GIPSY
(Georeferenced Information Processing SYstem)

Allison Gyle Woodruff
Computer Science Department
University of California, Davis
woodruff@cs.ucdavis.edu

Abstract

In current information retrieval systems, geographically-based searches such as "Find me all documents whose content pertains to location X" are not supported directly by indexing, query, and display functions. Instead, these searches at best exist only as side-effects of keyword indexing. Because of problems relating to ambiguity, synonymy, and other linguistic phenomena, the use of keyword indexing to indicate geographic location can lead to inaccurate and/or misleading references.

The initial portion of this paper presents the motivation for georeferencing text and previous work in this area. The next portion discusses a needs assessment presenting two studies documenting the frequency and composition of geographic queries in information retrieval systems. The first study is an examination of commands executed in a university online bibliographic catalog. The second study shows searches done by natural resources managers.

The final portion of this paper details a new model for supporting geographically-based access to text. In this model, words and phrases containing geographic place names and geographic characteristics are extracted from a document. These terms are then used as input to database functions which use spatial reasoning to statistically approximate the location being referred to in the document. Advantages and disadvantages of this model are discussed.

1. Introduction

A georeferenced document is a document which may be retrieved according to the geographic location of its content. There are several motivations for providing this type of indexing, and recent technological advances have made work in this area more viable.

1.1. Motivation for Georeferencing Text

A diverse and large group of library users desires geographic indexing of documents. Some of these include:

- **natural resources managers** who would like to retrieve information pertinent to specific areas. For example, the preparation of an EIR may require the examination of archaeological, biological, and geological texts referring to a location of interest.
- **Earth scientists** who would like to locate publications which discuss certain locations. For example, an Earth scientist studying climatology of the Western United States will want to retrieve meteorological research publications for that area.

- **historians** who would like to retrieve documents about specific areas.
- **journalists** who would like to locate documents pertinent to current events occuring at particular locations.
- **tourists** who would like to locate brochures, guide books, and historical information germane to areas being considered for trips.

Historically, indexing has been done primarily by subject and, to a lesser extent, by document type. However, in the types of searches discussed above, geographical location is the main focus of the search. In these searches, the documents being retrieved cross many subject areas, and are of many different document types (e.g. journals, newspapers, magazines, etc.). This often makes a comprehensive search difficult, as Byron [Byro87] observes:

> "Bradford's Law of Scattering does not apply only to the literature of science and technology: articles in the arts, humanities and social sciences also appear in unlikely periodicals, so that the task of achieving comprehensive coverage in a subject library is rendered impossible. No special librarians know this with more chagrin than those in local studies departments of public libraries; intent as they are on tracing, obtaining and conserving for posterity everything published about a given area or place, they find that potentially useful material appears in a multitude of publications... Instead of being able to search all the latent sources systematically to make additions to their collection, librarians are dependent upon serendipity and the goodwill of colleagues and library users..." (pg. 111)

Geographic location is the shared attribute which can draw together these disparate subjects and document types to allow a cohesive search. Authors such as Byron [Byro87], and the Research Libraries Group [RLG89] have therefore recognized the need for research, development, and implementation of geographic indexing.

Furthermore, in many cases, the users listed above will also wish to access and integrate text documents with other kinds of data such as photographs or maps. As Griffiths [Grif89] observes:

> "By linking data of different types one is able, for example, to relate disease distribution data to geological rock types or surname distribution." (pg. 261)

Geographic indexes are therefore also useful in joining together text documents with other geographically indexed data types [Gard93]. However, current georeferencing of text documents is inadequate in several ways. The most prominent needs will be presented below.

1.1.1. Need for Coordinate Indexes

Even in cases where a document is meticulously manually indexed, geographic indexes which consist of text strings have several well-documented problems, including:

- **lack of uniqueness** [Grif89], [Holm90]
  Confusion can result from a reference to **Cambridge**, which might refer to Cambridge, England or Cambridge, Massachusetts (to mention only two options).

- **spatial boundary changes** [Grif89]
  Political boundaries change over time, leading to confusion about the precise definition of the area being referred to.
- **name changes** [FaLe82], [Grif89]
  Geographic names change over time. This makes it difficult for a user to retrieve all information which has been written about an area during a lengthy time period.
- **spatial and naming variation** [FaLe82]
  Terms vary not only over time, but in contemporaneous usage as well.
- **spelling variation** [FaLe82], [Grif89]
  Geographic names vary in spelling, both over time and by language.
- **neologisms** [FaLe82]
  Study areas will often be given place names designated only in the context of a specific document or project. These occur particularly frequently for studies done in oceanic locations. These terms will be unknown to most users.

An information retrieval system must share a common language with the user [Batt89]. Lack of communication caused by the complications listed above has made retrieval by geographic indexes difficult. Inconsistency and ambiguity in Library of Congress geographic subject headings, for example, has been discussed at length in [Brin62], [Misc82], and [Well78]. Because coordinates, unlike text, are non-ambiguous [Holm90], they can be used to uniquely communicate locations to the user. Therefore, a georeferencing facility should make use of coordinate indexing.

Furthermore, many users require proximity queries which can not be supported by textual terms. Coordinates, particularly those representing polygonal entities rather than points, can be used to support these nearness queries.

1.1.2. Need for Spatial Interface

Texts indexed by coordinates may be accessed in two ways. First, the user may access them numerically, e.g. by specifying coordinate ranges. Second, the user may access them through the use of a graphical interface, e.g. icons placed on a map according to their index coordinates.

A graphical interface based on coordinates has several advantages over a model which uses text terms and over a model which uses numerical access to coordinates. First, it has been suggested that humans use different cognitive structures for graphical information than for verbal information, and that spatial queries can not be fully simulated by verbal queries [Furn91]. Because geographical queries are inherently spatial, a graphical model is more intuitive. This is supported by Morris' observation that users given the choice between menu and graphical interfaces to a geographic database preferred the graphical mode [Morr88].

Second, a graphical interface such as a map allows dense presentation of information [McTT88]. This addresses a problem which arises in the standard model of information retrieval. In this model, the user issues a query, and is subsequently shown only the subset of documents which the system judges to be relevant to the query. This modularization limits the user's view of the contents of the database, making navigation difficult.

If, instead, the user were able to interact with the full contents of the database, queries could be guided by the contents of the database. Korfhage [Korf91] has in fact argued that all information in a database should be on display to allow the user to make an informed query. This would allow the user greater flexibility and facilitate browsing [Korf86].

Third, in standard information retrieval systems, the display is not ordered by relevance. In this model, users receive insufficient information about what documents in the returned set are most pertinent to their searches. A map-based interface would provide implicit information about relevance through visual presentation of proximity.

Development of the behavior and characteristics of such an interface requires human factors research. As McCann, Taylor, and Tuori [McTT88] observe,

> "Humans have evolved well defined techniques for conversing through language with associated gestures; they have never before been confronted with a dialogue partner that draws pictures as readily as it uses language." (pg. 123)

1.1.3. Need for Automatic Indexing

Historically, indexing of geographic coordinates has been done manually. However, there are disadvantages to manual indexing. First, manual indexing is expensive and time-consuming. This difficulty is exacerbated by the desire to index subcomponents of documents such as paragraphs, motivated by the hypertext paradigm. Second, it has long been recognized that manual indexing is inevitably inconsistent. Third, many other data types have geographic indexes associated with them, e.g. satellite images or ground or aerial photographs. In the future, these data types many automatically be assigned geographic indexes through the use of GPS technologies [Holm90]. Therefore, automatic indexing of text documents will be necessary to integrate text with rapidly increasing amounts of data of other types.

1.2. Opportunity Statement

Aveney [Aven85] writes:

> "Our computerized catalog systems are still largely card catalog facsimile reproduction systems. While the specialized information retrieval services and many special libraries have been developing successful new methods of indexing and cataloging, the mainstream of American librarianship has steadfastly dodged the issues.
>
> Our current forms of subject access to library collections have proven inadequate. It is inappropriate for our expensive automation efforts to continue to perpetuate the already bankrupt subject approach now used in our new computer systems. We have great new tools in our grasp if we have the courage to use them.
>
> We have left behind our second generation computer systems and developed far more powerful software techniques using the capabilities of third generation

> devices. We must now face up to the intellectual challenge of providing new forms of access to library collections and cease being facsimile reproducers." (pg. 85-86)

Clearly, wide spread access to computing resources changes the nature of several problems which have traditionally confronted libraries, information systems and users. The physical size of the catalog limited the number of index terms for a document [Brin62]; the computational complexity of automatic indexing techniques were limited by cpu speed; authority control created consistency and update problems; and, since documents were not on-line, they could not easily be indexed through electronic means. These factors, and many others, are changing rapidly. For example, decreases in the cost of tertiary storage mitigate limitations on catalog size. Furthermore, new efforts being made in Project Sequoia 2000 both motivate and enable georeferencing of text documents.

The objective of Project Sequoia 2000 is to develop a new computer science work environment for Earth System scientists. As part of the project, many new technologies will be developed in the areas of tertiary storage, tertiary file systems, networking, scientific visualization, database management, and computer-facilitated collaboration [ChLS92]. The latter two issues pertain directly to information retrieval.

