, each term weighs in at 1, despite its frequency, versus in weighted indexing where a term assumes a different weight depending upon its importance, here determined by function or location in the document (Salton & McGill, 1983, p.73-74). In MapSearch, metadata words that recur, words found in particular locations in the article, and words that match with significant words in the ontology are assigned higher values. A single item may be assigned values in numerous categories within a facet. The category with the highest numerical value has the highest chance of being associated with the item, and is the most likely for classification. An item that scores high in two categories is assigned to both categories. This serves to improve measured recall, although some of the items retrieved might be less relevant than others.

The values for scoring were assigned on the basis of judgment rather than experiment. Determining the optimum balance among options requires a separate study, so an approximate balance is accepted here on the basis of preliminary testing. Many more examples must be considered to determine whether these would be the most effective parameters for the general case. Ten was used for a double-starred word, 5 for a single-starred word match and 1 for an ordinary match, rather than assigning the weights as 3, 2 and 1, because the importance of a double-starred word in determining to which category the item should belong is much greater than other words in the ontology that will not necessarily help in assigning items to categories. Multiplying terms by the number of occurrences is one of the standard strategies for weighting (Salton & McGill, 1983, p.205).

Below is an example of how weighting enables an item to be assigned to a category:

```
MAP "sweet potato new zealand.jpg"   Assignment: ArchAnthro (should be
Archaeology and Anthropology) OK
ArchAnthro.t 46
History.t 34
Politics.t 22
Religion.t 18
Society.t 10
Arts.t 6
Science.t 4
Technology.t 4
Commerce.t 4
Military.t 4
Medicine.t 2

CAPTION:  fig  2  map of new zealand
History.t:  zealand/8
Politics.t:  zealand/8

TITLE:  experimental archaeology gardens assessing the productivity of ancient
maori cultivars of sweet potato  ipomoea batatas  l   lam  in  new zealand
Religion.t:  ancient/4 archaeology/4
Society.t:  ancient/4 experimental/4
ArchAnthro.t:  archaeology/44
Science.t:  experimental/4
Commerce.t:  productivity/4
Arts.t:  potato/4
History.t:  ancient/4 zealand/4
Politics.t:  ancient/4 zealand/4

REFERRING:  of the four pre european kumara cultivars   considered by maori
informants to be of pre european origin or introduction    yen 1963 33    rekamaroa
was collected from ruatoria  the east coast  and the bay of planty   hutihuti  from the
east coast  bay of pleny  and northland   taputini  from northland  and  houhere  from
two locations in northland  yen 1963   see fig  2
```

```
Religion.t:  collected/2 east/4 european/4
Society.t:  origin/2
ArchAnthro.t:  collected/2
Medicine.t:  collected/2
Technology.t:  coast/4
Arts.t:  collected/2
Military.t:  coast/4
Politics.t:  east/4 see/2
History.t:  coast/4 collected/2 east/4 european/4 see/2 two/2
```

Fig. 3 Example of weighting: Sample output of classification by theme algorithm. The item received more points for the category of Archaeology and Anthropology (ArchAnthro total=46) than for any other category. The output shows which words were the category indicators and provided the points: it was mainly the word archaeology found in the title, which gave 44 points, and the word collected from the referring sentence, which gave 2 points.

5.4.3 How to rank results

Ranking of results becomes important when a database contains a large enough collection of items that several screens worth of results could match a query exactly. So it was necessary to determine which items match best and list these first. Users in a hurry probably will consult only the items that list first, making ranking critical.

Top ranking: relevance of scale or granularity. Within the set of all items that match a query, items will list in order of how closely they match the query term(s). Hierarchies in the ontology can be used to determine similarity to the query. The current version of MapSearch has at least two levels in each hierarchy: the category, and words that indicate and most often subordinate to that category. Further development of the categories and ontologies will provided additional hierarchical levels, such as demonstrated in the time category Modern which subdivides by decade, and these can be used to rank results.

Example query: Spain

Retrieved matches

1st (exact) —map of Spain
2nd (subordinate) —map of Madrid (part of Spain)
3rd (superordinate)—map of Europe (contains area outside query)
4th (coordinate) —map of France (because within Europe)

Those listed at the top of retrieved results will contain within their metadata an exact match to the query. Those listed next among the retrieved results will contain within their metadata matches within the same category as the query. After these, are listed those whose metadata matches within the same category as the query but at a different level of the hierarchy. So that, for example, the query "Spain" will list a map of Spain first (an exact match), then a map of France (same level: country in Europe), and then a map of all of Europe (different level of the hierarchy).

Second level (or higher with user option) ranking: quality relevance**.** The system in some cases will be unable to discriminate among granularity levels, and in some cases it will retrieve multiple items on the same hierarchical level. The system will draw then upon other characteristics of the data.

- Publication date of article
- File resolution
- Color/monochrome

Sorting results based on these characteristics of the data should the prerogative of the user. Options to change the sorting could be in a drop-down menu (as in Google Image Search) at the top of the result screen.

5.4.4 How good is a given classification? Scoring

The point of reference for machine classification has been set as the manual classification. This decision was made because in some cases, there is one category that fits exactly while in other cases the choice of category is ambiguous. Beyond this, the automatic classification does not need to exceed human capacity, it need only perform comparably well. So for these reasons, the measure of machine classification is set at the manual.

Automatic classification can match the manual classification exactly, but in some cases mainly due to category overlaps, the classification can be plausible. Plausible matches do not occur in categories for time, which are discrete by year. Categories for region have been set up to correspond with geography such that, for example, countries in the Caribbean and West

Indies are not included in Central America. But for theme, each item scores in multiple categories, and the category with the highest score delivers the theme assignment.

Plausible interchanges of classification category for theme are set up mostly because one category is a subset (Medicine can be said to be a subset of Science) or overlaps another (Technology overlaps with Science). The following interchanges were set up for the purposes of scoring:

History and Travel with Archaeology and Anthropology;
History and Travel with Military;
History and Travel with Politics and Law;
Society with Religion and Education;
Commerce and Finance with Politics and Law;
Science with Agriculture with Technology and Transportation;
Science with Technology and Transportation;
Science with Medicine;
Technology and Transportation with Military;
Technology and Transportation with Commerce and Finance.

The large number of overlaps for a single category suggests that the boundaries or perhaps the categories themselves should be changed. Technology and Transportation, overlapping with four categories as shown above, and History and Travel overlapping with three categories, present themselves immediately as candidates for change.

5.5 Improving classification algorithms

The purpose of the training set is to train the algorithms (here with manual rather than machine learning) in the better application of the classifications. The algorithms are created and improved through inductive reasoning: these principles seem to apply to the training set, therefore, they will apply in unseen cases. Re-running the training set and improving the algorithms is not to manipulate results, but rather, to improve the algorithms. The algorithms are evaluated on items unseen—the set of items known as the testing set.

Potential sources of error in causing misclassification may be attributed to data mining (insufficient metadata compromising precision or overly much metadata compromising recall), ontology (too much noise or insufficient detail to resolve metadata, compromising precision), correlation of ontology with classifications (systematic error), length mis-matches (shortening metadata by stemming or lengthening metadata to allow match phrases), or algorithm bugs (minor programming errors). Careful analysis of misclassifications in the training set hints at which parameters should be changed to improve the algorithm.

The algorithm clearly could be further improved by analyzing more training items for misclassifications to try to determine what can be altered or how to add heuristics so that the misclassified item(s) will be classified correctly in future. A larger test bed allows firmer generalizations with higher validity. One danger is that the algorithm will be overfitted to accommodate the test bed. Wang, Hodges and Tang (2003, p. 564) believe that the primary reason they have not been able to improve their classification algorithm is that they used a relatively small number of training documents—772. The same could be a problem here, with the training document set of only 150.

How accurate does automatic classification need to be ultimately? Or: when should one stop tinkering in the hope of improving the protocol? Larson found in conducting a similar automatic classification that only 46.6% of the sample could be classified correctly (1992, p.147). Bates commented that it is typical for researchers to present a new system as 70% accurate (Bates, 1998, p. 1186). She pointed out that achieving the last 30% is vastly more difficult—all the more so as the collection grows.

It seems clear that systems do not need to perform at 100% accuracy; they only need do work reliably that many find useful and that could not be done otherwise. Take Google, for example. Google does not perform at 100% accuracy; in fact, its recall is poor. But even if an army of web indexers could be found and paid, that army could not keep abreast of the billions of web pages and return results instantaneously, as does Google. This excuses Google's imperfection. The same may be said of MapSearch. Maps within publications that are the subject of this study are unseen by most web crawlers, so any retrievals that are relevant are better than none.

6 Classification architecture

This chapter elaborates sections 5.4.1 and 5.4.2 in the previous chapter on automatic classification of maps in order to enhance clarity, facilitate comparison among the three algorithms, and expedite meta-analysis for future scholarship. The three algorithms--region, time and theme—developed in this dissertation for retrieval of maps could be applied, with only minor adjustments, to indexing and classification of other data types.

After a section on methods common to all classification mechanisms, the chapter subdivides by facet. It discusses by facet the choice of classification categories, the referring source of ontology for the referent, and heuristics comprising the retrieval algorithm. (A heuristic is an approach to solving a problem that has no provable justification but that has been found to work.) The chapter concludes with a chart to compare results of classifications in each facet, and a discussion of elements common to automatic classification in all facets.

6.1 Method

Automatic classification entails categorizing items without human supervision. It does not require the creation of categories, as was done for this study in order to have categories specific to the collection. Nor does it require the use of domain ontologies for each category, as was done for this study to improve the relevance of results retrieved.

The set of heuristics comprising each algorithm accomplishes two things: resolving metadata into useable terms to index, and indexing those terms into categories.

The beginning heuristics are meant to help weed noise and resolve relevant metadata into “knowledge”. These heuristics were deduced by examining 150 metadata cases in the training set manually. Several heuristics for resolving place names were incorporated from the works of others (Leidner, 2007) and (Amitay et al, 2004). No supporting literature was found to draw upon for heuristics for resolving time that could be applied to MapSearch, so these were deduced from scratch. Supporting literature on resolving by subject came mostly by way of understanding what makes a useful subject ontology. The fundamental rules that the program would use to assign maps to categories were devised at the time the maps were classified manually. However, certain difficulties the program would encounter such as with place names that were entirely capital letters, were not apparent. Only by running the maps through the system repeatedly could these problems be gleaned and the rules sharpened.

The final heuristics for each algorithm assign items to categories, and then finally discriminate among the items that match “better” than others to be listed higher in retrieval ranking. These classification algorithms are the project of armchair logic. These are tested by running maps repeatedly through the system to refine the logic.

The heuristics were sharpened by examining metadata for each misclassified item to determine what led to the misclassification and what, if anything, could be done so that a previously unseen item would be classified correctly in the future without producing other errors as by-products of the changes. If no source of error were found, the misclassification could have been a result of a coding bug. And in the odd instance, an item may have been counted as a misclassification, when in fact it had been classified correctly by the program and incorrectly by the researcher.

Examination of an extract from an output for classification by region will demonstrate the method of how the algorithms were improved. The program retains the images but works exclusively from the metadata for the purposes of classification.

> Sample output from region classification:
>
> ```
> Processing “med2.jpg
> Europe 10
> “med2.jpg” should be in Asia program: Europe wrong
> ```
>
> Analysis:
> (Question 1) What led this map to be classified as Europe?
> (Question 2) Why was it not classified as Asia?
>
> To answer these questions is to examine closely the mined data associated with the map.
> (Answer 1) The place reference in the caption is “(Rome, 1570)”.
> (Answer 2) The referring sentence in the metadata reads “depiction of the Island of Cyprus, with the limits of Caramania, Syria, Judea, and Egypt.”

Suggested change to the heuristics:
The referring sentence could be given the same weight as the caption. That way, the program would classify the map in both Europe and Asia. This is the only way to get a correct classification without overlooking potentially valuable metadata for other history maps.

Test of the suggested change:
The training set is run through the algorithms again to check that, in clearing up this error, more have not been introduced.

6.2 Region

6.2.1 Classes and their application

Categories were created for the continents and regions of the earth above water, with World given its own category. The categories were initially:

North America
South and Central America
Europe
Middle East
Asia
Africa
Australia
Antarctica
Arctic
Atlantic Islands
Oceania
World

These categories could not be applied satisfactorily in practice, and so they were emended. The Middle East category was rejected because every country within the Middle East required double classification: both Middle East and either Africa or Asia. Atlantic Islands was rejected because it became unclear what areas other than Greenland belonged in the category. In addition, Greenland is classed by some gazetteers as North America, so this classification was upheld. The South and Central America category seemed inadequate to hold Cuba which is close to the United States, and for the large number of islands in that part of the Atlantic, and so the category Caribbean and West Indies was added.

The labels were improved to remove ambiguity and add clarity:

North America
Caribbean and West Indies
South and Central America
Europe
Asia
Africa
Australia
Antarctica
Arctic
Oceania
World

Ultimately these categories will need to be subdivided because analysis of geographically-related user queries shows that people ask about very small regions. Sanderson and Han (2007) analyzed 188 geographic query words, of which 63 related to countries, 67 to smaller areas such as states, provinces, counties or special areas, and 41 to city or city borough

People sometimes employ vague geographical concepts, termed vernacular geography (Pasley, Clough & Sanderson, 2006). Examples are "downtown" and the "grim area around the docks." To capture such imprecise concepts in query terms, a graphical interface with a map that zooms to area and approaches street level would be called for. This refinement is left to future research.

6.2.2 Geo-ontology

English ontologies for geographic information have been prepared by the International Standards Organization (ISO), the Federal Geographic Data Committee (FGDC) and the Open GIS Consortium (OGC).[15] Online gazetteers include the World Gazetteer with cities, states and countries but without topographical features, the Getty Thesaurus for Geographic Names, and Geonames.org.[16] All are updated regularly. The Getty Thesaurus is not freely available as a complete download as are the World Gazetteer and Geonames. In order to accommodate others who might wish to replicate this research, and for the sake of the present project which is of indefinite length, the freely available sources were selected.

[15] http://loki.cae.drexel.edu/~wbs/ontology/list.htm, Retrieved February 4, 2008.
[16] They can be found on the web at http://world-gazetteer.com/, http://www.geonames.org/, http://www.getty.edu/research/conducting_research/vocabularies/tgn/, Retrieved February 21, 2008

Geonames includes place names (also called toponyms) and their coordinates in a hierarchy. It is an ontology in Web Ontology Language that uses ISO-3166 alpha2 country codes. The hierarchy has *narrower terms* called "children," *next to* terms called "neighboring," and *places in the area* called "nearby."

Geonames includes over 8 million geographical names.[17] This comprehensiveness is in some respects detrimental in that, when a place name is found in metadata, it may be unclear which region in Geonames is the correct match. Two methods are used to lessen the ambiguity. One is that the World Gazetteer is used for the first pass to resolve place names because it has a much smaller name database. The other is that, for the purposes of this research, Geonames has been limited to upper level administrative regions, or places with large populations. The codes selected appear in Appendix C.

The World Gazetteer is about a quarter the size of Geonames, but includes summary population and other statistics, and national flags as well as place names. The website explains that it uses official data, when such is available, and when not, it draws facts from year books, encyclopedias, atlases and other types of reference sources.

The ontology moreover should be refined by removing the most common English words called stop words that are also place names (Grove, Spain), (Bath, United Kingdom), (Buffalo, New York). MapSearch uses a word list originally created in the 1970s for Unix.[18] The difficulty is that some stop words cannot be removed from the ontology: "china," "world," and "union," for coupling with European, for example.

6.2.3 Algorithm for region

Purpose of algorithm: Given a map within a digital text, classify that map into a category provided. Heuristics below make up that algorithm.

[17] www.geonames.org/about.html, retrieved December 18, 2007

[18] The file /usr/dict/words ships with Linux systems and contains a list of several thousand common English words and names.

/H0/ (Heuristic) Distinguish place words from non-place words

Indicators for place words

(a) Place words begin with capital letters

(b) Multi-word phrases in capital letters

(c) Place word(s) follow "Map of…"

(d) Place word(s) precede "map" or "eco-region" or "region" or "locale"

(e) The top 100 words from Geonames (such as bay, stream, center, hill, mountain, north, east, south, west) indicate place, so that a word found within two words of one of these could be a geographical name.

/H1/ (Heuristic 1) Location of metadata. Go through metadata regions searching for place words in the following order:

1. map caption
   words in map (if any)
   article title
2. sentence in article that refers to the map (if any)
3. paragraph containing sentence that refers to the map
4. first sentence in article or abstract
5. first paragraph of article or abstract
6. additional paragraphs

/H2/ Amount of metadata (for optimum recall and precision)

Continue scanning metadata locations /H1/ from 1-6 until a classification has been assigned.

/H3/ Multivalent classification Match metadata in each location in /H1/with one or two classifications.

Example Metadata: France, the Colonies and the Revolutionary War is classed both in Europe and in North America

/H4/ Preferences

In metadata

(a) Prefer place names that are repeated.

(b) Stem metadata as needed to match with place referent.