To address the needs of Earth System scientists, a new browser paradigm is being proposed [ChLS92]. In this browser, information will be displayed graphically according to continuous characteristics which are attributes of the data. For example, documents will be displayed on a map according to their latitude and longitude. Documents may also be displayed according to the time at which they were generated and the time to which they refer, as well as by more abstract functions such as the reading level of the document, the author's attitudes as expressed in the document, etc. The initial browser will allow for display of two of these functions or dimensions simultaneously. Later versions will support simultaneous display of three dimensions. However, both versions will allow the user to index information in n-space by repeating a two-step process: Step 1: Display the documents according to two (or three) dimensions, to be selected by the user. Step 2: Select a group or region of the displayed documents. Redisplay this newly chosen group according to newly chosen dimensions (Step 1 repeated). At any point, the user can select a specific document and display its contents to the screen.

There are several utilities which will make the query/display interface simpler to use. For example, a history will be kept of the user's search, allowing him or her to backtrack or save a series of steps that have been taken in the browser to reissue them at a later time.

Display can also be made more effective by the highlighting of icons which are potentially of interest to the user. This highlighting can be done by color changes or by animation. It has been shown that animated icons are effective in communicating and disambiguating meaning [BaSm91]. They can also help to alleviate problems resulting from clutter.

The Sequoia browser will include not only text information, but other data types as well, e.g. satellite imagery, simulation model output, or point data. The browser design

was chosen to provide a cohesive interface to all these types. Many of these types will be explicitly stored by the users with spatial location. However, text documents will not have user-supplied geographic indexes, so these must be generated.

Because of the number of documents that will be stored in the Project Sequoia 2000 textbase, manual indexing of these documents is not feasible. Therefore, a system must be designed which will automatically assign geographic location to a text document based on its content.

Project Sequoia 2000 affords its text processing and retrieval components three significant advantages. The first is the large amount of data supplied by the Earth System scientists. The Project Sequoia 2000 database will contain tens of terabytes[1] of spatial and geographic data, including information on vegetative mapping, land use, demography, seismology, cadastral boundaries, water quality, etc. Although it has been argued that a huge knowledge base is necessary to perform large scale natural language processing [LGPP90], previous natural language processing systems have not utilized such a volume of information.

The second advantage provided by Project Sequoia 2000 will be the functionality uniquely allowed by extensions to the POSTGRES database management system (DBMS) [StKe91]. POSTGRES will not only allow integrated access to a number of data types, but will also contain geographic information system (GIS) functionality which will allow spatially-related queries.

The third advantage provided by Project Sequoia 2000 consists of research on segmentation of text documents into text-tiles. These logical units will be useful for identifying subcomponents of text documents which may refer to specific areas.

1.3. Objective

The goal of this research is to explore the utility and viability of indexing full-text documents using coordinates for spatial display. A prototype system will assign each text document an index entry consisting of one or more polygons signifying geographic location. This system will be called **GIPSY**, an acronym for **G**eoreferenced **I**nformation **P**rocessing **SY**stem.

2. Previous Work

Indexing of documents can be based either on text terms or on coordinates. Indexing can be done manually or automatically. In this section, the previous indexing techniques and their application to georeferencing of text documents will be discussed.

2.1. Text Indexing

Text indexes have historically consisted of manually assigned subject indexes, although more recently keyword indexing techniques have matured. This section details some of the previous work in these areas and their deficiencies in georeferencing text

[1] A terabyte is 10^{12} bytes.

documents.

2.2. Library of Congress Subject Headings

Although the Library of Congress did not originally intend to develop a standardized subject protocol [Misc82], the procedures used at the Library of Congress have impact on an international scale [Well78]. Therefore, perhaps the most significant example of manually-generated geographic indexes comes from the Library of Congress.

There are three types of geographic indexing in Library of Congress Subject Headings. The first consists of topical subject headings followed by a geographic subdivision (e.g. ART - PARIS). The second consists of a place name followed by a topical subdivision (e.g. U.S. - HISTORY). The third consists of a phrase which begins with a geographical adjective (e.g. NORWEGIAN LITERATURE) [Brin62]. These have several well-documented problems:

Inconsistent Subheadings

A text document is most often assigned only one of the three types of indexes. Historically, this limit has been enforced to limit the size of the entire catalog. However, the procedure for choosing which type of index is appropriate for various documents has not been well-specified. The consequent inconsistency makes it exceptionally difficult for a user to locate reliably items of interest [Brin62].

Scatter

Subdivision can also fragment references. Mischo [Misc82] writes that:

> "Single entry specific headings with more than one subdivision cause serious problems of scatter. A patron seeking information on "automation of library processes" will need to search under the headings LIBRARIES - AUTOMATION; LIBRARIES - UNITED STATES - AUTOMATION; LIBRARIES - OHIO - AUTOMATION; LIBRARIES - GREAT BRITAIN - AUTOMATION; LIBRARIES - AUSTRALIA - AUTOMATION, etc." (pg. 110)

Updating

Several authors have documented problems with terminology in the Library of Congress Subject Headings. Many of the terms used are not kept up-to-date, as observed by Mischo [Misc82], who writes that:

> "In addition, there are any number of archaic, awkward, and even embarrassing subject headings in LCSH. Often mentioned are the headings **MOTOR-TRUCKS** instead of "trucks", **BIOLOGICAL CHEMISTRY** rather than "biochemistry", **REAL PROPERTY** in place of "real estate", and **EUROPEAN WAR - 1914-1918** as opposed to "World War I"... the uncomfortable situation of directing a student to the heading **MISCEGNATION** for titles on interracial marriage.
>
> Delays in accepting more current nomenclature to replace older terms are

> frequent. For example, it wasn't until 1972 that **ELECTRONIC CALCULATING MACHINES** became **COMPUTERS**." (pg. 108)

Furthermore, geographic terms have content which is dependent upon the time period to which they refer. Wellisch [Well78] discusses some of the problems which arise from attempts to use a single geographic term to refer to an area:

> "The instructions tell us to assign the heading **Banks and Banking - Zaire - Kinshasa** to a work on banking in Leopoldville, Belgian Congo, published in 1950. A scholar interested in the financial institutions of the Belgian Congo may or may not necessarily be interested in what happens now in one of its successor states, Zaire, but the fact remains that the status, the scope of activity, the influence on the economic life of the country, even the names of the banks in Belgian Congo and in Zaire are quite different. Most important of all: Kinshasa in Zaire is *not* what the author of the book dealt with, because it did not exist when he wrote the book. By the same token, a work on the school system in Breslau in 1900 is by no stretch of the imagination a work about education in Poland, for the simple reason that Poland did not exist in 1900, and the city of Breslau, which had not been under Polish rule since the fourteenth century, was in 1900 a part of Germany, becoming the Polish Wroclaw only after World War II, in 1945. A historian of German education at the turn of the century would certainly miss the book on Breslau, because it would not occur to him to look under the heading **Education - Poland**, knowing that there was no country named Poland at the time in question, nor would he be likely to look for additional material under **Education - France - Alsace**, because he would know that the province in question belonged at that time to Germany and not to France. Conversely, the investigator of the subject of education in Poland is badly served: if, during his search under the heading **Education - Poland** he encounters the further subheading **Wroclaw** and finds there a book on education in Breslau in 1900, he will be puzzled or annoyed, but he will certainly not wish to use the book, since it is clearly on education in Germany and has nothing whatsoever to do with Polish Education... One might as well say that a book on mammoths should have the subject heading **Elephants** because these animals are 'now' the representatives of that extinct species...
>
> While the valiant struggle of the Polish people for more than 120 years to reestablish their sovereign country has the admiration of most people, including catalogers, it ought not to be the business of subject indexers to retroactively invent a Polish state for a period during which it did not exist, or to make people think that Alsace belonged to France during a time when it was actually part of the German state.
>
> The new instructions regarding local subdivisions of topics, when those subdivisions were under different sovereignties during their history, have the effect of completely falsifying the representation of the subject matter of a book, thus negating the most fundamental rule of subject cataloging..." (pp. 162-164)

Subdivision Scoping

Problems also result from the hierarchical characteristics of geographic terminology. This is exemplified by inconsistencies in the Library of Congress Subject Headings practice in assigning indexes. For example, the procedure for deciding whether islands should appear as high level entities, or as subdivisions, is discussed in [Well78]:

> "Now, catalogers in the Library of Congress can, of course, arbitrarily decide what is to be an island in its own right, entered directly, and what is to be subordinated to another geographical entity. But the more important question is: how can *users* (other than specially trained reference librarians) know what decision has been made? Does not use of **Agriculture-Italy-Sicily** lead one to expect **Agriculture-Italy-Sardinia**? Assuming there is no such heading, will the user know or be able to guess that the heading is **Agriculture-Sardinia** (because the distance from the mainland has been arbitrarily designated as 'some')?" (pp. 164-165)

As a consequence of the above issues, the Library of Congress Subject Headings exemplify difficulties with the use of text terms for geographic indexes. These issues have been noted by both researchers and users. Reporting on a survey of geological librarians, Mulvihill and Eaglesfield [MuEa87] write:

> "Problems [reported] include LC's use of inverted headings, e.g., Mexico, Gulf of; failure to use formation names; old fashioned terminology; and preference for subject over geographic area -- the last the most common complaint in this survey." (pg. 152)

Although some of these issues can be resolved in part by assigning multiple indexes to documents, the use of manually-generated textual subject headings plainly has many inherent difficulties.

2.2.1. Keyword Indexing

Because the volume of full-text documents available on-line is growing rapidly, automatic indexing is an increasingly desirable strategy. Currently, automatic indexing is limited to the use of keyword text indexing, in which georeferencing is merely a side-effect of the existing paradigm. In this model, indexing of text documents is done by matching words which are extracted from the document (either title words, or in more advanced systems, keywords from the abstract or full text) with words employed by the user in a query.

There are many problems with this simplistic approach. For example, ambiguous place names such as **Paris** (which could be in France, Texas, etc.) can not be resolved by keyword indexing strategies. Additionally, in keyword indexing, synonyms for a location are not recognized as representing the same geographic location. Furthermore, the topic of a document can not be clearly identified by keyword indexing or examination of author and title fields. For example, the fact that a document contains the string "Davis, California" does not mean that "Davis, California" is the topic of the document; the term "Davis,

California" may refer to a location which has little or no significance for the document. In contrast to keyword indexing techniques, a system which utilized spatial reasoning and geographic knowledge would be able to draw more sophisticated inferences about the topic of a document and display them to the user in a more intuitive, cartographically-based interface.

2.3. Coordinate Searching and Spatial Interfaces

Several bibliographic systems have allowed use of coordinates in searching for text documents. In fact, GeoRef used manually-generated coordinates as early as 1978 [Tahi89]. In these systems, the user generally specificies the coordinates of a box or point of interest. More recently, prototypes have been developed which allow map-based interfaces to documents including images and tabular data as well as text [RLG89], [Holm90]. However, none of these systems has included automatic assignment of coordinates. Therefore, the popularity of these systems has been limited by the lack of automatic indexing and by their primitive interface mechanisms.

3. Needs Assessment

This section details two empirical studies. In the first study, transcripts of bibliographic searches are analyzed to determine the number of bibliographic searches which have geographic components. A taxonomy of these components is presented. The second study illustrates the types of bibliographic searches that are required to support projects done in natural resources management.

The purpose of these studies is to demonstrate that information retrieval systems have a significant need for spatial indexing. Since this type of indexing is not supported in current systems, these models do not communicate with the user as effectively as they might.

3.1. A Study of Geographic Constructs in Domain-Unspecific Searches

Data from an on-line library catalog was analyzed to assess the frequency and composition of spatially-related queries.

3.1.1. Description of Data

A transcript was made of bibliographic searches done at university public-access terminals throughout California. The transcript recorded the searches performed on a weekday afternoon during the month of November, 1991. Approximately 6,700 searches were done in the database during this time.

The database examined was Melvyl's CATALOG database, one of several databases from which a user may issue bibliographic queries. This database was chosen because it contains a large number of records and is used for a broad variety of queries, as opposed to some of the other databases in the Melvyl system.

This database contains information on a number of document types, including books, audio-visual materials, maps, etc. held at University of California campuses, the California State Library, and the Center for Research Libraries. As of June, 1992, the database contained nearly seven million titles. The CATALOG database allows searches

by indices such as author, subject, and title.

3.1.2. Characteristics Examined in the Data

The following three questions were explored:

Q_1: Do people routinely use geographic terms when searching for documents?

Q_2: Can the geographic queries people issue to search for documents be categorized into distinct types?

Q_3: Is the effectiveness of bibliographic searches affected by the presence or the type of geographic searches?

3.1.3. Identification of Types of Geographic Constructs in Searches

The transcript was examined to create a taxonomy of searches which pertain to geographic location. The types of searches were organized into the following four categories:

Modification

In these searches, general subjects were constrained by geographic names. Three different syntactic structures of modifiers were observed in the data. The first was the use of a geographical adjective, e.g. **find su[2] italian divorce** or **find su australian weather**. The second was the use of a place name in an adjectival position, e.g. **find su california mills** or **find su utah ecology at[3] ucsu and csl**. The third structure was the use of a prepositional phrase containing a place name to constrain the locality of a search, e.g. **find su domestic violence in italy** or **find su weather in southern hemisphere**.

Triangulation

This type of search did not contain an explicit place name, but gave information which could be assigned a spatial location. For purposes of analysis, only data pertaining to *cadastral boundaries* (e.g. states, counties), *land use*, *bioregions*, and *endangered species* was counted. This type of information can commonly be referenced in a geographic information system (GIS). When several of these objects are referenced simultaneously, the intersection of their locations can be used to identify regions of interest to the user; hence the name triangulators. Some examples of searches classified as triangulators include: **find tw[4] biological significance of estuaries**, **find tw forest service timber**, and **find tw ostrich at ucr**.

Several objects were excluded which might have been construed as triangulators; the primary reason for their exclusion was not their lack of geographic portent, but rather their lack of utility in the system proposed in section 4 of this document. Because the

---

[2] **su** signifies a subject search in the CATALOG database.

[3] **at** followed by an argument in the CATALOG database restricts the search to the location of the argument; here, only documents which are at UCSU (University of Southern California) or CSL (the California State Library) would be shown to the user.

[4] **tw** signifies a title word search in the CATALOG database.

data for human-made objects such as railroads, mills, dams, etc. is not currently available in the Project Sequoia 2000 database, these items were not counted. Also excluded were cultural items from which geographic location could be inferred: *peoples* (e.g. Sioux), *religions*, *individuals* (e.g. Salvador Dali, Shakespeare), *diseases*, or *holidays* (e.g. Thanksgiving).

Authorship

This type of search requested information organized by the location of the origin of the document (as opposed to the location of the subject of the document). For example, the following searches were done: **find xpe[5] new york times**, **find pa[6] ann arbor**, and **find pa japan[7]**.

Subjects

Some searches were done on specific geographic entities. For example, the following searches were done: **find su kilimanjaro**, **find se[8] africa south of the sahara**, **find su hot springs, arkansas**, and **find su san francisco airport**.

Melvyl also provides indexing of documents by the physical location of the actual document, as above in **find tw ostrich at ucr**. Because the focus of this project is on full-text retrieval systems which will display documents directly to the user's screen, or queue them to the user's printer, this type of search was not included in the taxonomy.

---

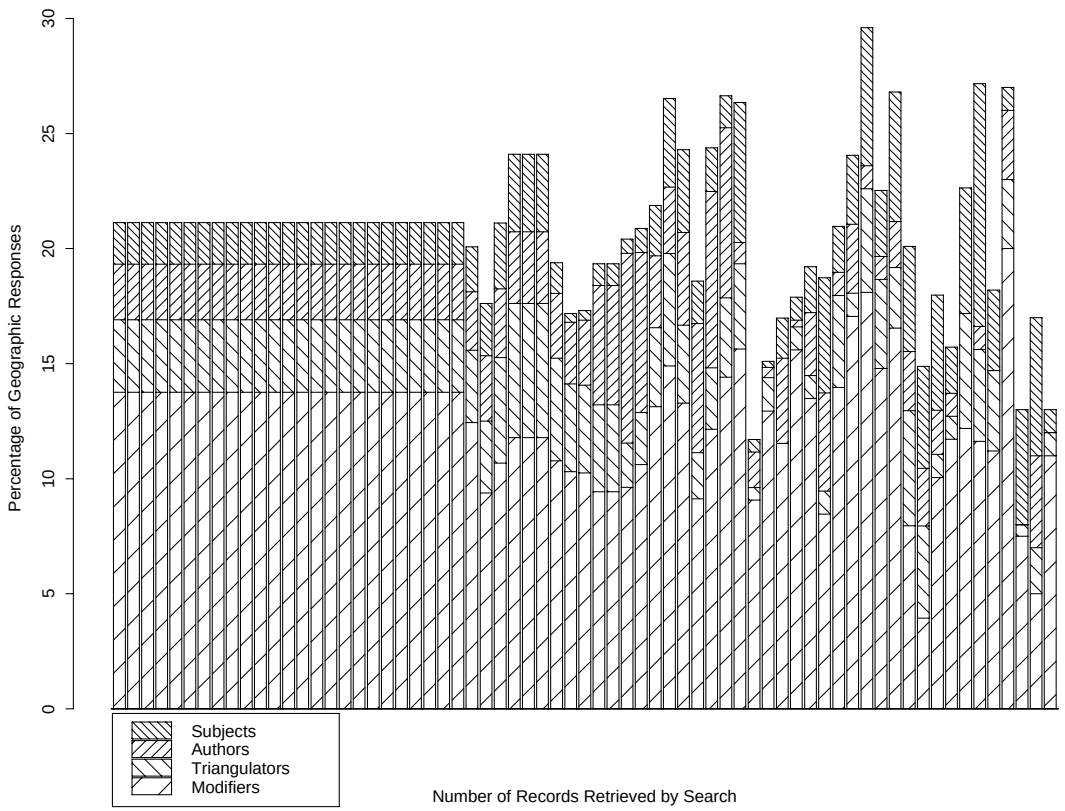
[5] **xpe** signifies a search for an exact periodical in the CATALOG database.

[6] **pa** signifies a personal author search in the CATALOG database.

[7] This search in the CATALOG database yields documents authored by the emperor of Japan.

[8] **se** signifies a search for a series in the CATALOG database.