(c) If two places are found in one metadata location as listed in /H1/, prefer metadata not in parentheses

(d) Consider a name a place if it is qualified within two words by another place name

Example: In "Sydney, Australia," Sydney is not a person but a place in Australia
Example: St. Paul is not a place in Minnesota when found in metadata "the Epistles of St. Paul"

(e) If two places are in one metadata location as listed in /H1/, and one or both correspond to more than one place in the referent, select the two places that are closest in distance. This is called "geometric minimality" by Leidner (2007, p.104).

Example: Lincoln, Nebraska in metadata
We know Lincoln is a place not a person because of (d)
We know that Lincoln is in the United States and not in the U.K. because of (e)

(f) If two places are found in one metadata location as listed in /H1/, prefer the one higher in the referent hierarchy

Example: New Brunswick in metadata
Will resolve to New Brunswick, Canada rather than New Brunswick, New Jersey because of (f)

In referent Match the metadata name to the most common occurrence of the place name, as is accomplished expediently by using a referent that is less rather than more complete, or by filtering a more complete referent.[19]

/H5/ Weeding out noise in metadata

(a) Exclude phrases of the sort "published at/in [place]"

(b) Exclude newspaper names such as "New York Times" or "San Francisco Chronicle"

[19] This rule replaces the rule of "largest population" in Leidner (2007, p. 103).

/H6/ Classification The ontology hierarchy must be correlated with the MapSearch categories

(a) Use table of country correlations between Geonames countries and MapSearch categories

(b) Use table of waters for correlations between major oceans, seas and rivers and MapSearch categories

(c) If no other place names are found but "world" "globe" or "global" appear in any of metadata locations H1 1-5, assign to category "World"

(d) If three classification regions seem to match, assign to category "World".

Example. Metadata: In the latter map, those same goods move in two lines, one across the eastern Mediterranean and the other across Anatolia, connecting western Europe to West Asia, East Africa, and India. This item should be classified as "World"

/H7/ Ranking by relevance of scale Rank first those matches that represent the smaller scale and correspond to the higher place in the referent hierarchy.

Example:

Browse query: North America and Europe
Retrieved: map of Spain and map of Manhattan
Ranking: map of Spain, map of Manhattan

/H8/ Ranking by relevance of data attributes When relevance of scale is not clear from the metadata or when items are equivalent in scale, rank according to data quality. Several options are suggested

(a) Color – grayscale – black and white line

(b) Most recent publication date first

(c) Highest resolution first

The user has the option to overturn ranking by scale /H7/ in preference to one of these rankings.

6.2.4 Findings

Number of maps classified by region	Number region classifications agreeing automatic with manual	Percent region automatic classifications agreeing with manual
150	123	82%

Table 1. Findings when the training set is classified for region.

In answer to the first research question, findings show that it is possible to classify automatically 82% of maps in the sample into the same region categories that a person would classify them.

Difficulties associated with automatic classification by region include how to distinguish place names from non-place names, how to determine which place of many with the same in the gazetteer ontology localize the metadata, and how to tell which of the possibly many places named in the metadata are relevant to the map. Each problem and its attempted solution via heuristics is discussed in turn.

What is involved in distinguishing automatically between geographic and non-geographic names? The section on related research mentions others who have worked on this problem, and the MetaCarta software that resolves geographical names in documents. This section focuses on heuristics. Two heuristics from Amitay, Hare'El, Sivan and Soffer (2004) adapted for MapSearch are capitalization and stop word removal. The Web-A-Where system by Amitay et al. attaches a location to each place named in a web page. The authors prepared their own gazetteer with about 40,000 places around the world (compare this to the more than 8 million names in Geonames used in MapSearch), and with a separate section listing place names that are also commonly used words. MapSearch uses a stop word list with place names that are also common words removed. Capitalization also helps to discriminate between place words and common words (Reading, Pennsylvania).[20] But in instances in which the whole word is capitalized as is sometimes the case in an article title, this indicator is erased. To the solutions of stop words and capitalization to distinguish between geo- and non-geo names used also by Amitay et al., Professor Lesk had the novel idea of extracting as place indicators the 100 most-used words in the Geonames gazetteer, such as hill, stream, center, mountain, north, south, etc. Then, for example, BEAR MOUNTAIN clearly is not an animal but a mountain named Bear.

Another aspect of resolving a place name requires recognizing that there are multiple names for the same place, or *aliasing*. Difficulties come with differences in spelling between the item metadata and the gazetteer, or when there are two accepted names for the same place such as Los Angeles and LA. Even when a list of alternate place names spellings from Geonames was incorporated into the MapSearch algorithm, the metadata Herakleion did not match with the alternate transliteration of the Greek Heraklion and so the item was not classified correctly. This problem of equally acceptable spellings for place names that are not in Roman alphabets remains unsolved.

The program does not necessarily know what place is meant when it spots a place name in the metadata because the gazetteer includes many different places with the same name, or *polysemy*. The problem of named entity recognition (Zhou and Su, 2001), indexing geographic locations (Vaid, Jones, Joho, & Sanderson, 2005) and toponym resolution (Leidner, 2007) has become of interest. Geonames includes 26 different places named Greenland, for example. The MapSearch project benefited from Leidner's comparison of heuristics in numerous programs that attempt to solve the problem. For example, the heuristic which he refers to as "geometric minimality" (p. 104) instructs that when two places are named, they should be resolved to minimize the pairwise distance between them. If an article mentions Rome and Albany, according to the heuristic, that article is referring to New York rather than both Rome, Italy and Albany, Oregon or Albany, Australia.

[20] Capitalization is not a reliable indicator of subject or time. For example, as an adjective, Renaissance art is upper case, but the renaissance period is lower case.

A place, once localized, must still be classified. Correct classification depends upon a matching of hierarchy between the Geonames gazetteer and the MapSearch categories. Correlation lists were provided to link countries to MapSearch categories, and rivers, seas and oceans (some of which had to be associated with several regions). Protectorates represent a lesser problem in classification. This is because the Geonames gazetteer defers to the legal sovereignty of a nation for its hierarchy, but the MapSearch hierarchy is purely geographical. For example, Guam is within United States jurisdiction and Geonames classifies it in North America, but for the purposes of MapSearch, Guam is in Oceania. The few instances of countries with holdings in distant parts of the globe will cause classification errors. This problem will be fixable with additional correlation lists between Geonames and MapSearch. The problem of distant protectorates comes up rarely, however, so that it was not adjusted for in the prototype.

Another problem resolving place names is specific to the situation of map extraction and data mining. A place, even if resolved correctly, is not necessarily relevant to the map. False leads come from references to the “New York Times” or “Published in X place” that might be found in near-map areas such as the caption. Analysis of maps in a larger training set would give a broader understanding of how this problem is manifest and what sort of fixes for it can be included in the algorithm.

On an optimistic note, the addition of the algorithm to extract words from maps should improve classification by region. In this study, words-in-map data was confined to words that are oversize. The program to extract words from each map will able to mine words of a point size of 12 and smaller, provided the resolution is adequate.

6.3 Time

6.3.1 Classes for time content and their application

The categories were chosen to cover the span of earth time. However, these categories will not cover the *collection* in a balanced way. The spans of the first four categories, therefore, are wide to maintain balance. The application of the categories showed that more than 60%

of the maps belonged in Modern. Subdivisions were created to afford some aggregation among maps that are recent.

Prehistory	(—801 B.C.)
Antiquity	(800 B.C.—476 A.D.)
Middle Ages	(477 — 1450)
Early Modern	(1451 — 1914)
Modern	(1915 —)

- World War I [21]
- 1920s
- 1930s
- 1940s
- 1950s
- 1960s
- 1970s
- 1980s
- 1990s
- Current, 2000—

Category breaks were created on the basis of events in Western Civilization that demark an age and are thus biased toward our culture. Correspondence of labels with dates makes them classify perfectly well for the civilization of China, for example, but a study of a heterogeneous body of users would be needed to determine whether the labels become a source of confusion, for example, in using “Middle Ages” in referring to events in a non-Western civilization.

Prehistory contains geologic time and mankind before writing up to the recording of the Greek poems, *Iliad* and *Odyssey*. Antiquity embraces the time of Homer, who flourished about 800 B.C.E., and the flowering of classical Greece and then Rome. The fall of Rome in 476 A.D. marks the end of the era. The Middle Ages category spans the dark ages of barbarian invasions of Europe through the high middle ages. The Early Modern period opens with the early renaissance in Italy, about 1450 and it continues up to the 20th century. The Modern age, seared with mass casualties of Great War technology, begins in 1915. It was determined later in the course of the evaluation that an additional category is needed for Future. This has yet to be implemented.

[21] The first but not second world war is included as a subdivision because it corresponds to breakdown of the century by decade.

6.3.2 Chrono-ontology

Alonso, Gertz and Baeza-Yates (2007, p. 38) have identified three categories of temporal expression: explicit (such as by date), implicit (such by names of administrations or empires), and relative expressions (such as “today” or “X years ago” that can be anchored only with respect to an explicit or implicit expression). The MapSearch system depends mostly on explicit expressions in extracting numbers, but also uses a word list to find implicit expressions.

The markup language specifically for time is not yet used widely. A web ontology language for time by the World Wide Web Consortium is still in draft form.[22] Pressing uses of the time ontology will be linked to place in that daylight and clock time are distributed differently throughout the world. A directory for time need not work to the hour for historical work. Petras, Larson and Buckland (2006) devised an XML schema describing web time concepts. They intend to populate the directory by both extracting suitable headings from catalog records, and by adding data manually, but no directory is presently available on their research website.[23]

A simple word list for time was built here for information retrieval purposes. Simplicity is not a drawback because the classification rules are based upon numbers. Some of the time words were excerpted from the History class of the Library of Congress Classification; other words were added that are associated with a style (Romanticism) or a period of political stability (Latin Christendom).

6.3.3 Algorithm for time

Purpose of algorithm: Given map within digital text, classify that item into a MapSearch time category. Heuristics below make up that algorithm.

[22] Markup language for time at http://www.timeml.org, and draft ontology template for time, retrieved February 6, 2008 from http://www.w3.org/TR/owl-time/

[23] http://ecai.org/imls2004/timeperiods.html, retrieved January 12, 2008

/H0/ Distinguish time numbers from other numbers

Signifiers with numbers Scan 2, 3, 4 or 5 digit numbers with modifiers either preceded or followed by B.C. or B.C.E. or BC or BCE or C.E. or CE or B.P. (Before Present) or Mya (Million Years Ago) or AD or A.D. or ca. or circa, or within two words of "year" or "date"

"X years ago" is a date unreliable for classification with the methods outlined below, although it might be built into the algorithm in due course.

Example: "…distribution of pine from 18,000 to 500 years ago" should not be classified in Middle Ages because of 500 years but rather should be in Prehistory

Dashes with numbers Take numbers in a span (3 or 4 digit numbers separated by a hyphen (1900-1950), en dash (1900–1950) or em dash (1900—1950).

/H1/ (Heuristic 1) Location of metadata. Go through metadata regions searching for numbers or time words in the following order:

1. map caption
   words in map (if any)
   article title
2. sentence in article that refers to the map (if any)
3. paragraph containing sentence that refers to the map
4. first sentence in article or abstract
5. first paragraph of article or abstract
6. additional paragraphs

/H2/ Amount of metadata (for optimum precision and recall)

Continue scanning metadata in the order given in /H1/ from 1-6 until a classification has been assigned.

/H3/ Multivalent classification Match metadata in a location with one or two categories.

Example. Metadata with "Figure 1. Locations of the 1862, 1890 and 1994 Land Grant Universities" is classified both in Early Modern and in Modern

/H4/ Preference in metadata

Repetition Prefer numbers or time words that are repeated

Classification with numbers Take every number regardless of parentheses as detailed in /H5/ if the item is classified in History

Metadata location Take *all* numbers in caption and title metadata

/H5/ Weeding out noise in metadata

(a) The following sub-rules attempt to avoid using numbers in bibliographical citations as dates, so the following forms are not harvested:

a. (19##a) OR (19##b) OR (20##a) OR (20##b)
b. (name, 19## Or 20##) OR (see Or e.g. Or see also name, 19## Or 20##)
c. (name 19## Or 20##) OR (see Or e.g. Or see also name 19## Or 20##)
d. (name et al. 19## Or 20##) OR (see Or e.g. Or see also name et al. 19## Or 20##)
e. (name et al., 19## Or 20##) OR (see or e.g. Or see also name et al., 19## Or 20##)
f. (name and name, 19## Or 20##) OR (see Or e.g. Or see also name and name, 19## Or 20##)
g. (name, with or without comma 19## Or 20##**;** name, 19## Or 20##)
h. (name, with or without comma 19## or 20##)**;**
i. Any of the above with additional words Or p. Or pp. within the parentheses so that the end parentheses does not close after the date
j. Name (19##) Or Name (20##)
k. Name (19## Or 20##, 19## Or 20##) → two dates in parentheses
l. (19##). Or (20##).
m. (19##): Or (20##):
n. (month-month, 19##) OR (month-month, 20##)
o. (month, 19##) OR (month 19##) OR (month, 20## or (month 20##)
p. (word, abbreviation for state, or major publishing city Berlin, London, Amsterdam, 19## Or 20##)

(b) Do not harvest numbers followed by M. or meters or km. or kilometers or acres or miles or any other unit of distance

(c) Do not harvest numbers followed by degrees

(d) Do not harvest numbers with decimal points or fractions

(e) Do not harvest “more than ####” or “less than ####”

(f) Do not harvest numbers preceded by currency symbols

/H6/ Classification

By number

```
If a number is found, assign to category
    If xxxxx or xxxx B.C. or C.E. or B.C.E., or if xxx BC or BCE > 800
     ➔ Assign Prehistory
If $0 \leq$ x or xx or xxx BC or BCE $\leq 800$  OR   if  $0 \leq$ x or xx or xxx $\leq 476$
    ➔ Assign Antiquity
If $477 \leq$ xxx $\leq 999$, or $1000 \leq$ xxxx $\leq 1450$
    ➔ Assign Middle Ages
 If $1451 \leq$ xxxx $\leq 1914$
    ➔ Assign Early Modern
 If $1915 \leq$ xxxx $\leq 2040$
    ➔ Assign Modern, and to subdivide Modern by decade,
            $1915 \leq$ xxxx $\leq 1919$   → Assign Modern, World War I
            $1920 \leq$ xxxx $\leq 1929$   → Assign Modern, 1920s
            $1930 \leq$ xxxx $\leq 1939$   → Assign Modern, 1930s
            $1940 \leq$ xxxx $\leq 1949$   → Assign Modern, 1940s
            $1950 \leq$ xxxx $\leq 1959$   → Assign Modern, 1950s
            $1960 \leq$ xxxx $\leq 1969$   → Assign Modern, 1960s
            $1970 \leq$ xxxx $\leq 1979$   → Assign Modern, 1970s
            $1980 \leq$ xxxx $\leq 1989$   → Assign Modern, 1980s
            $1990 \leq$ xxxx $\leq 1999$   → Assign Modern, 1990s
            $2000 \leq$ xxxx $\leq 2010$   → Assign Modern, Current

If two numbers appear in a span separated by dashes
            →  Assign one category for the first number and one category for the second
            number (provided the two numbers fall into two categories)
```

By word

```
    If $1915 \leq$ xxxx $\leq 2040$ if includes "today" Or "current"
            → Assign Modern, Current
```

Classify in Prehistory, Antiquity, Middle Ages, Early Modern and Modern using the time word lists

By default

```
    If neither number nor time word is found,
            →  Assign to Modern and subdivide based on article publication date
```

/H7/ Ranking by relevance of time Rank matches first that are closest in time to query, then when time is known, list matches chronologically.

Example: Query 1850 France
Results Map of 1865 France, then map of 1890 France, then map of 1900 France

/H8/ Ranking by relevance of data attributes When specific time is unclear from the metadata or when items retrieved are equivalent in time (both dated to the Carter administration, for example), rank according to data attributes. Suggested:

(a) Most recent publications first, listing backwards in time, with date unknown items last

(b) Color first, then grayscale, then black and white

(c) Highest resolution first

6.3.4 Findings

Number of maps classified by time	Number time classifications agreeing automatic with manual	Percent time automatic classifications agreeing with manual
150	135	90%

Table 2. Findings when the training set is classified by time.

Findings show that it is possible to classify automatically 90% of maps in the sample into the same time categories that a person would classify them, in answer to the second research question.

The algorithm relies mostly on numbers to classify items into the categories of Middle Ages, Early Modern and Modern. Only classification within Prehistory relies mostly on matches with terms in the domain ontology, such as terms for geologic time and the early ages of man. Including "prehistory" in the Prehistory ontology became problematic because the term is used to mean before written history, or before writing, and so for some peoples might indicate periods classified in Antiquity or in the Middle Ages categories.

The difficulties in automatic classification by time resemble but are less complex than those in classification by region. Problems include how to distinguish numbers meant to designate a date from those that do not, and how to distinguish date numbers that describe time content of the map from dates that are unrelated to the map. Attempted fixes for these problems written tersely into the heuristics in the previous section, are described in more detail below.

The problem of how to distinguish numbers that designate time could be solved if authors consistently used B.C.(E.) or A.D. after dates, or M.y.a. (million years ago) or B.P. (before present) for geologic time. But authors do not invariably include these. Symbols found next to a number for percent, currency, temperature or distance indicate that the number is not a date.