Figure 1: Geographic Queries in Melyvl



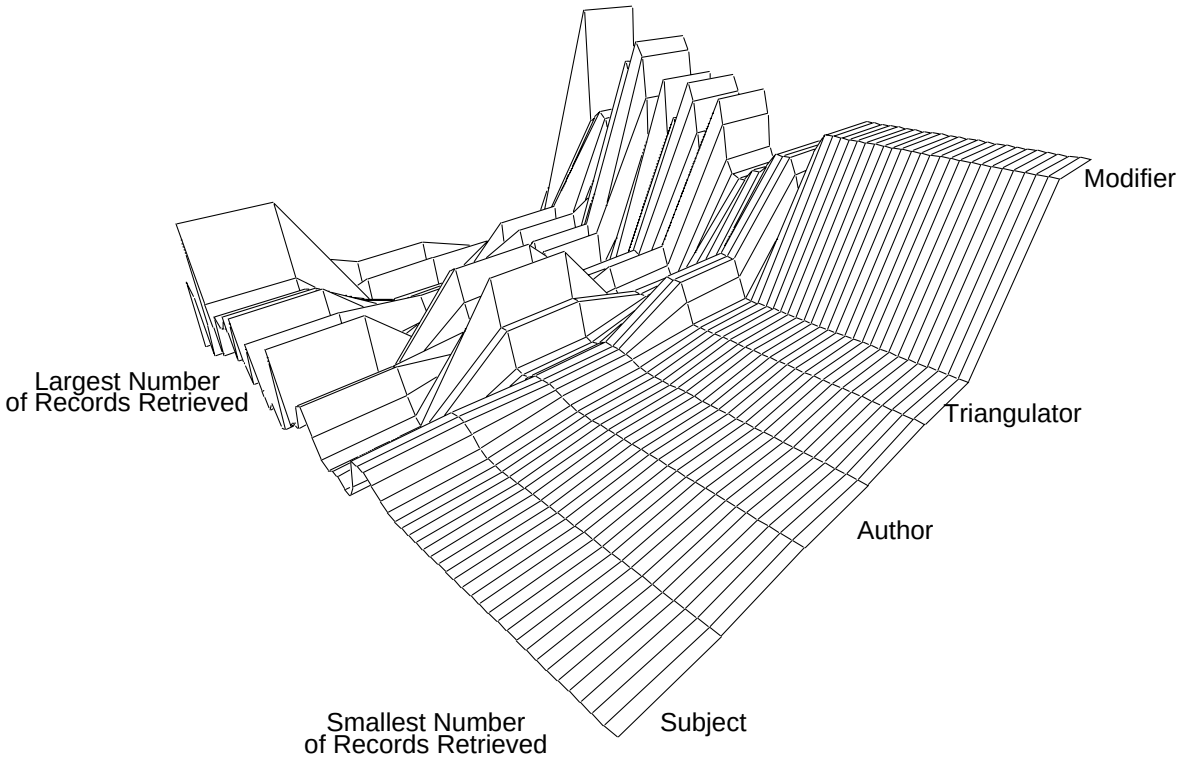
Procedure for Data Analysis

Several basic properties of the data were examined. The author looked at the data to see what percentage of each of the above four types occurred in the data set as a whole.

The author also took into account information about the effectiveness of the search, and studied whether successful searches had different geographic components than unsuccessful searches. Effectiveness can be inferred from the number of records retrieved per search. In the data set, the log of each search in the transcript reported the number of records returned by issuing that search in the CATALOG database.

The CATALOG database (like most systems) simply presents retrieved records in alphabetical order, giving no information about which records might be most pertinent to the user's search. As a consequence, the retrieval of too many records is virtually as ineffective as the retrieval of none at all. Larson [Lars91] reviews the literature on this problem further. He discusses the zero-infinity problem, the phenomena that in Melvyl and other systems the majority of searches return either zero records or so many records as to be useless.

Figure 2a: Types of Geographic Queries in Melvyl



Two properties which relate to the effectiveness of a search were examined. First, the author looked to see whether the percentage of geographic versus nongeographic searches varied by the number of records retrieved by a search. Second, the author examined the distribution of searches from each of the four taxonomies to see whether this distribution varied by the number of records retrieved by a search.

The tabulation of the data was necessarily a subjective process. Some of the difficulties included misspellings. Surprisingly, most queries with a misspelled word were not followed by a query with the term spelled correctly. Therefore, in the case of misspellings, the author inferred the meaning and counted the item which seemed to have been intended. This gave more accurate information about the number of geographic queries attempted. However, it slightly skewed the data about the number of records returned by the misspelled query, since the number of records returned by the misspelled query was used. Furthermore, searches with modifiers such as Utopia (of which there were a very limited number) were counted since these searches had the same syntactic structure as other modifiers.

The author counted the number of searches of each type done for each number of records retrieved, i.e. how many searches which returned zero records were modifiers, how many searches which returned zero records were triangulators, etc. In cases in which a search could fall into more than one taxonomy, a fraction was alloted to each taxonomy. In other words, if a search represented both a modifier and a triangulator, the

count of each of those categories was incremented by 0.5.

The searches were organized into 67 bins of 100. These bins were ordered by the number of records retrieved by the searches. Many of the bins represent a wide range of number of records returned. For example, the final bin contains searches which returned between 1,185 and 38,597 [9] records. In contrast, a few of the bins contain identical information because there were many searches which returned the same number of records. For example, more than 2,500 searches, nearly the first third of the searches, returned zero records. (In general, roughly 1/3 of Melvyl searches return zero records [Berg92].) Therefore, the first third of the bins have the same value. This method of presentation was chosen to more accurately represent the distribution of the data values.

3.1.4. Presentation of Data

Figure 1 shows the number of records retrieved per search versus the percentage of geographic searches. Each bar represents the percentage of geographic searches in one of the 67 bins. The percentage of geographic searches in the bins ranges from 11.7% to 29.6%, with 21.1% being the overall average. Each bar is broken down into one of the four types of geographic searches.

[9] The search in the CATALOG database which returned this many records was **find pe journal**.

Figure 2b: Types of Geographic Queries in Melvyl

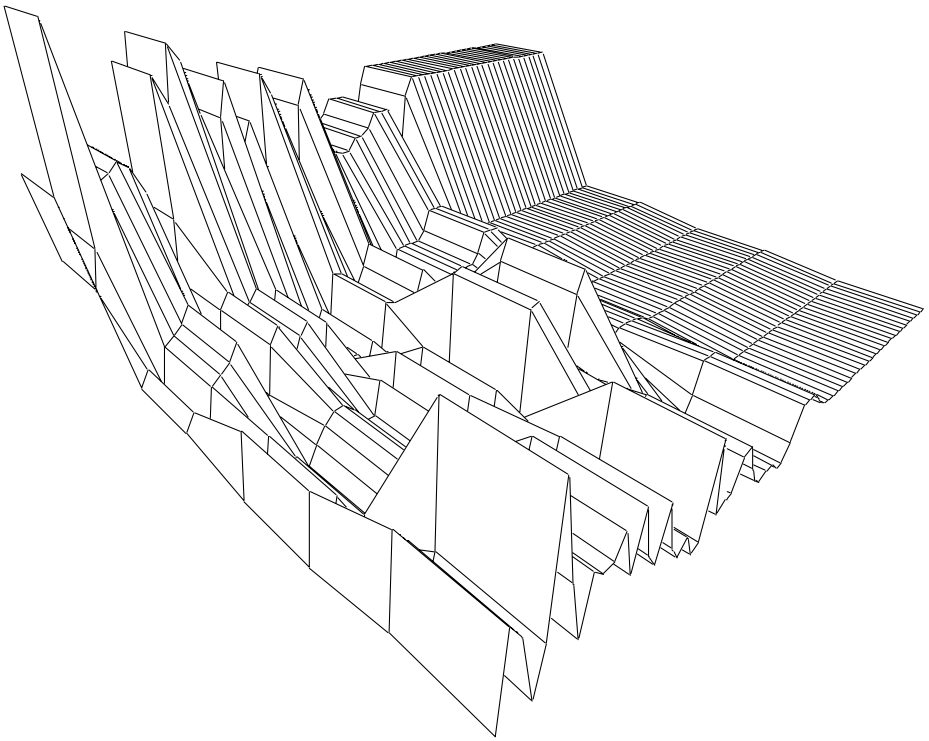


Figure 2a shows a contour plot of the distribution of the types of geographic queries in Melvyl. There are seven ribbons; the four flat ribbons represent each type of search, subject, author, triangulator, and modifier, from nearest to furthest. The height of the ribbon at any given point represents the percentage of geographic searches of that type done in that bin. The three unmarked slanted ribbons show differences between each adjacent category of geographic search. The location on the ribbon shows which bin is being referenced; the closer bins represent the searches which returned zero records, while the far bins represent increasingly higher numbers of returned records.

Figure 2b shows the contour plot of Figure 2a from a different perspective.

3.1.5. Interpretation of Results

Although a formal statistical analysis was not considered necessary, there are several conclusions which can be drawn from the data.

Q_1: Do people routinely use geographic terms when searching for documents?

A_1: A meaningful percentage (over 21%) of searches which were studied contain a geographic construct.

Q_2: Can the geographic terms people issue to search for documents be categorized into distinct types? If so, do some of these types occur more frequently than others?

A_2: A taxonomy of distinct geographic constructs has been presented in section 2.1.3. It was shown in section 2.1.5 that some constructs occur more frequently than others. For example, the modifier construct is clearly the most common.

Q_3: Is the effectiveness of bibliographic searches affected by the presence or the type of geographic searches?

A_3: Examination of the data did not reveal any meaningful patterns of variation related to the number of records retrieved by a search. Neither the percentage of geographic searches nor the distribution of the types of geographic searches appears to have impacted the number of records retrieved. The conclusion that geographic indices are not significant points of reference for retrieval does not follow from this, however. One plausible explanation is that since the CATALOG database does not support true geographic indexing (using only direct matching of key words in title and author fields), the geographic component of a search is not used by the database as a real index, and so does not impact the effectiveness of the search in a significant way.