The problem of how to distinguish a date potentially relevant to a map from date that is irrelevant is complicated by bibliographic citations (name, date) that may appear in the map caption or referring sentence. Heuristic /H5/ identifies and excludes dates in many bibliographic formats in close-to-map metadata. Articles classified as history, however, often have more dates than articles in other categories, and experience with the training set shows that many of the these dates are relevant to the map. So when the item is classified as history, /H5/ is not invoked.

Oftentimes no date appears in the item metadata. In many of these cases, the understood time is the present, and the article publication date is used to indicate the classification category.

Minor difficulties in classification obtain when an item is assigned more than one date. For example, according to the algorithm, dates in a span (1811-1920) are automatically classed in two categories, even though most of the period falls into a single category (Early Modern). Also, when three dates are found in a single metadata area that fall into three categories (or two Modern subcategories), the item is limited to two categories. These classifications, although suboptimal, appear correct.

6.4 Theme

6.4.1 Classes for theme

The Library of Congress Classification system includes 18 main classes, which were compressed for MapSearch into 12. Problems arose in the application of the categories in that the reach of the domain was not obvious from the label. Television, radio and film, for example, are classified within Arts, but it could as easily have been classed within Society (for communication) or Technology (for how they are produced). More specific labels were created in answer to this problem. The categories of Arts, History, and Technology have been revised to Arts and Media, History and Travel, and Technology and Transportation. Further, an unraveling of subtopics within each category will be provided in the interface.

The categories revised:

Arts and Media
History and Travel
Archaeology and Anthropology
Society
Commerce and Finance
Politics and Law
Science
Technology and Transportation
Medicine
Agriculture
Military
Religion and Education

6.4.2 Concept ontology

The MapSearch ontology for theme had to be built because, as discussed in section 3.3 earlier, the often-used WordNet subject ontology has known drawbacks.

The ontology backbone was taken from the Library of Congress Classification system Main Categories and Subdivisions because the full system is unavailable in digital form. The Library of Congress Classification System (LCC) was selected as a backbone for being comprehensive (as WordNet), hierarchical (as WordNet), and logical in organization (where WordNet falls short). The Library of Congress Classification needed to be adjusted for information retrieval purposes along the lines of a thesaurus. Shearer (2004) and Nielson

(2004) consider how best to construct a thesaurus. Foremost is that a thesaurus should contain terms that are relevant, and the Library of Congress Classification which is applied to a universe of (book) topics fits this requirement. Also like a thesaurus, it controls for synonyms in its “use for” category. Unlike a thesaurus, however, the Library of Congress Classification contains terms that are ambiguous and could belong to more than one category, and words irrelevant to classification in that they are instructions to the classifier. These ambiguities were lessened by adding weight to unambiguous terms with a star weighting system.

The weighting system was implemented to strengthen the ontology. Starring terms makes them “count” more for indexing purposes, with the intent of keeping the indexing words above the noise. How much noise is included in the non-starred words in the subject ontology? To test, the maps were run through the classification algorithm with the not starred words assigned a low weight, and again with the not starred words unweighted. Results were slightly higher when the not starred words were assigned a low weight.

As mentioned above, the initial ontology for theme was created by condensing the 18 main classes of the Library of Congress Classification into 12 categories, adding double or single stars to make words that best predict each category stand out, and isolating compound terms to act as single words. Stars were added manually: double stars for essentially unambiguous class determinants, and single stars for possibly ambiguous category determinants. Phrases suggested as class determinants (such as “first aid” for the class Medicine) were set on a separate list so that the program would look for them in compounds. Please see figure 4 for an example of how the ontology backbone was constructed.

Archaeology and Anthropology ← category created for MapSearch

CC1-960	Archaeology**	← double stars added for MapSearch
CC140	Forgeries of antiquities	
CC200-260	Bells. Campanology. Cowbells	
CC300-350	Crosses	
CC600-605	Boundary stones	

Fig. 4 Library of Congress Classification excerpt adjusted for use as an ontology

Not enough words made the theme ontology weak. It was supplemented with Library of Congress subject headings. A random sample of 130,000 catalog records for books classified with the Library of Congress Classification system was dissected to yield subject heading per book.[24] This offered another 8000 terms distributed among the 12 domain ontologies. The randomness of the sample left the politics category shy of words, so a dozen or so relevant terms were added manually to balance the lists. Testing demonstrated that this augmented ontology is a better tool for MapSearch.

The ontology can be used for MapSearch result ranking. The degree of similarity among terms can be calculated automatically using the notation. An exact match has a weight of 0. A match on the same level weights 1, although entries on the same level of each array[25] cannot be used for determining similarity as they tend to vary in concreteness rather than semantic relatedness. A match a level above or below the query word weights 2.

6.4.3 Algorithm for theme

Purpose of algorithm: Given map within digital text, classify that item into a MapSearch theme category. Heuristics below make up that algorithm.

/H1/ (Heuristic 1) Location of metadata. Go through metadata regions searching for matches with the theme ontologies in the following order:

1. map caption
   words in map (if any)
   article title
2. sentence in article that refers to the map (if any)
3. paragraph containing sentence that refers to the map
4. search the text of the entire article

[24] In MARC format, this is the equivalent of a single 6xx field per book, with only the $a heading of each string and none of the $x, $y, $z or $v subfields.

[25] The term array for the subdivision of the facet is attributed to Ranganathan (J. Mills, Faceted classification and logical division in information retrieval, *Library Trends 52*(3), 541-570, (Winter 2004), p. 550.

/H2/ Amount of metadata

Continue through metadata locations from /H1/ 1-4 until a classification has been assigned.

/H3/ Multivalent classification Assign item to the category for which it gets the highest score, but items that get within perhaps 25% of the highest scoring category should be assigned to both categories. This 25% figure must be further tested.

/H4/ Preferences

In metadata

(a) Repetition. Prefer metadata words that recur

(b) Stemming. Stem metadata to match ontology words

In referent: Prefer matches on double-starred, and then single-starred words in the ontology.

/H5/ Weeding out noise

(a) No metadata matches on stop word list composed of the articles, particles and pronouns that are the syntactic glue of the English language.

(b) No matches on words repeated in the metadata such as “map” or “figure.”

(c) No matches on metadata words isolated because they recur in multiple domains of the ontology such as “method.” The word “history” recurring throughout the ontology domains is a special case, and will be processed in metadata only for the domain of history.

/H6/ Classification

Assign item to category for which it gets the highest score, using the rule

Ontology (to weed out noise)		
	Word in label	12 x number of occurrences
	Word **	10 x number of occurrences
	Word *	5 x number of occurrences
	Word no*	2 x number of occurrences
Metadata location		
	Caption words	8 x number of occurrences
	Words-in-map	8 x number of occurrences
	Title words	8 x number of occurrences
	Referring sent.	8 x number of occurrences
Metadata frequency		
	Word	x number of occurrences

/H7/ Ranking by relevance of theme While the domain ontologies include different numbers of starred words, the frequency of term occurrence and metadata location are balancing factors. Rank first those matches in a category receiving the highest score, that suggests the greatest semantic relevance.

/H8/ Ranking by relevance of data attributes When two or more items are classed in the same category, rank them according to other data attributes. Suggested:

(a) Most recent publication date first, with date known before date unknown

(b) Color first, then grayscale, then black and white

(c) Highest resolution first.

6.4.4 Findings

Number of maps classified by theme	Number of automatic classifications agreeing with manual	Number of automatic classifications that are plausible	Percent classfications agreeing automatic with manual	Percent classfications that are plausible	Total percentage classifications that agree or are plausible
150	91	32	60.7%	21%	82%

Table 3. Findings when the training set is classified by theme

Findings show that it is possible to classify automatically 82% of maps in the sample into either the same subject categories or plausible alternative categories that a person would classify them, in answer to the third research question.

Iterative testing on the training yielding poor classification results for theme was traced to insufficient metadata indicating subject in particular items, and to insufficient words in the ontologies to match with the metadata and classify those items. The problem of some items having insufficient metadata to indicate subject will continue regardless of database size. The problem of insufficient words in the ontologies, however, will be corrected as

MapSearch is developed and its ontologies expanded. Hence, it is possible that classification results for theme might even improve when a larger collection is supported by more exhaustive ontologies.

More correct classifications were obtained when the entire article was considered than when the metadata were limited to caption, title and referring sentence. This implies that it is the entire article rather than the map that is being classified. Even more correct classifications were obtained when the initial domain ontologies were supplemented by Library of Congress Subject Headings. It is strongly believed that even larger hand-culled ontologies would improve retrieval further, and that this would be a profitable avenue for future research.

One untried but promising way to improve classification by subject would be to use the journal title. This could not be tested readily because journal titles were not included with the metadata. However, the indexers who performed the evaluation turned to the periodical title, so this might be a useful indicator of subject.

Domain ontologies generated automatically from Library of Congress Subject Headings did not perform as well as the domain lists built manually with stars. Even so, any ontology with good category indicators is preferable to none (which leaves relying on keyword search), and so continued research in the automated generation of ontologies is useful.

6.5 Summary of results

There was less ambiguity in classifying maps by region and time than by theme, so the results are more satisfactory. Relative weakness in automatic classification by theme might reflect to some degree the overlapping of theme categories and the insufficiency of terms in the domain ontologies, both of which could be improved by further tinkering and testing.

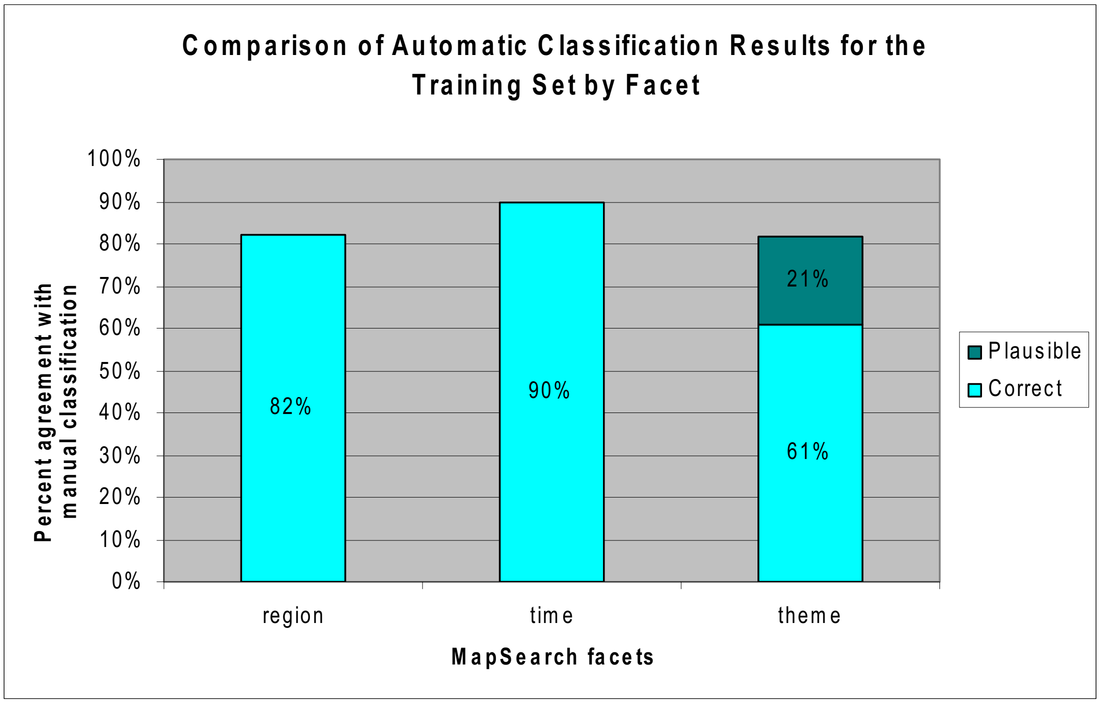


Table 4. Comparison of automatic classifications in the training set by facet

Results are measured in percentage correctly classified rather than the standard recall and precision because these percentages are easier to understand. In fact, the percentages are a measure of recall in that they represent the number of relevant classifications out of the total number of items. Calculating precision, or the number of relevant classifications out of the number of relevant + irrelevant classifications, is problematic as a measure of MapSearch efficiency. This is because the system was set up under the assumption that it was preferable to get a classification right than to miss it entirely (see Classification rules, section 5.3.1). This leans to recall rather than precision in the recall-precision balance in which raising one serves to lower the other.

6.6 Discussion

Problems particular to the classification of each facet are presented in Findings for Region 7.1.4, Findings for Time 7.2.4 and Findings for Theme 7.3.4. Commonalities among classification algorithms for each algorithm are discussed below.

Recall that the algorithms are composed of heuristics. Each heuristic by its very nature is based on probability, and will be correct only a certain amount of the time. This is important when considering the size of the training set required to hone the algorithms.

A larger training set would help determine which heuristics have the highest probabilities of success and which would help refine the algorithms. Yet, the advantages of improved algorithms would be diluted were the actual population of the database to be itself biased, which is likely. The ultimate population of a full MapSearch database will be contingent in part on publisher's rights which are not equally distributed over the disciplines. So a training set of 150 items seems well-conceived in random coverage and adequate in size.

In developing the algorithms for classification, the amount of data mined is key. Too little data will give insufficient terms to match with the domain ontologies, whereas too much data can introduce noise that will potentially match to the wrong domain ontologies. Many runs of the training set showed that better results were gained by restricting data harvesting to near the map when indexing region and time, but no so for subject. It was preferable for subject to scan the entire article before delivering a category, meaning that the subject of the article was considered to be the subject of the map. The assumption that the subject of map is the same as the subject of the article appeared valid in most cases. This assumption also governed how participants indexed the maps. Limiting metadata mined does not work in classifying by subject as it does in classifying by region and time because the restricted harvesting of metadata compounds the problems caused by the relative sparsity of subject indicators.

The next question in developing the algorithms is: should a category be assigned as a sort of compromise among all the data mined, or should it be assigned as soon as a match is found with the metadata closest and most relevant to the map? In the balance between too much or too little metadata, tests for region and time showed that it was better to prefer too little. This produced different situations per facet. To classify by region, there was never a lack of place indicator among the training set, so it seemed as though the metadata mined will be adequate. To classify by time, when no specific date or era was mentioned in the mined metadata, it was assumed that the time content was roughly the present, and so the publication date of the article could be used. To classify by subject, the mined metadata regions proved insufficient and it was necessary to search the entire article, essentially classifying the entire article rather than the map.

Another key to assigning classifications in one facet might be correspondences between facets. In other words, if a certain time period and subject (irrespective of region) tend to occur together, they might continue to follow the pattern and a rule could be established. The table with output for manual classifications for the 150 map-training set (some maps being assigned to more than one category) is one step to testing this assumption.

	Prehis	Antiq	Middle	EarMod	Modern
Agriculture	0	0	0	0	9
Archaeology & Anthropology	9	2	1	0	1
Arts	0	0	0	1	2
Commerce and Finance	0	0	1	7	9
History	0	0	9	14	9
Medicine	0	0	0	1	4
Military	0	0	0	0	10
Politics and Law	0	0	4	3	33
Religion and Education	0	0	4	2	6
Science	4	0	1	1	21
Society	1	0	0	2	17
Technology	0	0	0	1	12

Table 5. Time and theme assignments for the training set to suggest correlations between time and theme

The table lends some evidence to support the rule that items classified in Archaeology and Anthropology should be classified in the time period Prehistory, although it is seen that the rule will be correct only 69% of the time. Further heuristics that could be made based on the table are that Technology and Transportation could be classified in the time period Modern and it would be right 92% of the time.

The planned expansion of the collection of maps will necessitate an expansion of both the categories and the ontologies. Categories in region, time and theme will offer the user a finer selection, just as the category Modern was subdivided to offer a choice of decade. Subdivisions for place will amount to the political boundaries of country, and then city as found in Geonames. Subdivisions for theme may be taken from subheadings in the Library of Congress Classification. Each of the newly added subdivisions, in turn, will require its own domain ontology. The fact that the given domain ontologies aid classification into the given categories suggests that further expansion of domain ontologies along with categories following this method should be similarly successful.

7 Interface design

The MapSearch front end was developed for this dissertation to demonstrate the classifications of the back end. This chapter discusses the rationale behind the design of the MapSearch interface and results display. The tripartite division of facets was suggested by examination of user queries and has been tried in the past (Perry, Hakimpour & Sheth, 2006). Basically, the design was created from the top down.

The purpose of the prototype was to demonstrate the efficacy of automated classification only. Attention given to interface design should improve test conditions, but is a sideline to the core of this dissertation. This is why, although further evaluative testing might improve the interface, such testing has been postponed.

7.1 Method

The design of the MapSearch interface was informed largely by principles from design experts Shneiderman and Plaisant (2005) and Krug (2006), as well as by personal experience in previous interface design projects.[26] Knowledge of actual search systems such as the Yahoo directory, Google Image Search, and the Alexandria geolibrary and the National Geologic Database suggested further how controls might look on the main and the result screen, and how relevance options could be offered.

The preliminary design mockup on paper was similar to that in the current interface illustrated in Figure 5. Categories changed and subdivisions for the Modern time period were added. Preliminary searches with the controls on the small document test bed showed that even more display options were needed. The state of the system is described below.

7.2 Keyword search

Today's Google-raised searchers are accustomed to a simple keyword search box. To accommodate familiarity preference, therefore, the prototype includes a keyword box (please see Figure 5). The keyword search option in MapSearch yields more relevant results with

[26] A paper on Information Visualization published in Knowledge Organization (2007) 34(3), 128-143, and a proposal "Interface Lite" submitted with Professor Lesk to NSF in 2006

better recall than in comparable, non-ontology supported systems. The keyword search works by comparing the query terms to domain ontologies of all three facets. Items are retrieved in all facets when there is ambiguity. So, for example, the search string "John Quincy Adams administration" which could indicate either the Early Modern period, or a theme of politics under a certain president, would retrieve maps entered in both categories.

Unavailable in most interfaces is the ability to enter a query term and a browse category at once. Simultaneous use of both features allows the search to be refined in a way that is more specific to the collection than a user otherwise might know to construct.

7.3 Faceted category search

The categories will appeal particularly to those unfamiliar with database contents and those unclear as to what exactly they seek. Usability experiments show that users are uncomfortable when offered an actual thesaurus (Greenberg, 2004, p.117), but that they do prefer a classified approach when they lack in-depth knowledge of a field (O'Connor, 1978, p.152). These requirements are anticipated in MapSearch by offering subdivision selections arranged in a hierarchy.

Selections are offered as radio buttons. These are preferred to drop down menus because they are easier to use. Each button selected potentially will open another level of subdivisions from which to choose, making the design expandable. In the prototype, however, only the time category of Modern includes subdivisions because the small size of the test bed would leave further subdivisions empty. The radio button design prevents error, one of the "golden rules of interface design" of Shneiderman and Plaisant (2005 p. 75), in that buttons are not selectable when the limit as to the number of choices has been reached. Alternative to binary yes/no radio options would be to allow facet weighting. Such is the purpose of the slide bar, as demonstrated in the JSTOR advanced search. The added complexity to the MapSearch interface by introducing facet shading did not seem worth the subtle gain in relevance that presumably would result.

Alternative to a wholly semantic, radio button interface would be a visual interface that allowed selection of time along a timeline, or region within a map. Such a region selector is found in the GeoSearch engine of Geotags.[27] Another sort of visual display would allow the selection of an era along a timeline in addition to a keyword search (as in Google Labs new "view:timeline" command) or browse category. The option of uniting region or theme by time has not been adopted in MapSearch in acknowledgement of how it is believed users rely on words in information searches. Research in this area is needed. But what would be lost in combining visual with semantic search would be the balance among the facets, and the long-term potential to unify search parameters seamlessly with a broader document assortment.

Many users would appreciate knowing how many maps are contained within each category. The question is how to show this. If we put a number after each category, (Europe [553]), the significance of that number will not be immediately clear. The Alexandria Digital Library experimented with a map interface that showed the coverage and density of the maps in the collection by color, and also showed the footprint of a retrieved set. MapSearch follows the message model of the National Geographic Database that "Current selected criteria will find X publications."

7.4 Results display

It has been recommended for the sake of usability is to provide an overview first, with the capability of zooming in and filtering, and giving the details of each result only on demand (Shneiderman & Plaisant, 2005). MapSearch to some extent follows this recommendation. An overview of maps retrieved along with article title is provided with the many-map results display (six maps across the screen per line), as alternative to two-map wide or simple vertical results list. The map overview is enhanced by the map caption and the article title. Options provided that do not filter per se, sort results (such as by map sharpness or color variety) so that the most relevant are filtered to the list top. The resolution of the database maps in most cases is too low to offer the option of zooming in to adjust scale or panning from side to side to change the display focus, although there is an option to enlarge the

[27] http://geotags.com/frameset.html

thumbnail images. Based on the map thumbnails, article titles and, in some cases, the caption (see Figure 6), the user can decide whether or not to go the extra click to examine the map at full screen enlargement or to open the full text of the article that contains the map.[28] One key to usability is that the user is not drowned in details, but rather is offered enough background per item to decide whether or not to view any in particular.

Some sort options which would be appreciated cannot be achieved through data mining such as geodetic accuracy or date of map creation. Metadata on geodetic accuracy would require verifying author-supplied data map by map—challenging in peer review and expert cataloging let alone in this secondary application—so it cannot be a searchable criterion. One clue to measurement accuracy would be to know the date of map creation. Maps that were made in the 16th century could appear in a different category than modern maps that show that same region during the 16th century, for example. While this distinction in date of map creation is simple for the cataloger, it seems impossible to gather when relying on metadata harvesting.

Above the results display, the user could be oriented with a memory of the search. Both the mode of search (keyword or selection of a button) and the specific query are retained at screen top (User request: keywords > *language*). A "Return to search page" button allows users who have changed the result display options to go immediately back and try again so that they do not need to replay their past choices with the back button in the browser.

7.5 Feedback for the user

One of the eight golden rules of interface design is for the system to offer feedback (Shneiderman & Plaisant, 2005, p. 74). MapSearch responds politely in circumstances when the user might otherwise feel frustration: when the system slows and when no results are found.

[28] Most articles derive from open access journals or journals to which Rutgers subscribes. The articles are owned by the subscription service, but it is understood that this prototype is non-commercial, and that attaching these few articles constitutes no commercial loss to the subscription services and therefore should constitute fair use.

Consulting the article opens Adobe Reader and is likely to be slow. A user might even suspect that the system had malfunctioned. Selecting the link to the article, therefore, calls up a message to the effect that Adobe Acrobat Reader or some other .pdf viewer must be installed in order to read the document, and the user may have to wait while the document loads.

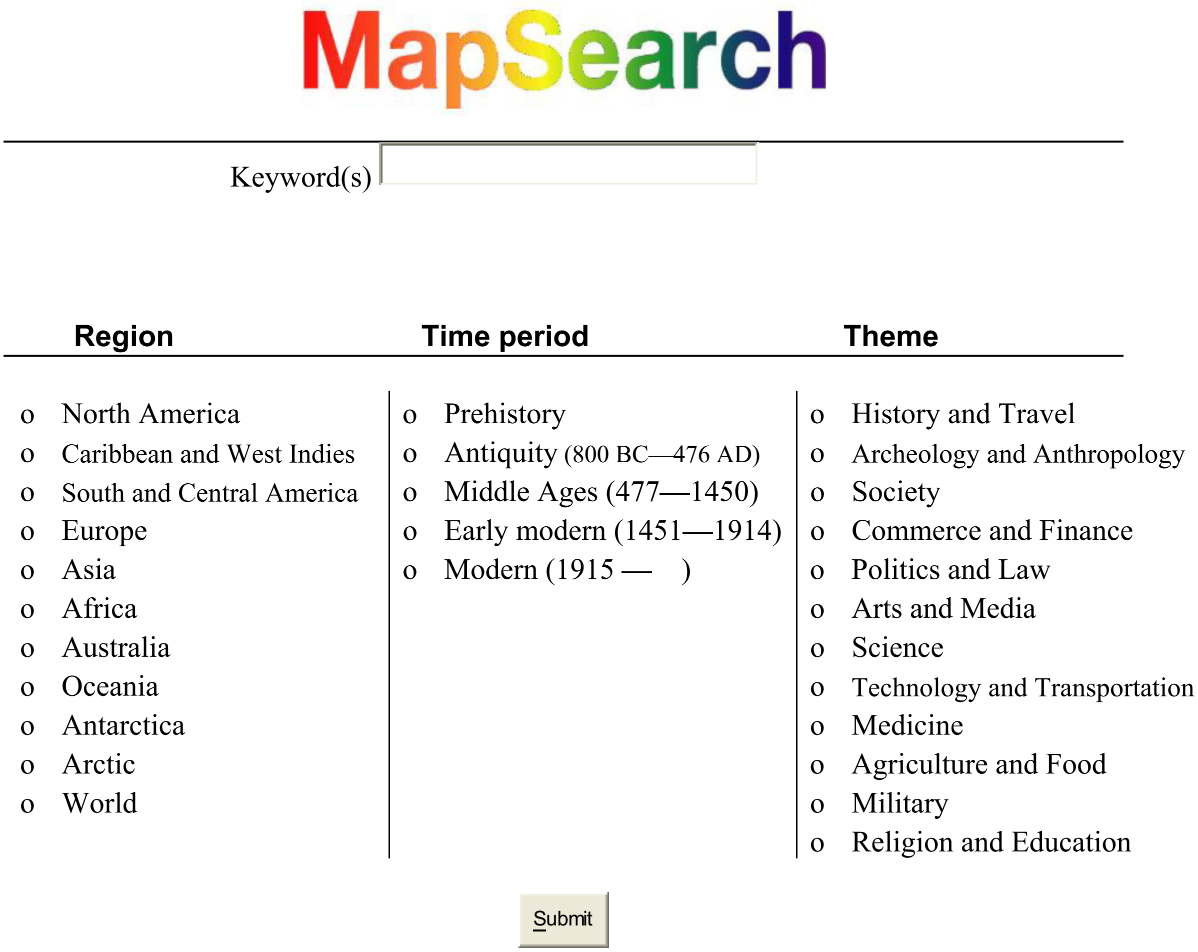
Entering both category and keyword often will over-specify results for the prototype's small test bed such that nothing will be retrieved. The user is notified that "No maps match the criteria specified. Please try fewer criteria." When a keyword is entered by itself, items should be retrieved because the keyword will be classified via ontology into a category, so that retrieved results, if not matching the keyword exactly, will be relevant. Instances in which the system is unable to classify the keyword will call up the message

Your search retrieved no results.

Please check the spelling,
Or try similar words
Or broaden search with fewer words
Or shorten term, using * for left-out letters
example: instead of *pharmacy* use *pharm**

7.6 Further testing

This interface design comes mostly from top down requirements and knowledge of others' design experiments. Requisite usability testing on MapSearch to suggest modifications and enhance the design must wait until a later stage of the project.



MapSearch

Keyword(s)

Region

o North America
o Caribbean and West Indies
o South and Central America
o Europe
o Asia
o Africa
o Australia
o Oceania
o Antarctica
o Arctic
o World

Time period

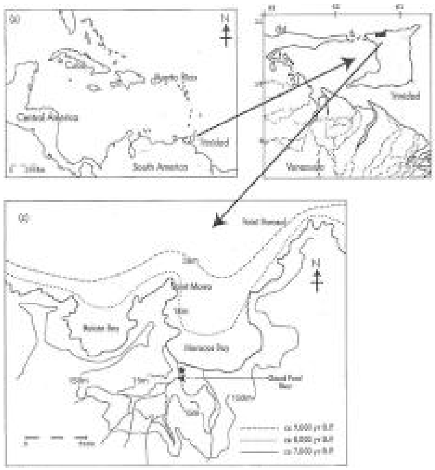
o Prehistory
o Antiquity (800 BC—476 AD)
o Middle Ages (477—1450)
o Early modern (1451—1914)
o Modern (1915 —)

Theme

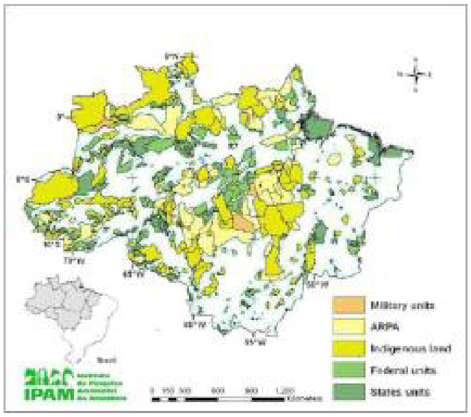
o History and Travel
o Archeology and Anthropology
o Society
o Commerce and Finance
o Politics and Law
o Arts and Media
o Science
o Technology and Transportation
o Medicine
o Agriculture and Food
o Military
o Religion and Education

Fig. 5 MapSearch interface—original at http://scilsresx.rutgers.edu/~gelern/maps/

...Large country labels and lists of artists per country dominate the colored map.
[Art Special: The Fairs]()



Map of Trinidad showing its general location within the Caribbean (a), northeast of the Orinoco River delta (b), the location of the Maracas Swamp, and (c) the position of the core taken from Maraca
[Holocene Development of Coastal Wetland in Maracas Bay, Trinidad, West Indies]()



Current and future planned protected areas (ARPA) in the Brazilian Amazon.
[Integrating Ecosystem Management, Protected Areas, and Mammal Conservation in the Brazilian Amazon]()

Fig. 6 MapSearch results display with large thumbnails for the category “South and Central America” —original at http://scilsresx.rutgers.edu/~gelern/maps/

8 Evaluation

This section discusses potential approaches for evaluating an information retrieval system. Following it describes the evaluation of MapSearch retrieval accuracy, with a chart showing accuracy of classifications by region, time period, and theme. In that there are numerous evaluation approaches as discussed in section 8.1, some alternate tests that could be performed are outlined, such as tests of automatic data harvesting (instead of hand cataloging), understandability of classification categories chosen and overall satisfaction with the system. Finally it is show that MapSearch passes easily what is probably the most important evaluative measure–cost-effectiveness—because this above all would justify costs of large scale set and implementation.

8.1 Potential approaches for evaluation

To evaluate an object is to estimate its value. The same object will be valued differently by different people. Value is modulated by the marketplace. Suffice it to say that markets may be strict (where the consumer enters a department store and pays the item sales price), or flexible (where the potential buyer sends a suggested bid to the Ebay seller, and the two negotiate a price). Just as in the economic market, in the technological marketplace there is no absolute value to measure.

Evaluation of information retrieval systems is a topic to which articles, books, courses and entire dissertations have been devoted. Any discussion here, therefore, must remain superficial. Methods of evaluation are wide-ranging. The January 2008, volume 44, issue of *Information Processing and Management* devoted to evaluation of information retrieval systems features articles which evaluate based on affective and cognitive search behaviors, interface design, and willingness to pay.

Despite the range of evaluation methods in the literature, information retrieval systems most often are evaluated on the basis of retrieval recall and precision based on a series of search tasks. Obviously, these are only some of many aspects on which one could evaluate. Here are some others:

Collection
    Internal factors: collection coverage and quality
    Comparison to other collections

System
    Internal factors: ontology referents, data mining, classification categories, time to produce results, recall and precision (retrieval relevancy)
    Comparison to other systems
User
    Internal factors: intuitiveness, ease-of-use, or cognitive load in interface use
    Comparison among user groups

Experiments show user satisfaction often is not highly correlated with traditional IR metrics (Turpin & Hersh, 2001). What then is the aspect that is most important to evaluate? "In the final analysis, the cost problem is of overriding importance in any operational situation, since the most effective system will not avail if the operations are too costly to be performed. However, costs are often difficult to measure…" (Salton, 1968). In short, evaluation is irrelevant if a system is too costly to implement.

A selection of experiments is proposed. Each experiment requires a hypothesis, a task, and a means to measure results. The final discussion assumes the worth of finding maps (map libraries justify their existence) in terms of a cost-comparison to present systems.

8.2 How well do the algorithms work?

The research questions concern the effectiveness of the automated classification of maps. The algorithms were established on a map training set, while they are evaluated on a test set. Participant indexing of the test set was used as a benchmark to determine how close MapSearch classifications come to indexers' assignments. The unit of analysis is the classification of each map.

8.2.1 Method for manual classification

The 55 maps in the testing set were selected randomly. The same method was used to collect these maps as was used to collect maps in the training set, with perhaps an even greater attempt in this smaller sample to vary the range of journal topics and limit the number of maps per article. While a larger testing set would increase validity, it was known that it would take time to index each manually, and so collecting a large number would be impractical.

Graduate students in library and information science were paid to index the 55-map sample. Two individuals were chosen particularly because they were available during the period that the test was being run. Both have professional indexing experience.

Participants were asked to index maps according to MapSearch categories of region, time and theme. A brief instruction sheet offered the rules of indexing (one or two assignments per category only) and the category labels. Categories for time are delimited by dates, but categories for region and theme are less precise. Therefore, materials to explain each category within region and theme were provided the indexers.

Participants were given the articles with the maps in each article flagged and a spreadsheet having the maps listed in the same order with blanks for categories of region, time and theme. These were accompanied by the description of the categories for each facet and rules for indexing (Appendix D1), and an elaboration of categories for region (Appendices D2 and D3) and theme (Appendix D4).

The items were ordered randomly, and that order was retained for both people. Retaining the map order made it easier to keep track of the maps. Any negative consequences that might result from keeping the same order, such as indexers fatigue causing significantly lesser accuracy by the final maps in the sample, or by the indexers becoming familiar with the process by the end to do a significantly better job by the final maps in the sample were mitigated by the small sample size, and allowing the indexers the freedom at the end of the study to return and re-think classifications for questionable maps. Each indexer filled out the spreadsheet on her personal laptop computer and worked at her own speed. One took 75 minutes and the other took 90 minutes to complete the task. The study was complete when all blanks on the spreadsheet had been filled in.

8.2.2 Assembling the results of manual classifications

The next step was the compounding of participants' categories. The choice of classification was unambiguous and both selected the same category for the majority of cases. When categories did differ, it was mostly in theme. All categories were included in the benchmark list when indexers' results differed, despite the fact that the rules for manual and automatic indexing assign each map a maximum of two categories per facet.