3.2. A Study of Geographic Constructs in Domain-Specific Searches

The goals of the California Department of Water Resources (DWR) are essentially the following: (1) predict how much water of what quality is needed where [census data, land use analysis] and the priorities for supplying this water [health and safety, economics, legislation]; (2) find the means to supply as much of this water as possible [hydrology, water quality studies, meteorology, construction], taking into account the impact on environmental and cultural resources [plant and animal species, archaeological data]; and (3) provide the public with information [publications, movies, press releases].

To determine the types of queries DWR makes on these topics, sample queries stated by staff during interviews were noted and DWR contracts relating to studies and literature searches were reviewed.

Table 1: Sample Queries Issued by DWR Staff

- Find all tech reports on archaeological findings **in area X**.
- Find all legislation related to archaeology **in area X**.
- What dredging [deepening] **of the Delta** has occurred during the last 20 years?
- If we put a barrier **at location X**, how will it affect salinity and water elevation **at location Y**? [The engineers have to ensure that sufficient water is released into the Delta to prevent the sea water from intruding.]
- If we enlarge **forebay X**, or pump a different amount of water out **of forebay X**, what will happen to flows, elevations, and salinity **at location Y**?
- Find all research **on Twitchell Island** that says Lilaeopsis masonii [an endangered plant in the Delta] exists there.
- Show the distribution of **kit foxes in CA**.
- Given a map of all endangered species, show all regions where **blunt-nosed leopard lizards** live, then show all studies on the **lizards in this area**.
- Show **land ownership** and **land use change** through animation.
- Find all articles on obstruction of **stream** flow.
- What are the types of pesticides used on a certain **crop**?
- Is food limiting to **striped bass larvae in estuaries** where **striped bass** are found?
- Are there fewer **kit foxes** now than last year?
- What is the incidence of **ponds** in relationship to **species X**?
- Find all newspaper articles which discuss **the State Water Project**.
- Find all newspaper articles which oppose **the Coastal Branch of the State Water Project**.

3.2.1. Sample Queries

DWR staff were interviewed to determine their data retrieval needs. In informal discussions, their data analysis and retrieval processes were discussed, and sample queries elicited. These conversations did not focus on any specific topic, e.g. geographic indexing, but on the database requirements of their work.

It can be seen that all DWR searches which were elicited contain a geographic component. The type of geographic search is almost always modifier, triangulator, or some combination thereof. The pertinent components are represented in bold font in Table 1.

Many of the searches are combinations of the categories outlined in section 3.1.3. The searches are more complex because (1) DWR engineers are experts and (2) the Melvyl search language limits the syntax and scope of queries.

As part of the interviews, staff made it clear that the ability to reference text and other data by geographic location would be extremely useful. Many staff said they would like to begin searches by selecting a geographic area, and then ask questions about documents pertaining to that area.

3.2.2. Excerpts from Contracts

Consultants often do bibliographic searches for DWR, as well as unique studies. These contracts are not a comprehensive representation of DWR activities because DWR tends to contract in areas in which it has less expertise. If better research tools were available, it might be cost efficient to do more of these studies in-house.

In Appendix 1a-f are excerpts from some of the Division of Planning contracts (in which DWR contracted with private institutions or other State Agencies, or was the contractor itself). My selections are based on a quick perusal of the material from the past decade or so. For the sake of brevity, only short passages have been included, some of which have been paraphrased.

Since the Department of Water Resources is responsible for managing water in California, staff queries are by nature project-oriented. This pragmatic bent means that many queries pertain to the design or maintenance of a proposed or existent project. Each of these projects is associated with a spatial location. It can be seen that the questions detailed in Appendix 1a-f each contain a spatial modifier of some kind; in some cases this modifier is simply the State of California.

For example, in Appendix 1a, the economy and agriculture of California and specific regions within California are discussed. Appendix 1b shows questions relating to the impact of pesticides, oil fields, etc. on water quality in particular regions or at particular projects. Appendix 1c contains questions on the behavior of named water bodies. In Appendix 1d, queries relating to the environmental characteristics of different regions are shown. Appendix 1e illustrates searches on the cultural heritage of specific locations. Finally, Appendix 1f shows questions regarding the feasibility of constructing projects in precise locations.

Furthermore, although each of these project-oriented searches inherently contains a geographic reference to the project location, examination of the feasibility of the project may also require more abstract searches pertaining to theoretical properties of geographic characteristics rather than precise location, e.g. "Find me all articles on obstruction of **stream** flow". Presumably, the project-oriented searches would tend to contain modifiers, while the more theoretical searches would have a higher percentage of triangulators.

Finally, searches might also be structured to find analogous projects in other locations, e.g. "Find me articles on **dams** that are similar to **Shasta Dam**." These searches entail not only an initial modifier (here, Shasta Dam), but a triangulator (if the other item is not specified), or a subject or modifier (if the desired analogous project is known by name).

The data above demonstrate that DWR has a significant need to reference text by geographic location.

3.3. Summary

The Melvyl searches demonstrate that geographic constructs are common components in general academic searches. The DWR queries demonstrate that in the domain-specific searches of natural resources managers, geographic delimiters occur frequently.

4. Solution

This section describes the general philosophy of GIPSY. A prototype which implements a subset of the features described in this section is presented in section 5 below.

4.1. General Strategy

GIPSY's basic strategy is to extract place names as well as more general geographic indicators from documents. The intersection of these referents is used to generate estimations of the area to which a document refers. For example, if a document discusses a proposed dam site near the habitat of eagles and near a large city, the system should return a list of locations of water bodies near the habitat of eagles and near an urban area. The method to return this list of locations consists of three primary steps:

Step One: Parsing

The first step is to extract relevant key words and phrases. As an example of the desired output, consider a document generated by the California Department of Water Resources. The Department of Water Resources stores water below ground in the San Joaquin Valley. This water is placed in the water table by the intentional flooding of a specific region, the Kern Fan Element, in Southern California. A document regarding the Kern Fan Element might contain the following geographic words and phrases to be extracted:

Information Type	**Examples**
place names and relationships	**in Kern County** **in the State of California** **north of the Tehachapi Mountains** **proximate to the San Joaquin Valley Water Basin** **not related to the Kern River**
geologic features	**aquifers**
climate	**semi-arid**
land use	**agricultural land taken out of production** **barley growth**
endangered and threatened species	**blunt-nosed leopard lizard** **Tipton's kangaroo rat** **San Joaquin kit fox**
size	**20,000 acres**

The parser for this strategy must have a way to identify geographic terminology. Relevant terminology is of two types:

(1) terms which relate to objects or attributes in the data set. This step entails a significant amount of thesaurus work.

(2) lexical constructs which contain spatial information, e.g. "contiguous", "adjacent", etc. To implement this, a list of the most commonly occuring constructs must be created.

Noun phrases and adjectival phrases such as "coastal zone shrubland" or "south of the Delta" should also be recognized; this can be done by extensions to existing part-of-speech taggers.

Once the geographic components of phrases have been identified, a certain amount of syntactic work will be required to omit the irrevelant portions of phrases, e.g. "south of the Delta" versus "literature on the Delta". In the former, the entire phrase should be preserved, while in the latter, all but the keyword "Delta" should be omitted.

Another issue pertaining to parsing is the granularity of the document to be indexed. Furthermore, extraction of the relationships among the key words and phrases will also be of interest. One useful relationship should be extracted by segmentation work being done by Hearst [Hear92]. In segmentation, consideration of locality and density of keywords is taken into account. This will facilitate making distinctions between distinct geographic areas which are represented within a document. For example, if sycamores and Los Banos Grandes are mentioned in one portion of a document, and snow and the Sierra Nevada Mountains are mentioned in another portion, sycamores and Los Banos Grandes should be construed as representing one geographic entity, and snow and the Sierra Nevada Mountains should be interpreted as referring to another.

Step Two: Locating pertinent data

The output of the parser is passed to a function which locates data pertinent to the extracted terms and phrases. For example, the data sets available may contain a list of coordinates describing habitats in which blunt-nosed leopard lizards live. Spatially-indexed data used in this step will include, for example, name, size, and location of cities, states, etc.; name and location of endangered species; name, location, and bioregional characteristics of different climatic regions; etc. The system must identify locations which most closely match the geographic terms extracted by the parser.

There are several types of data which will be relevant to extracted keywords and queries: (1) tabular data, e.g. verbal (Redding: Shasta County), point (Redding: latitude X, longitude Y), in situ (Redding: 110 degrees); (2) polygonal data, e.g. vector (Redding: vector X, vector Y, etc.); and (3) raster data, e.g. satellite , model output (grid block: vegetation type X). It is necessary to develop a method of mapping extracted keywords and phrases into the terminology used in these data sets. This can be done with a thesaurus incorporating the following types of information: (1) synonyms (e.g. latin and common names); (2) hierarchical information (e.g. an egret is a type of bird); (3) membership information (e.g. wetlands contain...); and (4) quantification of terms (e.g. "brackish").

Some of the difficulties of constructing such a thesaurus are discussed in Tahirkheli [Tahi89]:

> "In theory a simple hierarchical relationship sounds neat and clean. In the real world things get a little sloppier. Geographic terms are a sloppy area. As non-unique terms, a simple hierarchical relationship is not sufficient to deal with the complexity of geographic place names. Counties are a good example of overuse of a single term. Washington County appears in thirty-two states. Terms can be altered to accommodate overuse. The Washington County in Arkansas can be indicated by a compound term -- Washington County (Arkansas). In GeoRef, terms have traditionally been used in their simplest form. All Washington counties are identical in form. Users are able to identify the Washington County of their choice through the use of coordinate searching... terms that are part of multiple hierarchies where it is desirable for more than one broader term to be automatically added to each record. This applies, for the most part, to minerals. Some minerals are classifiable as more than one type of major mineral group..." (pp. 90-91)

The thesaurus must also consider the difficulties of integrating different types of data and combining multiple (possibly conflicting) taxonomies.