8.2.3 Evaluation of automatic classification

Finally, indexers' categories for the 55 maps were fed into MapSearch as the "right" answers, and MapSearch was asked to index the same. The same evaluation was used as in the training set in which partial credit was given to categories that can be said to overlap (Science and Medicine, for example). A comparison of MapSearch and indexers' categories gave the final results for MapSearch classification accuracy (Table 6).

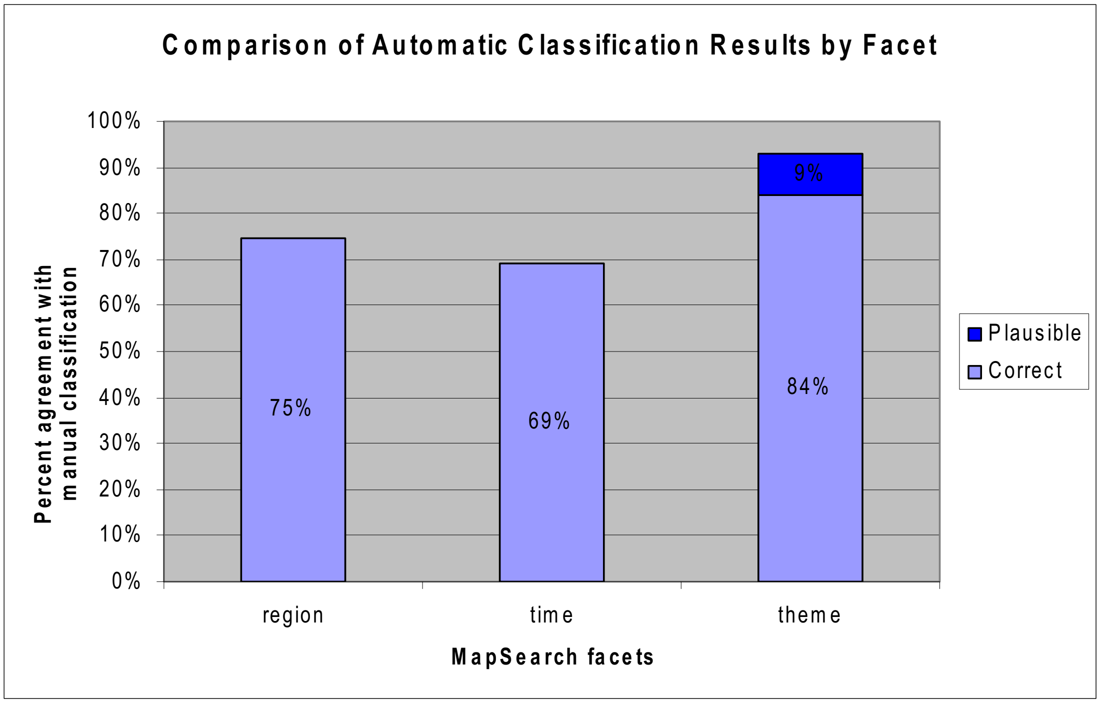


Table 6. Comparison of automatic classifications in the testing set by facet

The indexers themselves commented that, of the three facets, they had most difficulty assigning themes. Both relied on the article to assign a subject when the map seemed to be an illustration of the article, as was suggested in the instructions (Appendix D1). Another problem shared by the indexers was to assign time content to a map of historical sites when the map showed where the sites are today, while the sites themselves might fall in the categories Prehistory of Antiquity. Both elected to use the historical time period rather than the modern. Their difficulties seem to have been reflected in the results of the classification done automatically.

Compare the evaluation results of the automatic classification for the testing set (Table 6) to the results of the automatic classification for the training set (Table 4). Similarity suggests that adjustments to the algorithm made during preliminary stages were not the result of overfitting to collection particulars. These adjustments were for the most part well-taken and had some generalizabilty that genuinely improved the program.

Differences among automatic classification results between the training and testing document set might be attributable to a number of factors. Sampling practices in creating the training set and the testing set were similar, although even more effort was put on making a heterogeneous mix of maps in the testing set than in the training set. The manual classification of maps might also cause some measure of difference. This could be because, for the training set, the person who actually created the categories was also assigning the categories. Those assigning categories for the testing set were less familiar with the categories and so might not have assigned them in the same way, although the materials offered indexers (Appendices D2, D3, and D4) attempted to allay these problems.

In particular, the largest difference between training set and testing set results is in the time category, which dropped more than 20 points from 90% for the training to only 69% in the testing. A detailed analysis of the results has not been done at this point because, for the moment, the algorithm cannot be changed. However, it is possible that the two maps showing the earth 50 and 250 million years in the future could account for some of this difference. These two maps presented a problem less because of the algorithm than because no time category was had been created for Future – which, in the classification of these maps, flummoxed both the indexers and the algorithm. The region and theme facets exhibited far less difference between training and testing results. Between the training 82% and testing 75% for region, there was only a 7 point difference. And the 23 point increase for theme between the training 61% and the testing 84% is probably attributable to the way the manual classification was assembled. Different answers from each indexer were equally acceptable, making a possible four categories correct in ambiguous cases, whereas researcher-assigned categories in the training set were in all cases limited to two.

8.3 Proposed experiment for data mining

Hypothesis Harvested metadata works as well for information retrieval as manually entered metadata in either a standard spatial metadata schema or a general metadata scheme such as Dublin Core.

Participants Two professional catalogers or master's students in library and information science

Task One participant catalogs 100 maps using ISO 19115 for geographic information. The other will judge the relevance of these cataloged maps retrieved in response to given queries.

Protocol 100 hand-cataloged maps are fed into MapSearch, and 10 search queries (mixing browse and keyword search) are asked of the maps. Results are recorded. The same 100 maps are then fed into MapSearch and metadata is harvested using the MapSearch data mining algorithm.

Measurement The participant-as-judge will examine results of the 10 queries for recall (were all of the maps that should have been called up indeed called up), and precision (are all the maps that were recalled appropriate for the given search). Outliers, or data points with some value, should be differentiated from noise, data with no value.

Analysis The relevant results from the mined maps should be roughly equivalent to those from the hand-cataloged maps to demonstrating the adequacy of the mining procedure.

8.4 Proposed experiment for classification categories

This is an evaluation of categories themselves rather than the retrieval system. The analysis of librarian-reported and actual map-related queries suggested the three facets. The question becomes: are the subdivisions of the facets in line with user expectations?

Hypothesis Terms used for classification subdivisions will match user expectations.

Participants 50 people, ideally mixing ages and ethnicities to highlight cultural differences in vocabulary usage

Task Each person will provide a phrase or sentence definition for each category, and assign five maps to each category

Protocol All participants can work simultaneously

Measurement The experimenter should check categories for wide discrepancies in definitions and map assignments. Any categories that vary widely among participants are candidates for re-naming. The work of Furnas, Landauer, Gomez and Dumais showed that the likelihood that many will assign the same name to the object is less than one in five (1987, p. 966). The theme categories suggested for the maps should therefore be assessed not on the basis of whether the "best" term had been chosen, but only whether the category labels selected are understood widely.

Analysis Any category labels in MapSearch that are not well understood should be changed.

8.5 Proposed experiment for user satisfaction

Research question How do people view MapSearch?

Participants 30 people, ideally of varying ages and backgrounds

Task Ten spatially-related questions would be written. Eight of the ten would be answered with MapSearch, and two would be answered with any resources but MapSearch. The two non-MapSearch questions act as a control.

Protocol Participants would be timed in answering all 10 questions. Each then would be asked a series of evaluative questions about the interface following the survey instrument template in Shneiderman and Plaisant (2005, chap. 4).

Measurement Time and accuracy of responses would be measured and survey answers tabulated to arrive at an average satisfaction level for MapSearch.

Analysis Comments about the interface should be considered in an attempt to improve the interface. After the interface is modified, further comments should be elicited.

8.6 Cost effectiveness

The hypothesis is that the MapSearch system will be more cost-effective than a system based on manual cataloging.

Reasoned proof of this hypothesis is clear from the following ordered assumptions:

1) Cataloging and classification accuracy of items must be measured per unit time.
2) The cost of automatic cataloging is almost nothing (past the cost of creating the program in the first place), whereas the cost of paying a person to do the cataloging is charged by the hour
3) The cost of hand-cataloging grows as the collection grows, but the cost of automatic cataloging remains almost nothing as the collection grows
4) The cost of cataloging is repaid as items are examined
5) Even if fewer relevant items are retrieved per automatic record than per manual record (because recall and precision are less), users will "satisfice" and find something useful per search
6) Fewer items are actually looked at as the collection grows.
7) The financial gap between automatic and manual processing widens as the collection grows.

Therefore, automatic cataloging must be more cost effective than hand cataloging.

8.7 Importance of testing

Design should be an iterative process. Leidner identifies sought-after factors in language processing software to include: efficiency, accuracy, productivity, flexibility, robustness, and scalability (2007, p. 157, quoting his own earlier work). Martins, Silva and Chaves point out that performance factors presumably such as Leidner's accuracy and productivity are often not correlated with ratings from user interaction (2005, p.68). It takes only a few participants in pilot studies (Krug, 2006, p.134) to yield valuable insight for designers. User input on various system factors is perhaps less valuable for passing “objective” judgment than it is for giving designers feedback for system improvement.

9 Limitations

Some aspects of the prototype limit its usefulness. This section points out which limitations will vanish when the system is scaled up – such as a small map collection and incomplete ontologies for subject and time period – and which limitations will remain as the best possible alternative.

9.1 Collection, current and expanded

Maps have the potential to be useful without the entire article, even though MapSearch provides the article full text. The programs that will mine maps and separate basemap from the text layer are still being developed. Assuming that these programs are effective, it will be possible to increase the number of maps from .pdf open access journals more easily. A bit of further programming then will allow the programs to extract maps and articles in different file formats. This would allow the system to hold a more heterogeneous collection. A wider collection, in turn, offers users more options.

Even a very large map collection will not necessarily hold the right map for all MapSearch users. One theory is that people "satisfice" (Simon, 1976/1997), sacrificing their original search goals to become satisfied with results available. If the theory is true, most MapSearch users will find something relevant. For those who do not, an option to make your own map could be added. Websites such as GeoCommons[29] supply basemaps and overlay data sets separately to allow users the possibility of creating a new map, often called a mashup. Such a link added to MapSearch would widen possibilities for users.

9.2 Categories

Any selection of facets or subdivision categories directs search toward those categories and away from other legitimate categories. Were the right facets chosen? It could be tested whether region, time and theme codes apply to very large numbers of user queries. Further, it could be tested whether the facet subdivisions might be applied to actual user queries. Subdivisions alternatively might come from the users themselves by encouraging the users to

[29] GeoCommons had been on the web at http://www.geocommons.com but has been temporarily disabled in spring 2008 to create the site's next generation.

add subject tags to maps and taking those tags for category labels. The sum of the categories created by users is called a folksonomy (Sturtz, 2004). Typically, folksonomies are single layered, rather than hierarchical, so were a folksonomy to be implemented, a different sort of interface would be needed to accommodate a new category arrangement.

9.3 Classifying maps

The algorithms are meant to classify some general map, but each unseen map is an individual. It has been pointed out in section 6.6 above that a larger training set would have allowed the algorithms to generalize to an even larger percentage of items unseen. But it was also pointed out that increasing the generalizability of the algorithms would be wasted on a collection that is biased in scope.

The classification rule of how to assign items to MapSearch categories could be viewed as a limitation, but it is probably the best option available. Presently in MapSearch, an item is assigned to the one or two categories in which it scores the highest. Different results would obtain were items to be assigned to a single category only, or to be assigned to every category in which they score. If items were assigned to every category in which they score, the category boundaries would be vague rather than clearly defined. But category labels are words. The meaning of words in natural language is not absolute, but “fuzzy,” hence the use of fuzzy logic in computing with words (Zadeh, 1999). This logic is overtaken by the reality of automatic metadata harvesting. What would happen if fuzzy logic were implemented would be that non-relevant metadata taken automatically would cause items to appear in categories in which they do not belong. The end result would be that precision would decrease substantially.

9.4 Ontologies

The number of words and the specificity of words in a controlled vocabulary both influence the information retrieval properties of the vocabulary, as mentioned in section 3.1 above. MapSearch employs a full ontology for region but its ontologies for time and theme are abbreviated. Although time and theme ontologies are adequate for the several-hundred map collection of the prototype, they would be inadequate for a more complete collection.

The method of combining Library of Congress Classification classes and subdivisions with Library of Congress Subject Headings could be used to create a more extensive ontology for subject. A systematic method to create a more extensive ontology for time period should be devised.

Controlled vocabularies must be kept current to optimize their retrieval benefits. But presently, the MapSearch ontologies are static. The time and theme ontologies would need to have new terms added and ambiguous terms removed regularly, and the gazetteers maintained on the web that were used for the region ontology should be uploaded periodically. Methods to update ontologies automatically would require further research.

9.5 Interface design

The keyword and facet search, display options and user feedback should be submitted to testing before being considered adequate to the task of mediating between map collection and users. Aspects of the present design, therefore, might be limiting to some.

The present method of semantic query entry is less precise for time and theme than would be a graphical query entry. Time period could be entered along a sliding scale, accurate to the nearest month, day or even hour. Nothing would be gained, however, with time queries entered more precisely than year because there is no comparable precision in the item metadata. On the other hand, region queries could be entered by sizing a bounding box, and the user could specify whether items wanted were contained in, near, touch, or overlap the given area. Gains in information retrieval would result. Entering a query so that the map requested is “contained in” a region, for example, would retrieve maps of smaller scale than the region requested. This option could work in concert with the ontologies too. The disadvantage of using graphical input for region is that it contrasts with semantic input for time and theme, and so the interface would no longer balance and the facets would seem to lose their equivalence.

9.6 MapSearch is a question-answering tool—but what are the questions?

Automatic assessment of data quality in a map is nearly impossible. So questions answered using the maps are only probably correct. When the map collection is combined with maps that have been cataloged manually, cataloger data on accuracy should be retained. The FGDC standard includes information about data quality, horizontal and vertical accuracy, and data source.

A debate raised during the making of the Alexandria geolibrary applies here: should the system find maps, or should it answer the user's questions? (M. Goodchild, personal communication, March 24, 2008). The ideal surely is to answer questions. This might be accomplished on the highest level by including in the interface an "Ask a Librarian" button. Selecting the button would link a user to a map librarian in real time, with the system giving both the ability to view the same MapSearch item at the same time. A lesser alternative would be to broaden the collection beyond maps to include documents and reference sources. The National Geologic Map Database is one example of such a collection. Then a question such as "what is the length of the Susquehanna?" might be answered not by retrieving a map of the eastern United States and calculating river length, but by retrieving a descriptive document about the river or a comparative table of the extent of rivers.

10 Contribution: what is new and why it matters

This dissertation has contributed in the domain of information science, and to a lesser extent, geographic information science and library science. The following subsections discuss a hybrid method for automatic classification, an expandable subject ontology alternative to WordNet, classification algorithms for region, time and subject, and the potentially expandable MapSearch prototype, and why each matters.

10.1 Hybrid method for classification

Classification typically is performed wholly automatically. Documents are classified with respect to each other, or clustered. The clusters change as the database contents expands, so it is often performed on-the-fly. Here, items are classified with respect to pre-figured manually drawn categories, and are assigned to categories with the help of manually devised ontologies. The method proves robust.

The advantages of the hybrid method are that the classification can be done as soon as items enter the database, so for large scale collection it has the potential to speed query processing. The other advantage is that the ontologies improve classification results. Ontologies improve recall because they find synonyms, and they improve precision because their choice of synonyms help to disambiguate ambiguous terms.

10.2 Subject ontology

The reigning general purpose subject ontology presently is WordNet. The current research shows that the ontology constructed by mixing Library of Congress Classification Headings and Library of Congress Subject Headings is quite useful. Here, it is on only one level. But with the addition of lesser Classification subdivision headings and additions Library of Congress Subject Headings, it could be greatly expanded. Moreover, the Classification subdivision are clear break points between domains, and with the addition of headings, domain ontologies could be created for information retrieval in specific subjects.

A straightforward method to create ontologies matters greatly because ontologies are the essence of intelligent information retrieval. Web Ontology Language seems to be stalled in an early stage of development. Were a robust general-purpose ontology to be easily creatable, however, we are quite close to intelligent information retrieval on the World Wide Web.

10.3 Algorithms

Two types of protocol that could be used for maps or adapted to other document types are made available for the information science community.[30] The protocols cover data mining for items in articles, and automatic classification by region, time and theme. These algorithms, with only minor adjustments, could be applied to extracting and classifying other sorts of documents. Of particular interest are the heuristics added to the problem of toponym resolution, or how to determine which place is meant in a document.

The scope of this problem is enormous, in that many web pages have geographic indicators. The commercial MetaCarta software resolves place names, and research projects by Amitay et al. (2004), Leidner (2007), and others are making strides to improve results. Further refinements should nonetheless be welcome.

10.4 MapSearch prototype expandable

A contribution to geographic information retrieval rests in the faceted search system to find maps by region, time period or theme as well as the standard keyword. It is hoped that others who have collections will benefit directly from this work by feeding their maps in MapSearch, and the larger collection, in turn, would make the resource more valuable to users. To this end, the possibility of working with an industrial company to create a large-scale version of the prototype is being explored.

Thousands of journal articles have maps. The potential scope of the MapSearch collection is vast, which would make this in itself a useful tool. Barriers to database expansion are not in method, but in expanding the ontologies, subdividing the classification categories, and in legal permission.

[30] The Perl scripts probably will be mounted on or linked to the website of the MapSearch prototype.

10.5 Toward a simpler spatial metadata scheme

A preliminary study on how people ask for maps suggests that the rich spatial metadata schemas that are used in many countries of the world could be thinned substantially and still be useful for information retrieval, while being faster and easier to implement manually. The indexing points coded in the region—time—theme studies must be confirmed by larger query samples. To this end, additional queries have been requested from the Internet Public Library.

Should spatial metadata schemes with fewer fields become standard, many more maps would be able to be manually cataloged quickly. Simplifying the metadata scheme would thus retain access to current maps and probably increase greatly the number of maps that are known from catalog records. The 12-field Denver Core as the minimum searchable set of the FGDC scheme has yet to catch on, but this research makes it more attractive.

11 Future Research

Studies of retrieval effectiveness should be done with a much larger database of .pdf items in order to confirm effectiveness of the ontology-mediated classification method. The ontologies will need to be expanded along the lines that have been established here. The next stage will be to include items in other formats besides .pdf which will be found in other sources. This will require aspects of the information fusion problem to be solved such that data from multiple sources can be processed simultaneously. Future research directions concern adding complexity to the system side and adding features to the user side. The concluding research direction proposes that MapSearch be used as a model for creating an analogous system for other specific types of graphic data.

11.1 Expanding the search protocol in order to expand the collection

The information fusion problem, or search among diverse databases simultaneously, is the subject of conferences, a journal and an international society.[31] Collecting, or fusing resources from different sources is important because the larger and more heterogeneous the collection, the more potentially relevant will be the results per search, and the more potentially useful will be the tool for more people.

Substantial difficulties involved in collecting resources from different sources are technical, linguistic, legal and financial. Technical: Some digital libraries have their own search protocols and items cannot be looked at from other interfaces; different kinds of static or dynamic maps might require display by particular browsers. Linguistic: Ontologies need to be available in any non-English language that the system retrieves maps from, and the ontology domain must conform to the data domain in terms of its generality or specificity. Legal and financial: Some databases are proprietary and restricted such that it is necessary to secure legal access and to pay for the right to search among holdings. In acknowledgement of permissions, an expanded database probably would include not the actual article, but instead a link to the publisher site. MapSearch users would then enter to get the full article, if needed, or else they would be given publisher instructions as to how to subscribe

[31] International Society of Information Fusion, on the web at http://www.isif.org/

11.2 Adding features for users

Features for user interactivity should be added as the database grows. The browse categories will need to be subdivided such that, for example, regions will subdivide by country and possibly by city. Options should be added to search by map type (vector or raster), or file type (.kml, .html., for example) not just the .jpg used in the prototype. It will have to be determined whether options will be selected pre-search on the home screen, or post-search along with the controls on the results display screen.

Classification need not be hierarchical. Automatically created groups may be clustered on a single level. For the user to keep track of all the clusters of a collection, there must be fewer clusters. The larger the collection, the more limited would be the number of clusters, and so the less precise would be the grouping of each. Also, clusters would change as the database contents would change, making pre-set ontologies useless in mediating between cluster and query, thereby erasing the possibility of employing ontologies to enhance retrieval relevance. Creating new clusters also would prevent the user from gaining familiarity with the scheme.

Ranking items according to relevance in more than one category is an under-researched problem. One way to show relevance among different categories is with sliding scales. Sliding the scales manipulates the degree of relevance among categories, but can be hard for the user to understand. Another way would be to show relevance graphically in a chart showing the interrelationships. Yet another way would be to pre-determine how each of the different factors should rate and fold that into the retrieval algorithm by weighting factors differently so that items would be retrieved in that ranked order. For example, maps with a keyword matches with the query could be worth 50%, maps from a recent publication could be worth 25%, and map with color variety could be worth 25%. But this would give the user less flexibility, and also the reason behind the results listing would deviate from a direct match with the query entered and so the user would likely find the reasoning behind the result listing obscure.

MapSearch is expected to be a tool for research foremost, but were it to gain popularity, an option might be added for interactivity. An “Add your own label” button on the results display window would allow users to enter their own social tags. These tags, in turn, could be used to improve relevance of maps retrieved by keyword search.

Ultimately, a “push” component with the system pushing results to the users could be added to this “pull” technology that requires users to pull results out of the system by entering queries. The “push” or alert would help the user keep up with an expanding collection. The user could subscribe for a MapAlert, telling the system what region—time—theme combination is of on-going interest. Then the system would push maps to the user whenever a map meeting pre-set criteria became available.

11.3 MapSearch as a model

Google custom search engines devoted to a field of study (economics) or a sphere of influence (U.S. government) are often preferred by a particular community.[32] MapSearch resembles more a search by data type such as search for audio or image files, although maps come in a wide assortment of file types. Yet, creating a system that mines data and performs automatic classification following the MapSearch model will recommend itself to a particular community. For example, genealogists want cemetery records, geologists seek data about rocks and soil; ornithologists want bird locations, oceanographers need water temperature, mineral content and depth; astronomers appreciate photographs illustrating galaxies, and chemists consider graphics showing compounds and reactions. One very long term goal would be to use MapSearch as a template search system that could manipulate very specific data types within a single interface.

The conclusion of a survey conducted by the University of Michigan in the early stages of formulating its Open Archives Initiative (OAIster) catalog here:

> One user commented: “You will never beat Google. No way.”
>
> Probably true. However, it's not our intention to beat Google, but to provide an adjunct method for accessing information online. Our hope is that by providing a comprehensive service that caters to user needs – e.g., finding resources by subject, finding resources by format,

[32] http://www.google.com/coop/cse/examples/GooglePicks

> retrieving the full resource – and addresses multiple searching problems, we can provide access to more, and more varied, useful and informative digital resources that are currently difficult to find.[33]

[33] http://www.oaister.org, retrieved November 2007.

Appendix A. Glossary of terms in the dissertation

BAG OF WORDS Words "bagged" or harvested for use as metadata, such that the words may be parsed for their meanings independent of each other

CLASSIFICATION = CLUSTERING Aggregating or grouping

DATA MINING Extracting particular types of data in order to find patterns

DECISION TREE ALGORITHM Model in which decisions are taken at every decision-point, like the branches of a tree, in order to reach a larger decision or a classification

DISAMBIGUATION = RESOLUTION Clarifying or removing ambiguity

FACET Each leading division on the same hierarchical level is a facet of the whole (Compare to HETERARCHY OR ARRAY)

GAZETTEER Geographical dictionary. A typical gazetteer will include the place name, how the places relate to one another hierarchically, and the latitude and longitude coordinates of the *centroid* or mathematically computed geometric center of each region

GEOLIBRARY Digital library of spatial data

HETERARCHY (HETER= DIFFERENCE OR CONTRAST) = ARRAY elements share equivalent horizontal positions in a hierarchy. (Compare to FACET)

HEURISTIC Approach to solving a problem that has no provable justification but that has been found to work

INDEXING Something (here a term) that points to something else (here, a map), just as the index finger is used to point (Compare to WEIGHTED INDEXING)

INTERFACE The sum of the screen choices and layout of those choices used to interact with the system

METADATA Data about data is the customary definition, where the first "data" is meant words of description, and the second "data" is meant an object or file or item to be described. In the dissertation, data mined from articles becomes metadata to describe the map

NATURAL LANGUAGE PROCESSING concerns how a system delivers meaning from words

ONTOLOGY Classification of knowledge that includes interrelationships among the terms

PARSE To go through text to weed out noise and separate words into those that can and cannot be mined for knowledge, for example, geo-name/not geoname, date number/not-date number

POST-COORDINATE INDEXING When a document is assigned the terms such as “Antarctica” and “Modern” but no relationship is assigned to those terms. The user has the option of conducting a search that finds documents that include one or both of the terms. As opposed to pre-coordinate indexing in which both terms are combined into one subject heading. Retrieved February 2, 2008 from http://web.njit.edu/~robertso/infosci/pre-post.html

PRECISION measure of the ability of a system to present only relevant items

$$\textit{Precision} = \frac{\textit{number of relevant items retrieved}}{\textit{total number of items retrieved}}$$

See http://trec.nist.gov/pubs/trec15/t15proceedings.html, Appendix for the NIST Special Publication SP 500-272 for in-depth discussion of precision

QUERY The user’s browse selection or keyword term(s) for the system to search

RANKING FOR RETRIEVAL = Order in which relevant items are listed

RECALL a measure of the ability of a system to present all relevant items

$$\textit{Recall} = \frac{\textit{number of relevant items retrieved}}{\textit{total number of relevant items in a collection}}$$

See http://trec.nist.gov/pubs/trec15/t15proceedings.html, Appendix for the NIST Special Publication SP 500-272 for in-depth discussion of recall

REFERENT The source of the ontology

RELEVANCE Item(s) that answer the user query

STOP WORDS Commonly-occurring words in a language such as prepositions, pronouns and conjunctions that are excluded in a search engine because they do not contribute to relevancy

TEST DATA SET Items used to evaluate how well rules predict classifications (Compare to TRAINING DATA SET)

THEME MAP Visualization that shows both location (map) and subject (theme)

TRAINING DATA SET   Items used to induce classification rules (Compare to TEST DATA SET)

WEIGHTED INDEXING   An indexing mode in which some terms are assigned higher value for indexing than others because those terms are considered more useful for prediction.

Appendix B. Instructions to mine data from around map

A. Caption

1. Locate caption: Scan directly below map and directly above map for text that is different size that the article text. Also scan the side of the map, if the map does not take the entire width of the page

2. Mine data: Take entire caption, either above or below map, from start to end of different size text. The caption does not necessarily end with a period.

3. If "Source" or "Reproduced" or "Reprinted" or "©", "Courtesy of" "By permission" or "permission" appears in the caption, do not mine these word/symbols or the text that follows.

B. Title and subtitle of article

1. Look at the beginning of the article for the word or words that are larger than the article text

2. Mine entire word string until the word "by"

C. Referring sentence: mine the sentence that refers to the map

1. Locate referring paragraph by scanning the first word of the caption (see A2.)

a) Broaden first word to match. If data is associated with "Figure 2", for example, search within article for "Figure 2" or "Fig. 2" or "Fig 2" or "fig 2". For Fig 2a, search Fig 2a and Fig 2(a) and Fig 2(A). If data is associated with "APPENDIX B," search for keyword "Appendix B," etc. If data is associated with "Illustration," search for keyword "Ill." or "Ill" or "Illustration".

b) Fig 1 might match with Figs. 1-x, or Figs. 1, x or Figures 1, etc.

c) If step 1a) and 1b) find nothing, search for keyword match by pairs within words of caption. Search for (non-geographic word+non-stop word) pair. If this finds nothing, search for an exact match for non-stop word pair (w2+w3) or (w2 + stop word + w3) in the caption. Continue with this pattern until every pair in the caption has been used to query the text. If a keyword match to caption pair is found in the article text, harvest this sentence in lieu of a referring sentence. See step 3 as a caveat.

2. Do not mine as specified in steps C1a, b, or c if
    - i. there are more than one referring sentences. Mine only the first sentence that appears in the article.
    - ii. exact match is found in a footnote
    - iii. exact match is found in List of Figures, Table of Figures, Table of Contents or List of Illustrations
    - iv. if near exact match such that Fig. 3 matches with Fig. 3.1

Stop harvesting metadata unless C procedures yield nothing, then go to D.:

D. Abstract or beginning of article

1. Harvest up to the first period, if present OR
2. Harvest entire first paragraph

Stop harvesting metadata.

Appendix C. Ontology Building with Geonames

Geonames Feature Codes
http://www.geonames.org/export/codes.html

First Pass
ADM1 first-order administrative division a primary administrative division of a country, such as a state in the United States
ADMD an administrative division of a country, undifferentiated as to administrative level
PCL political entity

- * (removed) PCLD dependent political entity
- * PCLF freely associated state
- * PCLI independent political entity
- * PCLIX section of independent political entity
- * PCLS semi-independent political entity
- * PRSH parish an ecclesiastical district

TERR territory
ZN zone
ZNB buffer zone a zone recognized as a buffer between two nations in which military presence is minimal or absent

RGN region an area distinguished by one or more observable physical or cultural characteristics

PPL populated place a city, town, village, or other agglomeration of buildings where people live and work

Appendix D. Materials for evaluation
D. 1 Instructions to Indexers

I. Indexing of maps.
Please assign each (flagged) map to *one* category (or two if the map fits equally well in both) in each facet of region, time and theme. If you cannot determine a map time or theme, please refer to the article context, because the labeling of the map and its raison d'être hail from the research article.

II. Categories

REGION

North America
Caribbean and West Indies
South and Central America
Europe
Asia
Africa
Australia
Oceania
Antarctica
Arctic
World : Apply to maps of the world, or to maps showing more than two regions

TIME

Prehistory (— 801 B.C.) includes geologic eras
Antiquity (800 B.C.—476 A.D.)
Middle Ages (477—1450)
Early Modern (1451—1914)
Modern (1915—) Leave subdivision blank except if the map is assigned to Modern: In which case, assign one or two of the following (and the earliest decade that is applicable):

World War I (1915 —1919)
1920s
1930s
1940s
1950s
1960s
1970s
1980s
1990s
Current (2000—)

THEME

Arts and Media
History and Travel
Archaeology and Anthropology
Society
Commerce and Finance
Politics and Law
Science
Technology and Transportation
Medicine
Agriculture
Military
Religion and Education

Appendix D. Materials for Evaluation

D.2 Regions expanded

I. North America	Canada
	Mexico
	Saint Pierre and Miquelon
	United States
II. Caribbean and West Indies	Antigua and Barbuda
	Netherlands Antilles
	Barbados
	Bermuda
	Bahamas
	Cuba
	Dominica
	Dominican Republic
	Grenada
	Guadeloupe
	Haiti
	Jamaica
	Saint Kitts and Nevis
	Cayman Islands
	Martinique
	Montserrat
	Puerto Rico
	Turks and Caicos Islands
	Saint Vincent and the Grenadines
	British Virgin Islands
	U.S. Virgin Islands
III. South and Central America	Anguilla
	Argentina
	Aruba
	Bolivia
	Brazil
	Belize
	Chile
	Colombia
	Costa Rica
	Ecuador
	Falkland Islands
	French Guiana
	South Georgia and the South Sandwich Islands
	Guatemala

	Guyana
	Honduras
	Saint Lucia
	Nicaragua
	Panama
	Peru
	Paraguay
	Suriname
	El Salvador
	Trinidad and Tobago
	Uruguay
	Venezuela
IV.	
Europe	Andorra
	Albania
	Armenia
	Austria
	Aland Islands
	Bosnia and Herzegovina
	Belgium
	Bulgaria
	Belarus
	Switzerland
	Serbia and Montenegro
	Cyprus
	Czech Republic
	Germany
	Denmark
	Estonia
	Spain
	Finland
	Faroe Islands
	France
	United Kingdom
	Guernsey
	Gibraltar
	Greece
	Croatia
	Hungary
	Ireland
	Iceland
	Italy
	Jersey
	Liechtenstein
	Lithuania
	Luxembourg
	Latvia
	Monaco
	Moldova

	Montenegro
	Macedonia
	Malta
	Netherlands
	Norway
	Poland
	Portugal
	Romania
	Serbia
	Russia
	Sweden
	Slovenia
	Svalbard and Jan Mayen
	Slovakia
	San Marino
	Ukraine
	Vatican
V.	
Asia	Afghanistan
	Azerbaijan
	Bangladesh
	Bhutan
	China
	Georgia
	Hong Kong
	Indonesia
	Isle of Man
	India
	Japan
	Kyrgyzstan
	Cambodia
	North Korea
	South Korea
	Kazakhstan
	Laos
	Sri Lanka
	Myanmar
	Mongolia
	Macao
	Maldives
	Malaysia
	Nepal
	Philippines
	Pakistan
	Palestinian Territory
	Singapore
	Togo
	Thailand
	Tajikistan