Once the appropriate data sets have been retrieved, they must be processed by a probabilistic function to return a list of polygons, each of which has an associated relevance weighting. The relevance weighting is based on (1) the geographic terms which have been extracted, (2) the location within the document and the density of these terms, (3) knowledge of the geographic objects in the database and their attributes, and (4) spatial reasoning about the geographic constructs occuring in phrases such as "adjacent to California" or "within a one hundred mile radius of Los Angeles".

Step Three: Performing the polygonal overlay

The final step is to perform a polygonal overlay of the data output by Step Two. This step will identify areas which have a large relevance weighting. The polygons and their associated weights must be overlayed to identify areas which have a high overall relevance weighting. This requires computational functions which can identify the intersection of polygons, as well as create an accumulated relevance value for each polygon.

5. Implementation of GIPSY

A prototype implementation has been completed. The core of this prototype is a thesaurus which establishes a relationship between natural language terms and the real-world polygons they indicate. Because data and a text corpus could be compiled for California, California was chosen as the test area. The thesaurus contains references to three kinds of data about California:

Place Names

The first data set used is from the US Geological Survey's Geographic Names Information System [USGS87]. This data set contains latitude/longitude point coordinates associated with over 60,000 geographic place names in California. To facilitate comparison with other data sets, the GNIS latitude/longitude coordinates were converted to the Lambert-Azimuthal projection. Examples of place names with associated points include:

```
University of California Davis: -1867878 -471379
Redding: -1863339 -234894
```

Feature Types

Each place name in the GNIS data set also contains a feature type. Therefore, it was possible to generate a list of points associated with each of the feature types listed in Table 2.

airport	arch	area	bar	basin
bay	beach	bench	bend	bridge
building	canal	cape	cave	cemetery
channel	church	civil	cliff	crater
crossing	dam	falls	flat	forest
gap	geyser	glacier	gut	harbor
hospital	island	isthmus	lake	lava
levee	locale	mine	oilfield	other
park	pass	pillar	plain	ppl[10]
private	range	rapids	reserve	reservoir
ridge	school	sea	siding	site
slope	spring	stream	summit	swamp
toll	tower	trail	tunnel	valley
well	woods			

Table 2: GNIS Feature Types

Land Use Type

Another data set used in this implementation consists of land use data from the US Geological Survey's Geographic Information Retrieval and Analysis System (GIRAS). This data is categorized according to the Anderson Classification System [AHRW76]. This GIRAS data set contains over 60,000 entities in the Universal Transverse Mercator projection. These entities were processed to be explicitly represented as polygons in the Lambert-Azimuthal projection. The data sets contains polygons associated with each of the land use types listed in Table 3.

[10] populated place

Table 3: GIRAS Land Use Classifications

Urban or Built-up Land	Residential
	Commercial and Services
	Industrial
	Transportation, Communications and Utilities
	Industrial and Computer Complexes
	Mixed Urban or Built-up Land
	Other Urban or Built-up Land
Agricultural Land	Cropland and Pasture
	Orchards, Groves, Vineyards, Nurseries, and Ornamental Horticulture Areas
	Confined Feeding Operations
	Other Agricutural Land
Rangeland	Herbaceous Rangeland
	Shrub and Brush Rangeland
	Mixed Rangeland
Forest Land	Deciduous Forest Land
	Evergreen Forest Land
	Mixed Forest Land
Water	Streams and Canals
	Lakes
	Reservoirs
	Bays and Estuaries
Wetland	Forested Wetland
	Nonforested Wetland
Barren Land	Dry Salt Flats
	Beaches
	Sandy Areas Other than Beaches
	Bare Exposed Rock
	Strip Mines, Quarries, and Gravel Pits
	Transitional Areas
	Mixed Barren Land
Tundra	Shrub and Brush Tundra
	Herbaceous Tundra
	Bare Ground
	Wet Tundra
	Mixed Tundra
Perennial Snow or Ice	Perennial Snowfields
	Glaciers

Many relevant terms do not exactly match place names or the feature and land use types listed above. For example, **alfalfa** is a crop grown in California, and should be associated with the crop data from the GIRAS land use data set. The thesaurus was therefore extended to include the following types of terms:

synonymy

`= := synonym`

kind-of relationships

`~ := hyponym` (maple is a ~ of tree)
`@ := hypernym` (tree is a @ of maple)

part-of relationships

`# := meronym` (finger is a # of hand)
`% := holonym` (hand is a % of finger)
`& := evidonym` (pine is a & of shortleaf pine)

Place names do not have synonyms or kind-of relationships, but do have part-of relationships. An evidonym of a place name is defined as a word which occurs as a component of a place name. For example, **Cachuma** is an evidonym of **Lake Cachuma**. These evidonyms were calculated automatically by extracting words from the place name phrases. Because they had little significant content, tokens consisting entirely of function words such as **of** and **the** (e.g. from **Arch of the Navarro**), were deleted from the final thesaurus, as were numbers (e.g. from **2 C No 1 Drain**) and words consisting of single letters (e.g. from **A F Traver Ranch**)[11].

Feature types can have synonyms, kind-of relationships, or part-of relationships. For example, **airfield** is a synonym of **airport**, and **university** is a kind of school. The terms for these relationships were derived from the GNIS definitions, as well as from a variety of thesauri and dictionaries, including WordNet.

Land use types can have synonyms, kind-of relationships, and part-of relationships. Synonyms, kind-of relationships, and a number of part-of relationships were taken from the definitions of the Anderson Classification scheme. A few hundred part-of relationships were taken from the California Department of Water Resources land use classification definitions, including primarily lists of crops grown in California. Several thousand part-of relationships were also extracted from the Wildlife Habitats Relationships Database from the California Department of Fish and Game. These consisted of lists of common and scientific names of plants and animals which occur in specific land use types. For example, the **California Slender Salamander** is a meronym of the GIRAS land use data type **Deciduous Forest Land**. Once the synonyms, holonyms, etc. associated with the land use data were found, evidonyms of these terms were computed. For example, **Sequoia Sempervirens** is a meronym of **Evergreen Forest Land** which is further broken down into the evidonyms **Sequoia** and **Sempervirens**.

---

[11] There are some disadvantages to this, e.g. **C** appears in the GNIS data set as place name for a canal.

The thesaurus includes over 200,000 entries in the form:

<entry> := <char_str>:<level>:<attr>:<source>:<coord>

<char_str> := place name | word in place name
<level> := 1 | 2 ...
<attr> := <feature> | = | ~ | @ | # | % | &
<source> := gnis | giras | <char_str>
<coord> := <file-name> | <polygon>
<feature> := authorized gnis/giras features
<file-name> := unix file name
<polygon> := list of verticies

<char_str> is the natural language term which the parser searches for in documents. <level> indicates the amount the entry is removed from the original GIRAS or GNIS terms. A direct match is at level 1, e.g.

```
beach:1:beach:gnis:beach
```

Indirect matches have increasingly large levels, e.g.

```
apartment:2:#:commercial and services:giras12
```

indicates that the term **apartment** is a meronym at a level of two from the GIRAS data set which contains the **Commercial and Services** land use polygons for California.

Thesaurus entries such as those above are used in the steps outlined in Chapter 4:

Step One: Parsing

Parsing is done by matching tokens which occur in the thesaurus. Tokens are either single words, e.g. **Fresno**, or phrases which include spaces, e.g. **San Francisco**. Text to be input to the parser is stripped of all punctuation, and therefore consists of a string of single words separated by spaces. The parser, which is written in emacs-lisp, advances through the text. For each word in the text, if the word begins a token in the thesaurus, the longest exact matching (matching is case-insensitive) token is stored in the list of matched tokens, and the pointer in the text is incremented to the end of that token. If the word is not found in the thesaurus, the word is run through a morphology function to depluralize it and the depluralized version is checked against the thesaurus, e.g. **valleys** would fail to match, and would be transformed to **valley**, which would match tokens in the thesaurus.

Step Two: Locating Pertinent Data

Given a list of matched tokens, several operations must be done on the thesaurus, which serves as a look-up table. The operations are performed by a variety of functions written in emacs-lisp, C, and perl.

- The point, filename, or polygon associated with each token must be extracted.

- Points (which come from the GNIS data) must be expanded into polygons. A bounding box is therefore calculated for each of these points, the size of which is dependent on the feature type of the location. To calculate approximate sizes, feature types have been designated small (100 meters), medium (1 kilometer), or large (10 kilometers).
- Because the files containing the land use polygons are large (several megabytes apiece in some cases), the landuse polygons are not stored in the lookup table, but must be read in dynamically from files.
- The probability that a polygon is correlated with the document from which the natural language terms are extracted is dependent on a number of factors. Currently, the formula used to calculate the expected relevance of the polygon is equal to *the level of the term* multiplied by *the number of times the term occurs in the thesaurus* multiplied by *the number of polygons which are associated with the term* (this factor is only relevant for GNIS feature types or for GIRAS files with multiple polygons for a single land use classification) multiplied by a discrimination factor[12].