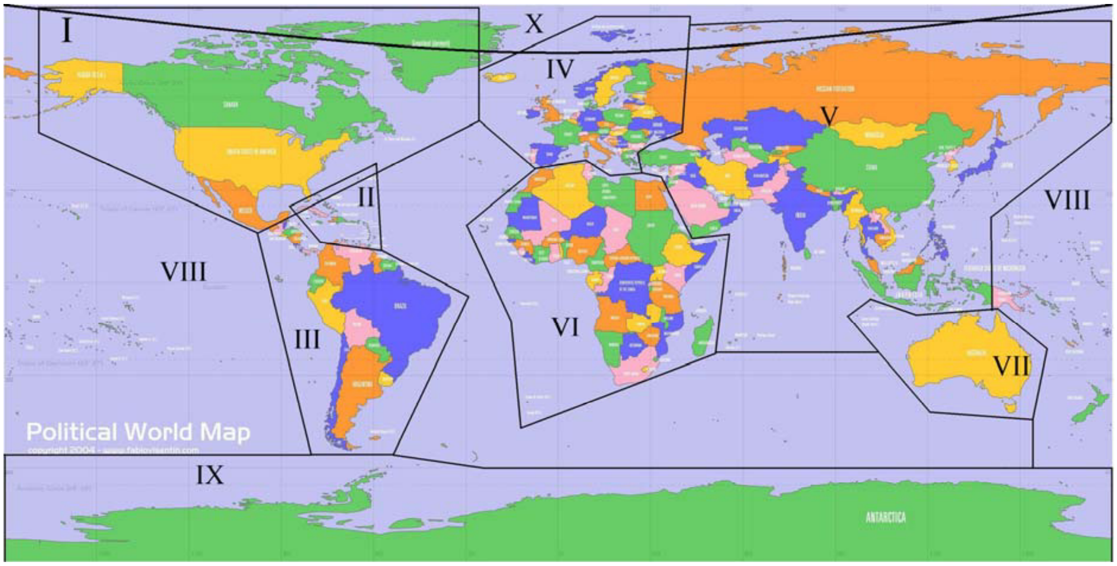
	East Timor
	Turkmenistan
	Turkey
	Taiwan
	Uzbekistan
	Vietnam
	Yemen
	United Arab Emirates
	Bahrain
	Brunei
	Israel
	Iraq
	Iran
	Jordan
	Lebanon
	Oman
	Qatar
	Saudi Arabia
	Syria
VI.	
Africa	Angola
	Burkina Faso
	Burundi
	Benin
	Botswana
	Congo - Kinshasa
	Central African Republic
	Ivory Coast
	Cameroon
	Cape Verde
	Djibouti
	Egypt
	Western Sahara
	Eritrea
	Ethiopia
	Gabon
	Ghana
	Gambia
	Guinea
	Equatorial Guinea
	Guinea-Bissau
	Kenya
	Comoros
	Kuwait
	Liberia
	Lesotho
	Libya
	Morocco
	Madagascar

	Mali
	Mauritania
	Mauritius
	Malawi
	Mozambique
	Namibia
	Niger
	Nigeria
	Reunion
	Rwanda
	Seychelles
	Sudan
	Saint Helena
	Sierra Leone
	Senegal
	Somalia
	Sao Tome and Principe
	Swaziland
	Chad
	Tanzania
	Uganda
	Mayotte
	South Africa
	Zambia
	Zimbabwe
	Algeria
	Tunisia
VII.	
Australia	Australia
	Cocos Islands
	Heard Island and McDonald Islands
VIII.	
Oceania	American Samoa
	Cook Islands
	Christmas Island
	Fiji
	Micronesia
	Guam
	Hawaii
	Kiribati
	Marshall Islands
	Northern Mariana Islands
	New Caledonia
	Norfolk Island
	Nauru
	Niue
	New Zealand
	French Polynesia

	Papua New Guinea
	Pitcairn
	Palau
	Solomon Islands
	Tokelau
	Tonga
	Tuvalu
	Vanuatu
	Wallis and Futuna
	Samoa
IX. Antarctica	Antarctica
X. Arctic	Greenland
XI. World	Classify here if item belongs in two or more of the above categories

Appendix D. Materials for evaluation

D.3 Regions mapped



The indexers’ copy was four times larger for clarity and so it had to be printed in black and white.

Appendix D. Materials for evaluation

D.4 Themes expanded

- Agriculture
  - Farming, soil, fertilizers
  - Conservation of natural resources
  - Nurseries, crops and harvesting
  - Horticulture
  - Gardens
  - Parks and reservations
  - Pests and weeds and pollutants
  - Hazardous waste
  - Street cleaning and sewage
  - Forestry
  - Animal culture: breeding, grazing
    - Cattle and dairy
    - Horses and racing
    - Sheep, goats, fur animals
    - Poultry
    - Pets
    - Insect rearing: bees and honey
  - Veterinary medicine
  - Fishing and shellfish, seafood, and whaling
  - Hunting, shooting, trapping
  - Wildlife

- Archaeology and Anthropology
  - Early human culture

- Arts and Media
  - *Arts* includes
    - Art and architecture
    - Museums
    - Photograph and painting
    - Dance
    - Music
    - Theater/drama
    - Literature
    - Crafts (woodworking, needlepoint, etc.)
    - Furniture, rugs and tapestries
    - Decoration
  - *Media* includes
    - Television

Radio
Motion picture (cinema)
Journalism

Commerce and Finance

Commerce includes

Commercial law
Manufacturing industries

Metal, fur, paper, rubber, cereal, textile, tobacco
Economics

Price, competition, production, wealth, capital, income, interest, profit, entrepreneurship, welfare
Industry management, innovation, public and social relations and team work
Economics of land use, agrarian reform, sharecropping
Trade associations and industry
Labor and trade unions
Commerce
Trade, tariff, tax, protectionism
Shopping, wholesale, shipping, purchasing, retail, selling, department stores, mail order, warehouses, fairs and markets, black market, shipping, delivery and advertising

Finance includes

Liquidity, money, banking, interest, bank accounts, stocks, credit, loans, debt, foreign exchange, trust companies
Investment, venture capital
Lotteries
Insurance (life, fire, health, accident)
Taxes, auditing, inflation
Banking
Money, loans and investments
Insurance
Public finance
Insolvency and bankruptcy

History and Travel

History includes

Military and naval history
Political and diplomatic history
Medieval history: crusades, migrations
Wars and battles with names
Periods of occupation, of dominance, dynasty, empire and administration

    - Personal history and history of people: genealogy and ethnography
    - Archives, seals and documents
    - Calendar
  - Travel
    - Voyages of exploration
    - Atlases, globes and maps
    - Geography and human geography (demography and statistics, etc.)

- Medicine
  - Public and personal disease, health and sanitation
  - Public health and hygiene
    - Immunity and immunization
    - Disease, epidemics, quarantine
    - Toxicology and poisons
    - Hospital and nursing homes
    - First aid, intensive care
    - Red Cross, Red Crescent
    - Legal aspects of medicine
    - Hazardous waste
    - Street cleaning
    - Sewage disposal
  - Internal medicine
    - Neurosciences, psychiatry, immunology, surgery, ophthalmology, otorhinolaryngology, gynecology, pediatrics, dentistry, dermatology, therapeutics, pharmacology, homeopathy, chiropractic
    - Health
    - Diet and vitamins
  - Personal hygiene

- Military
  - War and battle
  - Strategy and tactics and safety
  - Cavalry and troops
  - Vessels and planes for troops
  - Equipment and supplies and barracks for troops
  - Artillery
  - Navigation, sailing and shipwrecks
  - Heraldry
  - Flags, banners, standards and insignia

- Politics and Law
  - Political science, nationalism, sovereignty, patriotism
    - Executive branch, civil service

Legislature, Congress, House of Representatives
Political parties
Local and municipal government
Colonies, emigration and immigration
Diplomacy
United Nations
Socialism, communism, anarchism, utopia, democracy
Law
Jurisprudence, legal theory, trials, treaties, contracts, torts, arbitration, negotiation
Intellectual law, law over drugs and alcohol
Criminal law and national defense
Animal rights
Military law
Federal law
History of law
Judicial decisions and law reports
Law of space
Prevention of crime, police, detectives, traffic control
Criminology
Court and jury

Religion and Education
Religion includes
Philosophy, metaphysics, cosmology
Ethics, virtue and vice
Soul, monotheism, polytheism, doctrines
Hinduism, Jainism, Zoroastrianism, Confucianism, Taoism, Shinto
Judaism, Islam, Buddhism, Mysticism
Christianity
Bible
Places of worship
Liturgy and prayer
Sermons and creeds
Education includes
School: elementary, middle, high, college, graduate, vocational
Literacy
Testing

Science
Mathematics, geometry
Astronomy, solar system, stars
Physics, acoustics, thermodynamics (heat), optics, radiation, electricity and magnetism, meteorology, climatology

- Chemistry
- Geology
- Biology, genetics
- Botany
- Zoology
- Anatomy and human physiology
- Microbiology
- Cartography, remote sensing and Geographic Information Systems
- Physical geography: Hydrology and water
- Oceanography

Society

- Behavior
  - Psychology, personality, temperament
  - Etiquette and manners, fashion and style
  - Customs and dress
  - Recreation and leisure
  - Camping
- Customs
  - Sports and games
  - Death and dying (thanatology)
  - Home
  - Nutrition and Cookery
  - Hospitality industry (hotels, restaurants, clubs, taverns, pubs, saloons)
  - Laundry
- Sociology and behavior of groups
  - Family marriage women
  - Sexual behavior, homosexuality, etc. : life style
  - Erotica
  - Parents, Children, birth control, family planning
  - Adultery, divorce, polygamy
- Communal behavior
  - Societies and fraternities
  - Community and urbanization
  - Classes: caste system, serfdom, slavery
  - Social work
  - Refugees
  - Orphanages
  - Alcoholism, poverty, drug abuse, slums
- Language and communication
  - Present and past languages and linguistics

Technology and Transportation

- *Technology* includes

- Engineering
- Patents and Trademarks for inventions
- Environmental engineering
- Building construction
- Mechanical engineering and machinery
- Energy, Heating, Nuclear engineering
- Power, fuel and gas
- Agricultural machinery
- Domestic machinery (sewing machines)
- Electrical engineering
- Lighting
- Computers
- Telephone industry and wireless communications
- Mining and Metallurgy
- Chemical engineering
- Ceramics, glass, exterior paint, varnish

Transportation includes

- Highway engineering and infrastructure pavement, roads and sidewalks)
- Railways and bridges
- Motor vehicles
- Bridges, tunnels, waterways, shipping, boats and ferry
- Automotive, bus and taxi
- Airlines

References

Abdelmoty, A. I., Smart, P., & Jones, C.B. (2007). Building place ontologies for the semantic web: Issues and approaches. *Geographic Information Retrieval Proceedings of the Fourth ACM workshop on Geographic Information Retrieval, November 9, 2007, Lisbon, Portugal*, 7-12.

Agrawal, R., Imielinski, T. & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record 22* (2), 207-216.

Albrecht, J. (2007). *Key concepts and techniques in GIS*. Los Angeles: Sage.

Alonso, O., Gertz, M., and Baeza-Yates, R. (2007) On the value of temporal information in information retrieval. *ACM SIGIR Forum 41*(2), 35-41.

Amitay, E., Har'El, N., Sivan, R., & Soffer, A. (2004). Web-a-Where: Geotagging web content. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 25-29, Sheffield, United Kingdom*, 273-280.

Bates, M. (1998). Indexing and access for digital libraries and the internet: Human, database, and domain factors. *Journal of the American Society for Information Science 49*(13), 1185-1205.

Blair , D. C. & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-test document-retrieval system. *Communications of the ACM 28* (3), 289-299.

Borges, K. A. V., Laender, A. H. F., Medeiros, C. B., & Davis Jr., (2007). Discovering geographic locations in web pages using urban addresses. *Proceedings of the 4th ACM Workshop on Geographical Information Systems, November 9, 2007, Lisbon, Portugal*, 31-36.

Brodeur, J., Bédard Y. & Moulin, B. (2005). A geosemantic proximity-based prototype for the interoperability of geospatial data. *Computers, Environment and Urban Systems 29*, 669-698.

Buckland, M., Chen, A., Gey, F. C., Larson, R. R., Mostern, R., & Petras, V. (2007). Geographic search: Catalogs, gazetteers, and maps. *College & Research Libraries 68*(5) 376-387.

Burrough, P. A. & McDonnell, R. A. (2000). *Principles of geographical information systems*. New York: Oxford University Press.

Bush, V. (1945). As we may think. *The Atlantic Monthly*. Retrieved November 30, 2007 from http://www.theatlantic.com/doc/194507/bush

Callan, J., Allan, J. Clarke, C. L. A., Dumais, S. Evans, D. A., Sanderson, M., Zhai, C-X. (2007). Meeting of the MINDS: An information retrieval research agenda. *ACM SIGIR Forum, 41*(2), 25-34.

Chang, E., Huang, C-R., Ker, S-J., Yang, C-H (2002). Induction of classification from lexicon expansion: Assigning domain tags to WordNet Entries. *International Conference on Computational Linguistics COLING-02 on SEMANET: Building and Using Semantic Networks 11*, 1-7.

Chen, Y-Y, Suel, T., & Markowetz, A. (2006). Efficient query processing in geographic web search engines. *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, June 27-29, Chicago, Illinois*, 277-288.

Chua, S. & Kulathuramaiyer, N. (2004). Semantic feature selection using WordNet. *Proceedings of the ICCC/WIC/ACM International Conference on Web Intelligence*. 166-172.

Cleverdon, C. W., Mills, J. & Keen, E. M. (1966*). Factors Determining the Performance of Indexing Systems, Vol. 1—Design, Aslib-Cranfield Research Project*, Cranfield, England.

Dhar, D. B. & Chanda, B. (2006). Extraction and recognition of geographical features from paper maps*. International Journal of Document Analysis 8*(4), 232-245.

El Sayed, A., Hacid, H., & Zighed, D. A. (2007). Mining semantic distance between corpus terms. *Proceedings of the ACM first Ph.D. workshop in PIKM, November 9, 2007, Lisbon, Portugal*, 49-54.

El Sayed, A., Hacid, H., & Zighed, D. A. (2007b). A multisource context-dependent approach for semantic distance between concepts. In R. Wagner, N. Revell & G. Pernul (Eds.), *Lecture Notes in Computer Science 4653. Proceedings of the 18th International Conference on Database and Expert System Applications (DEXA), September 3-7, 2007, Regensburg, Germany*, 54-63.

Ercegovac, Z. (1998). Minimal level cataloging: What does it mean for maps in the contexts of card catalogs, online catalogs, and digital libraries? *Journal of the American Society for Information Science 49*(8), 706-719.

Furnas, G. W., T. K. Landauer, L. M. Gomez, & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM 30*(11), 964–71.

Gabrilovich, E. & Markovitch, S. (2007). Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization. *Journal of Machine Learning Research 8*, 2297-2345.

Goodchild, M. F. (2008). Epilog: Putting research into practice. From unpublished manuscript on spatial data quality.

Goodchild, M. F., Yuan, M. & Cova, T. J. (2007). Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science 21*(3), 239-260.

Goodchild, M. F. & Zhou, J. (2003). Finding geographic information: Collection-level metadata*.* *GeoInformatica 7*(2), 95-121.

Greenberg, J. (2004). User comprehension and searching with information retrieval thesauri. *Cataloging & Classification Quarterly 37*(3), 103-120.

Henrich, A. & Lüdecke, V. (2007). Characteristics of geographic information needs. *Proceedings of the 4th ACM Workshop on Geographical Information Systems, November 9, 2007, Lisbon, Portugal*, 31-36.

Heuer, J. T. & Dupke, S. (2007). Towards a spatial search engine using geotags. In F. Probst, C. Kessler (Eds.), *Proceedings of the 5th Geographic Information Days 10.-12 September 2007, Münster, Germany,* 199-204

Hill, L. L. (2006). *Georeferencing: The geographic associations of information*. Cambridge, MA: MIT Press.

Hochstein, C. (2006). TOXMAP: A GIS-based gateway to environmental health resources. *Medical reference services quarterly 25*(3), 13 ff.

Jain, A. K., Murty, M. N & Flynn, P. J. (1999). Data clustering: A review. *ACM computing surveys 31*(3) 264-323.

Kammersell, W. & Dean, M. (2007) Conceptual search: Incorporating geospatial data into semantic queries. In A. Scharl & K. Tochtermann (Eds.) *The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society*. (pp.47-54). Springer, London.

Kemp, Z., Tan, L, & Whalley, J. (2007). Interoperability for geospatial analysis: A semantics and ontology-based approach. *Proceedings of the eighteenth conference on Australasian database vol. 63. ACM International Conference Proceeding Series vol. 242. Ballarat, Victorian, Australia*, 83-92.

Khan, L., McLeod, D. & Hovy, E. (2004). Retrieval effectiveness of an ontology-based model for information selection. *The VLDB Journal 13*, 71-85.

Kim, J.-H. & Lee, K.-H. (2002). Designing a knowledge base for automatic book classification. *The Electronic Library 20*(6), 488-495.

Kim, S-B., Seo, H-C., & Rim, H-C. (2004). Information retrieval using word senses: Root sense tagging approach. *27th Annual International ACM Special Interest Group on Information Retrieval (SIGIR) '04, July 25-29, Sheffield, Yorkshire, United Kingdom,* 258-265.

Krug, S. (2006). *Don't make me think! A common sense approach to web usability* (2nd ed.) Berkeley, CA: New Riders.

Kuhn, W. (2005). Geospatial semantics: Why, of what, and how? In S. Spaccapietra and E. Zimányi (Eds.), *Journal on Data Semantics III, Lecture Notes in Computer Science 3534*, 1-24.

Larsgaard, M. L. (2005). Metaloging of digital geospatial data. *The cartographic journal 42*(3), 231-237.

Larson, R. R. (1992). Experiments in automatic Library of Congress Classification. *Journal of the American Society for Information Science 43*(2), 130-148.