For example, the term **facility** appears in the thesaurus seven times, as a level 2 evidonym of each of the following:

California Medical Facility
Delta Fish Protective Facility
Glen Helen Rehabilitation Facility
Hogback Island Recreation Facility
Lacy Security Facility
Mira Loma Detention Facility
United States Naval Missile Facility Point Arguello

facility is also in the dictionary. Therefore, the polygon associated with each of the above place names is assigned a weight of *2*7*8 = 112*. In another case, the term **western spadefoot** appears as a token in the thesaurus twice, because it appears in two habitats, e.g. it is a level 2 meronym of **Deciduous Forest Land**. Since the **Decidous Forest Land** data sets contains 538 polygons, and the term is not in the standard dictionary, the weight of **western spadefoot** is *2*2*538 = 2152*.

In this weighting scheme, the higher values indicate a lower probability of relevance, so the weights are inverted, and the result of the inversion is used as the relevance indicator passed to the polygon function. Because the polygon functions are compute intensive and to eliminate noise, polygons with a weight below a threshhold designated by the user are deleted.

---

[12] A discrimination factor based on the commonality of the word was added as a result of unnaturally high values for some of the evidonyms. For example, a town called **Quality** was hit in sentences such as "In addition to the potential degradation caused by sea water intrusion, further degradation could occur from the upward migration of inferior **quality** ground water found in deeper zones into the producing zones." or "Reclaimed water is the treated effluent from domestic waste water treatment plants that can be utilized for specific uses if it meets standards established by the California Department of Health Services and the discharge requirements of the California Regional Water **Quality** Control Board." To mitigate this effect, tokens are checked against the standard Unix dictionary, and if they are present, their weights are decreased.

For each entry, in addition to the polygon and its relevance weighting, the term from the text and the token it matched are also preserved for purposes of analysis. An example of the final input to the polygon function is:

```
("2D" "facility" "delta fish protective facility" 0.07142857
(-1873352 -551201 -1875352 -551201 -1875352 -553201 -1873352
-553201 -1873352 -551201))
```

Step Three: Overlaying Polygons

The objective of the polygon function is to take in a number of polygons, some of which will overlap, and calculate the cumulative relevance for each location in California. In this implementation, each polygon is represented as a **thick** polygon, a polygon with a base positioned in an *x,y* plane which extends upwards a distance of *z* to the same *x,y* coordinates in a higher, parallel plane (Figure 3a). These thick polygons are laid onto California to form a **skyline**. As polygons are added, three cases may arise. First, if the polygon being added to the skyline does not intersect with the *x,y* coordinates of any other polygons, it is simply laid on the base California map beginning at $z = 0$ (Figure 3b). Second, if the polygon being added to the skyline is completely contained within a polygon which already exists on the skyline, it is laid on top of that polygon, i.e. its base is positioned in a higher *z* plane (Figure 3c). Third, if the polygon being added to the skyline intersects but is not wholly contained by a polygon(s) in the skyline, the polygon being added is split, and the intersecting portion is laid on top of the existing polygon and the non-intersecting portion is laid at a lower level (Figure 3d). To minimize fragmentation in this case, polygons are sorted by size prior to being positioned in the skyline.

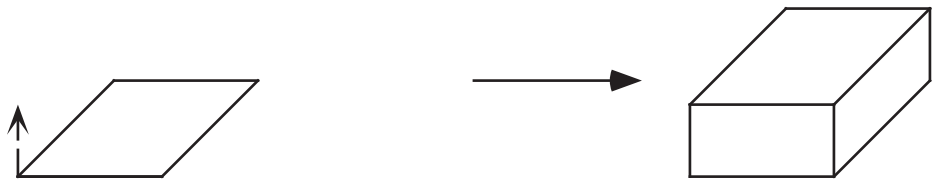


Figure 3a

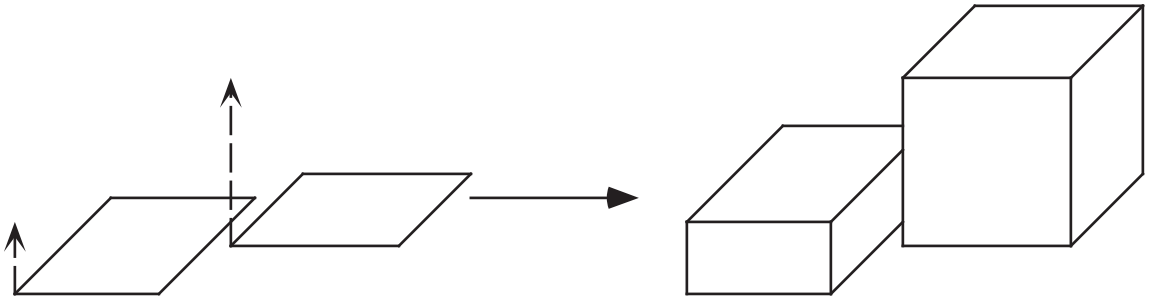


Figure 3b

Figure 3c

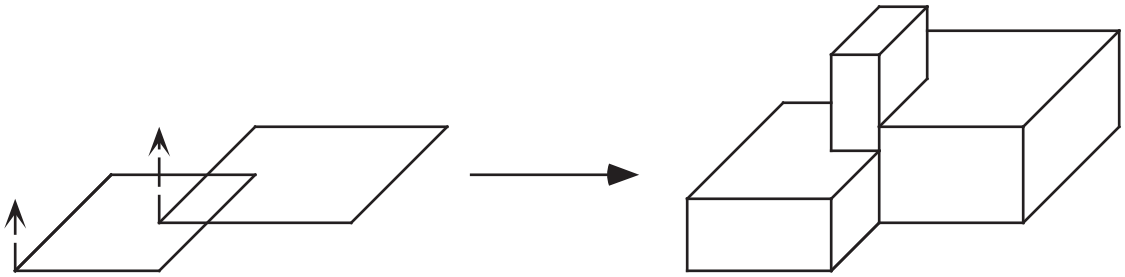


Figure 3d

Overlay is performed with a library of graphics functions written in C. With these functions, concave polygons are broken into lists of convex polygons which can be clipped against each other to determine overlap.

The resulting skyline is then plotted. Currently, the polygons are mapped onto a regular grid and displayed as a surface plot.

Throughout the process, difficulties arose from the size of the data sets, and many optimizations have been employed. For example, polygons with weights below a certain (user-specified) threshhold are not passed to the stacking program. Because the weights of these polygons are negligible (often in the millions), it is not expected that this will affect performance.

Performing all of the steps above on a single paragraph currently consumes between approximately 5-30 minutes of CPU time. The variance in time is controlled primarily by the number of polygons passed to the stacking program, which uses an n^4 algorithm. In 30 minutes, approximately 3,000 polygons can be stacked.

6. Results

Experimentation with GIPSY has begun, varying parameters such as the size of the threshhold below which polygons are deleted, the amount of the various weighting factors, and the size of the text block being parsed. Experiments have also been run with and without the GNIS and the GIRAS data to determine their usefulness.

Texts presented throughout this section are publications of the California Department of Water Resources. Examination of GIPSY's behavior on various passages has led to several general observations:

- Exact place name data is very useful.

There are many examples in which exact place name hits were correct indicators of the location of the paragraph. For example, GIPSY has processed the following text:

> Amphibians were found in water or relatively moist surroundings. Bullfrogs were found in Lake Cachuma or in the moist areas where the lake had receded. California slender salamanders preferred woodland habitat. Pacific tree frogs were found in and around the lake and in riparian areas of streams leading into the lake. A western toad was located in the lake and a western spadefoot was in grassland near the lake.

Given this input, GIPSY generates a surface plot of California in which the area being discussed in the text is represented as a high tower (Figure 4).

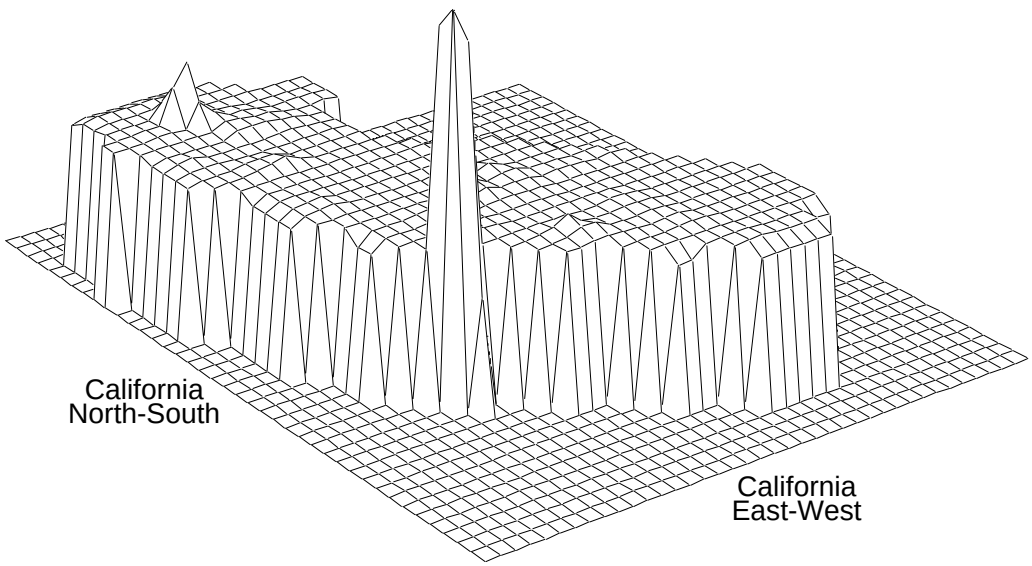


Figure 4

In most cases, place name ambiguity did not appear to be a problem with the California data set applied to this document corpus.