Leidner, J. L. (2007). Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names. Unpublished doctoral dissertation, University of Edinburgh, United Kingdom. Retrieved January 8, 2008 from http://hdl.handle.net/1842/1849

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. *Proceedings of the 5th Annual Conference on Systems Documentation. ACM Special Interest Group for Design of Communication*, 24-26.

Leveling, J. (2007). Exploring term selection for geographic blind feedback. *Proceedings of the 4th ACM Workshop on Geographical Information Systems, November 9, 2007, Lisbon, Portugal*, 43-48.

Lieberman , M. D., Samet, H., Sankaranarayanan, J., Sperling, J. (2007). *STEWARD: Architecture of a spatio-textual search engine*. Proceedings of the 15th International Symposium on Advances in Geographic Information Systems ACM GIS 2007, November 7-9, 2007, Seattle, WA, 1-8.

Longley, P. A., Goodchild, M. F., Maguire, D. J. & Rhind, D. W. (2001). *Geographic Information Systems and Science*. NY, Chichester: John Wiley & Sons.

Manning, C. D., Raghavan, P., Schütz, H. (2007). *Introduction to information retrieval*. Retrieved November 30, 2007 from http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html

Marley, C. (2001). The changing profile of the map user. In R. B. Parry and C. R. Perkins (Eds.), *The map library in the new millennium* (pp.12-27). Chicago: American Library Association.

Martins, B., Silva, M. J., Chaves, M. S. (2005). Challenges and resources for evaluating geographic IR," *Proceedings of the ACM Workshop on Geographic Information Retrieval November 4, 2005, Bremen, Germany*, 65-69.

Moellering, H. (Ed). (2005). *World spatial metadata standards: Scientific and technical descriptions, and full descriptions with crosstable*. Associate editors I. H. Aalders and A. Crane. Amsterdam: Elsevier.

Nielson, M. L. (2004). Thesaurus construction: Key issues and selected readings. *Cataloging & Classification Quarterly 37*(3/4), 57-74.

Nogueras-Iso, J., Zarazaga-Soria, F. J., & Muro-Medrano, P. R. (2005). *Geographic information metadata for spatial data infrastructures: Resources, interoperability and information retrieval*. Berlin: Springer.

Oberhauser, O. (2005). *Automatisches Klassifizieren: Entwicklungsstand – Methodik – Anwendungsbereich*. Europäische Hochschulschriften. Series XLI Informatik, vol. 43. Frankfurt am Main: Peter Land.

O'Connor, D. O. (1978). *The interactive influences of person and situation characteristics on expectations and their cumulative effects on relevance judgments and file preference choices by undergraduates search alphabetical and classed subject catalogs*. Unpublished PhD Dissertation, Syracuse University.

Pasley, R. C., Clough, P., & Sanderson, M. (2006). Geo-tagging for imprecise regions of different sizes. *Proceedings of the 4th ACM Workshop on Geographic Information Retrieval, November 9, 2006, Lisbon, Portugal*, 77-82.

Peng, Z-R. & Tsou, M-H. (2003). *Internet GIS: Distributed geographic information services for the internet and wireless networks*. Hoboken, NJ: John Wiley & Sons.

Perry, M., Hakimpour, F. & Sheth, A. (2006). Analyzing theme, space, and time: An ontology-based approach. *Proceedings of the 14th ACM International Symposium on Geographic Information Systems, ACM-GIS 2006, November 10-11, 2006, Arlington, Virginia, U.S.A.*, 147-154.

Peters, T. A. & Kurth, M. (1991). Controlled and uncontrolled vocabulary subject searching in an academic library online catalog. *Information Technology and Libraries 10*, 201-211.

Peterson, M. P. (2007). Hypermedia maps and the internet. In E. Stefanakis, M. P. Peterson & C. Armenakis, & V. Delis (Eds.) *Geographic Hypermedia: Concepts and Systems*.

(pp.121-136) in Lecture Notes in Geoinformation and cartography series. Berlin: Springer.

Petras, V., Larson, R. R., Buckland, M. (2006). Time period directories: A metadata infrastructure for placing events in temporal and geographic context. *Proceedings of the 6th ACM/IEEE CS Joint Conference on Digital Libraries, June 11-15, 2006, Chapel Hill, NC, U.S.A.*, 151-160.

Phelps, T. A. & Wilensky, R. (2000). Multivalent documents. *Communications of the ACM 43*(6), 83-90.

Prabowo, R., Jackson, M., Burden, P. & Knoell, H.-D. (2002). Ontology-based automatic classification for web pages: design, implementation and evaluation. *Proceedings of the 3rd International Conference on Web Information Systems Engineering*, 182-191.

Roddick, J. F., Hornsby, K., & de Vries, D. (2003). A unifying semantic distance model for determining the similarity of attribute values. *Proceedings of the 26th Australasian computer science conference Vol. 16, ACM International Conference Proceeding Series; Vol. 35,* Adelaide, Australia, 111 – 118.

Salton, G. (1968). *Automatic information organization and retrieval*. New York: McGraw–Hill Book Company.

Salton, G. & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw–Hill Book Company.

Samet, H., & Soffer, A. (1996). MARCO: Map Retrieval by Content. *IEEE transactions on pattern analysis and machine intelligence 18*(8), 783-798.

Sanderson, M. (2000). Retrieving with good sense. *Information Retrieval 2*, 47-67.

Sanderson, M. & Han, Y. (2007). Search words and geography. *Proceedings of the 4th ACM Workshop on Geographical Information Systems, November 9, 2007, Lisbon, Portugal*, 13-14.

Sanderson, M. & Kohler, J. (2004). Analyzing geographic queries. Workshop on geographic information retrieval, SIGIR, 2004, July 25-29, 2004, Sheffield, United Kingdom, n.p., Retrieved December 17, 2007 from http://www.geo.unizh.ch/~rsp/gir/abstracts/sanderson.pdf.

Sandusky, R. J. & Tenopir, C. (2008). Finding and using journal-article components: Impacts of disaggregation on teaching and research practice. *Journal of the American Society for Information Science and Technology 59*(6), 970-982.

Saracevic, T. (2006). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II. unpublished manuscript, 1-89.

Sensuse, D. I. (2004). *A comparison of manual indexing and automatic indexing in the humanities* [Abstract]. PhD. dissertation for the University of Toronto, Canada.

Shaw Jr., W. M. (1993). Controlled and uncontrolled subject descriptions in the CF database: A comparison of optimal cluster-based retrieval results. *Information Processing & Management*. 29 (6), 751-763.

Shawa, T. W. (2006). Building a system to disseminate digital map and geospatial data online. *Library Trends 55*(2), 254-263.

Shearer, J. R. (2004). A practical exercise in building a thesaurus. *Cataloging & Classification Quarterly 37*(3/4), 35-56.

Shiri, A. & Crawford, R. (2006). Query expansion behavior within a thesaurus-enhanced search environment: A user-centered evaluation. *Journal of the American Society for Information Science and Technology, 57*(4), 462-478.

Shneiderman, B. & Plaisant, C. (2005). *Designing the user interface: Strategies for effective human-computer interaction*. 4th edition. Boston: Pearson.

Simeoni, F., Yakici, M., Neely, S. & Crestani, F. (2008). Metadata harvesting for content-based distributed information retrieval. *Journal of the American Society for Information Science and Technology, 59* (1), 12-24.

Simon, H. A. (1997). Administrative behavior: A study of decision-making processes in administrative organizations (4th ed.) New York: Free Press. (Original work published 1976).

Smart, P. D., Abdelmoty, A. L, El-Geresy, B. A. and Jones, C. B. (2007). A framework for combining rules and geo-ontologies. *Lecture Notes in Computer Science 4524*, 133-147.

Spärck Jones, K. (2005). Some thoughts on classification for retrieval. *Journal of Documentation 61*(5), 571-581.

Sturtz, D. (2004). Communal categorization: The folksonomy. Retrieved March 4, 2007 from http://www.davidsturtz.com/drexel/622/sturtz-folksonomy.pdf

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Boston: Pearson.

Tezuka, T., Kurashima, T., & Tanaka, K. (2006). Toward tighter integration of web search with a geographic information system. *Proceedings of the 15th International Conference on World Wide Web, May 23-26, 2006, Edinburgh, Scotland*, 277-286.

Turpin, A. H. & Hersh, W. (2001). Why batch and user evaluations do not give the same results. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-12, 2001, New Orleans, Louisiana, U.S.A.*, 225–231.

Vaid, S., Jones, C. B., Joho, H. & Sanderson, M. (2005). Spatio-textual indexing for geographical search on the web. *Proceedings of the 9th International Symposium on Spatial and Temporal Databases*, August 22-24, 2005, Angra dos Reis, Brazil, 1-18.

Van Rijsbergen, C. J. (1979). *Information Retrieval*. 2nd ed. (pp.1–10) London: Butterworths. Retrieved September 28, 2005 from http://www.dcs.gla.ac.uk/Keith/Chapter.1/Ch.1.html

Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E., & Milios, E. (2005). Semantic similarity methods in WordNet and their application to information retrieval on the web. *7th ACM International Workshop on Web Information and Data Management, November 5, 2005, Bremen, Germany*, 10-16.

Velasco, M. and Boba, R. (2000). *Manual of crime analysis map production*. Report to the Office of Community Orienting Policing Services, Cooperative Agreement #97-CK-WXK-004. U.S. Department of Justice.

Wang, Y., Hodges, J. & Tang, B. (2003). Classification of web documents using a naïve Bayes method. *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, 560-564.

Wang, J. & Lee, M-C. (2007). Reconstructing DDC for Interactive Classification*. ACM Sixteenth Conference on Information and Knowledge Management, November 6-8, Lisbon, Portugal*, 137-146.

Wolfram, D. & Zhang, J. (2008). The influence of indexing practices and weighting algorithms on document space. *Journal of the American Society for Information Science and Technology 59* (1), 3-11.

Wright, P. (1998). Knowledge discovery in databases: Tools and techniques. Retrieved March 4, 2008 from http://www.acm.org/crossroads/xrds5-2/kdd.html.

Yager, R. R. & Rybalov, A. (1998). On the fusion of documents from multiple collection information retrieval systems. *Journal of the American Society for Information Science 49*(13), 1177-1184.

Yao, H., Etzkorn, L. H., & Virani, S. (2008). Automated classification and retrieval of reusable software components. *Journal of the American Society for Information Science and Technology, 59*(4), 613-627.

Zadeh, L. A. (1999). Fuzzy logic = computing with words. In L. A. Zadeh & J. Kacprzyk (Eds.), *Computing with Words in Information/Intelligent Systems 1: Foundations (Studies in Fuzziness and Soft Computing)* (pp. 3-23). Heidelberg: Physica.

Zhou, G. & Su, J. (2002). Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), Philadelphia, PA*, 473-480.

Curriculum vitae

Judith Gelernter

Education

Rutgers University	Info Science	Ph.D., 2008
Simmons College	Info & Lib Science	M.I.L.S., 1994
Harvard University	Fine Arts	A.M., 1992
Yale University	Medieval Studies	B.A., 1989, *magna cum laude,* distinction in the major

Certification

C.I.P.M. Certified Institutional Protection Manager, Security, from the International Foundation for Cultural Property Protection, 2004

Network+ Networking, Computing Technology Industry Association, 2003

A+ Hardware and Operating Systems, Computing Tech. Industry Association, 2003

Teaching Experience

Rutgers University
Research Assistant, fall 2007 – spring 2008
Teaching Assistant, fall 2005 – spring 2007

Queens College, City University of New York
Adjunct Assistant Professor, 2005

Professional Experience

Union Club of the City of New York, NY
Library Director and Art Curator, 1997–2004

Dance Notation Bureau, New York, NY
Librarian and Archivist, 1994–1997

Houghton Rare Book and Manuscript Library, Harvard University, Cambridge, MA
Cataloging Assistant, 1993–1994

Fogg Art Museum, Harvard University, Cambridge, MA
Curatorial Assistant, Department of Prints and Drawings, 1992

Publications

Gelernter, J. (2008). Neogeography, submitted in January 2008 to the Editors of the *Encyclopedia of Social Interaction Technologies*.

Gelernter, J. (2007). Information Visualization for Digital Libraries. *Knowledge Organization 34*(3), 128-143.

Gelernter, J. (2005). Why Theft Prevention Should be High Priority and Loss Prevention Strategies for the 21st Century Library, *Information Outlook* 9(12), 12-22 (by editor invitation)

Gelernter, J. (2004). Infoviz for Info Pros: Information Visualization Software Tools. *Searcher 12*(9), 52-61.

Gelernter, J. (2004). The Digital Library Edge. *Information Outlook 8(*8), 16-18.

Gelernter, J. (2003). Specialty Search Engines. *Searcher 11*(1), 26-31.

Gelernter, J. (2001). The Internet: Yesterday, Today and Tomorrow. *Information Outlook 5*(6), 67-68.

Gelernter, J. (1995-1997). Sharing Images over a Network. *Visual Resources Association Bulletin 24*(1) (Spring 1997), 37-38. Storing Images. *Visual Resources Association Bulletin 23*(1) (Fall 1996), 42-43. Images on Screen *Visual Resources Association Bulletin 22*(1) (Spring 1995), 26-27.

Conference proceedings and presentations

Gelernter, J. & Lesk, M. (2008). Creating a Searchable Map Library via Data Mining. *Joint Conference on Digital Libraries, June 16-20, 2008, Pittsburgh, Pennsylvania.*

Gelernter, J. (2007). A Quantitative Analysis of Collaborative Tags: Evaluation for Information Retrieval—A Preliminary Study. *The 3rd International Conference on Collaborative Computing: Networking, Applications and Worksharing November 12-15, 2007, White Plains, New York.*

Gelernter, J., & Old, C. (1995). The Metropolitan Museum of Art Performing Arts Index, In A. W. Smith (Ed.), *Dance and Technology III*, (pp. 9-16) Also presented at the conference.

Gelernter, J. (1994). Mannerist Aesthetics and the Court Dance of Fabritio Caroso *Proceedings of the Society of Dance History Scholars* 10-13, 129-139. Conference presentation, winter 1993.

Reviews

Gelernter, J. (2006). Modern Dancer: The World According to Tharp. *The Weekly Standard 11*(42), 34-36.

Gelernter, J. (2006). Pointes of View. *The Weekly Standard 11*(21), 46-47.

Gelernter, J. (2004). Balanchine's Stage: Terry Teachout on how we tell the dancer from the dance. *The Weekly Standard*, *10*(12) (December 6, 2004), 25-26. By editor invitation.

Gelernter, J. (2004). Same Old Song and Dance: How the Nutcracker Captured America. *The Weekly Standard 9*(16), 45-46.

Gelernter, J. (2003). The June TECHXNY Conference and Implications for Libraries. *Information Technology and Libraries 22*(1), 32-35.

Gelernter, J. (2003). Broadway Ballet: In Movin' Out, Twyla Tharp Creates a Dance to the Music of Time. *The Weekly Standard 8*(16), 37-38.

Gelernter, J. (1999). Taylor's Tangos. *Ballet Review 27*(3), 92-96.

Gelernter, J. (1997). Partsche Bergsohn and Bergsohn: Early Dance. *Dance Research Journal 29*(1), 91-93.

Gelernter, J. (1994). Database Management Software Review: askSam. *Special Libraries 85*(4), 297-298.

Book chapters/entries

Gelernter, J. & Rader, P. (2001). Dance Notation. in M. Edsall (Ed.), *A Core Collection in Dance*, American Library Association.

Gelernter, J. (1998). The Shakers by Doris Humphrey. In T Benbow-Pfalzgraf (Ed.), *International Dictionary of Modern Dance* (pp. 709-710). St. James Press.

Gelernter, J. (1998). White Oak Dance Project. In T. Benbow-Pfalzgraf (Ed.), *International Dictionary of Modern Dance* (pp. 819-821). St. James Press.

Professional activities

Professional service

Reviewer, *Encyclopedia of Social Interaction Technologies*, 2008.
Editorial Board, *Information Technology and Libraries,* 2002–2004.
Grant Reviewer, Institute of Museum and Library Services, spring 2003.
Editor, *Channels*, the World Dance Alliance, America's Center newletter, 1995–1996.

Grants and awards

Mobile Knowledge, July 7, 2007 – June 30, 2008, Academic Excellence Fund, Rutgers University, Project Directors: M. Lesk, K. Ozbay, and C. Rose
Institute for Scientific Information Scholarship, spring 2006
Lillian Moore Award for the book: *A Core Collection in Dance* (I co-authored the chapter on Notation), 2003
LSTA Technology Grant, spring 2001
Gadd/Merrill Endowment Grant, winter 1996
Simmons College Scholarship, fall 1993
Mellon Grant, spring 1992
Harvard Student Travel Grant, spring 1990
Harvard University Scholarship, fall 1989–spring 1992

Consulting

Cosmopolitan Club Library, NY, 2007– present
Kaye Scholer, LLP, NY, 2005–2006
General Society of Mechanics and Tradesmen, NY, 2005–2006
Yale Club of the City of New York, NY, 2005
National Minority Supplier Development Council, Inc., NY, 2005
New York Genealogical and Biographical Society, NY, 2004–2005
New York Yacht Club, NY, 2004
Polaris Arts, Ltd., NY, 2002–2004
New Haven Day School Library, CT, 2001–2002