- GIRAS data is currently most useful when place names are not present.

In many cases, the GIRAS data was removed by threshholding. This is partly because it was difficult to devise an appropriate weighting scheme to balance the GNIS and the GIRAS data. One issue is that dividing the weight by the number of polygons does not seem completely appropriate. Another difficulty is the granularity of the data; a reference to a single species could generate a reference to tens of thousands of polygons.

It appears that more specific data would yield better results[13].

GIPSY processed the following text, without dividing the weight by the number of polygons:

> The proposed extraction conveyance facilities consist primarily of buried PVC pipeline, small concrete-lined canals, inlet structures at the Cross Valley and Alejandro canals, road culverts, and other miscellaneous small canal structures.

The resulting plot appears in Figure 5. Note the density of reference in the north part of the state, in which there are many canals, and the low reference to the south-eastern part of the state (desert area, in which there are few canals, other than in the Colorado River area).

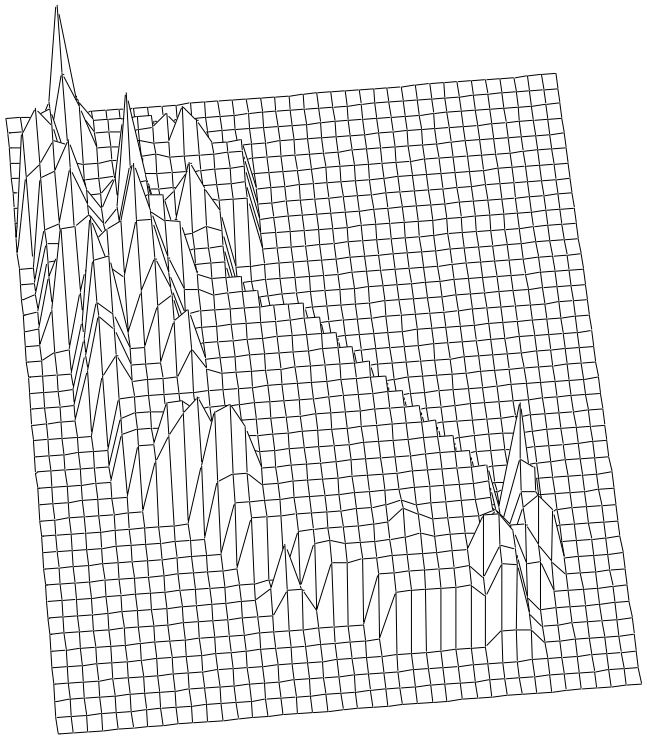


Figure 5

- Weighting is crucial.

[13] It is also hoped that the system will scale smoothly to different levels of detail, functioning on large geographic entities as well as small. For example, if specific detail about the plant distribution on Twitchell Island were entered into the database, and a document referring to that area were indexed, the georeferencing of the document could be extremely specific, stating which portion of the island was being referred to. In short, the more detailed and precise the data in the database, the more specific the indexing.

The weightings have been chosen modeled on techniques from a variety of previous works in information retrieval. All parameters chosen appear meaningful; varying virtually any parameter has significant impact on results. Minor experimentation has been done to explore the appropriate relative value of these parameters. The complex tradeoffs among these factors should be further explored. For example, lighter weighting of evidonyms is advantageous in some cases, but misleading in others. Further, higher weighting of GIRAS data should allow the GIRAS data to assume more significance, but currently introduces too much noise; a better relative weighting must be found for this data.

- Many of the evidonyms were useful, and others were not.

Many evidonyms seemed useful. For example, **Kern Water Bank** is not in the thesaurus. Fortunately, **Kern County**, which completely contains the Kern Water Bank, is in the thesaurus, and will be retrieved by the evidonym **Kern**. However, **Los Banos Grandes** is not found in the thesaurus, so GIPSY matches to the evidonym **Grandes** and returns **Piedras Grandes**. further, many evidonyms such as **year** in **New Year Mine** received unnaturally high weights. Although GIPSY currently uses the UNIX dictionary to evaluate the commonality of a term, this strategy is not adequate; more sophisticated discrimination techniques are necessary.

- Better parsing is needed.

The lack of sophisticated parsing led to a significant amount of noise, contributing for example to misleading evidonyms such as **buried** indicating **Buried Mountain**. Furthermore, all parsing was case-insensitive, which had several benefits, but also created noise.

- Polygon manipulation is resource intensive.

Stacking of polygons consumed a significant amount of coding time, and is by far GIPSY's most CPU-intensive task. In retrospect, it seems that prototyping should have been done by mapping the polygons onto a regular grid and manipulating the associated grid cell numbers.

- The overall strategy is promising, but needs to be more sophisticated.

GIPSY ran on the following text:

> The proposed project is the construction of a new State Water Project (SWP) facility, the Coastal Branch, Phase II, by the Department of Water Resources (DWR) and a local distribution facility, the Mission Hills Extension, by water purveyors of northern Santa Barbara County. This proposed buried pipeline would deliver 25,000 acre-feet per year (AF/YR) of SWP water to San Luis Obispo County Flood Control and Water Conservation District (SLOCFCWCD) and 27,723 AF/YR to Santa Barbara County Flood Control and Water Conservation District (SBCFCWCD). This is known as the downsized project. SBCFCWCD's full SWP entitlement is 45,486 AF/YR but they would not receive their full entitlement under the proposed project. Distributing SWP water from the proposed Coastal Branch to service areas within San Luis Obispo and Santa Barbara counties will require local distribution facilities. This EIR includes one such facility, the Mission Hills Extension, but other local

ancillary facilities are to be addressed in separate environmental documents. The proposed down-sized project would not deliver the full entitlement to serve areas beyond northern Santa Barbara County. If Santa Barbara County requests its full entitlement, an additional water distribution facility, the Santa Ynez Extension, would be needed. This extension would serve the South Coast and Upper Santa Ynez Valley. DWR and the Santa Barbara Water Purveyors Agency are jointly producing an EIR for the Santa Ynez Extension. The Santa Ynez Extension Draft EIR is scheduled for release in spring 1991.

GIPSY generates Figure 6, a useful index for the document.

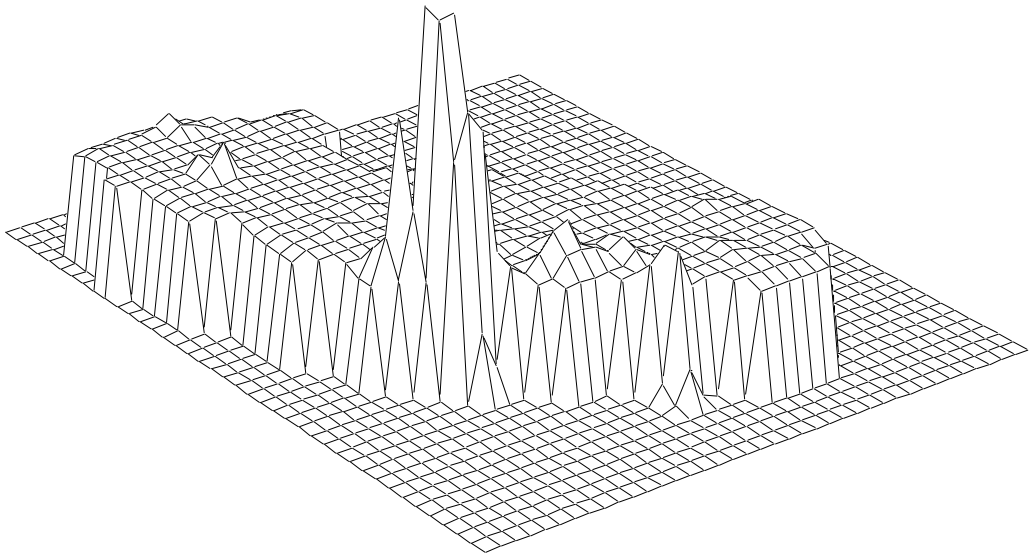


Figure 6

7. Future Work

Promising research includes several extensions to the existing GIPSY implementation:

Parsing

Because a geographic knowledge base and spatial reasoning are fundamental to the georeferencing process, they have been the focus of initial research efforts. As this fertile area is explored, the existing prototype can be complemented by sophisticated natural language processing. For example, spatial reasoning and geographic data could be

combined with parsing techniques to develop semantic representations of the text. Part-of-speech taggers could be used, particularly to identify noun phrases. Adjacency indicators such as "South of" or "between" could also be recognized by a parser. Finally, work being done in Project Sequoia 2000 to segment documents could be used to explore the locality of reference to geographic entities within full-text documents. For example, GIPSY's technique may work more effectively on paragraphs or other subcomponents of a text than on an entire document.

Locating pertinent data

To support the data management capabilities desired by Earth System scientists who perform complex analyses and simulations, POSTGRES will be expanded to contain the functionality of a geographic information system, and will allow searches on the types of attributes used in the current GIPSY implementation. Large amounts of data will be entered stored in POSTGRES.

GIPSY can be extended to take advantage of the new functionality. For example, the POSTGRES schema can be used to support ad hoc queries to find data of interest as it is entered in the database.

Polygon Overlaying

POSTGRES will support the traditional cartographic overlay operations such as intersection and union of polygons. It will also be extended to include libraries for map projection issues and conversion of GIS data formats. This will eliminate the need to spend hours or days converting data to formats which can be intercompared.

Once the polygons have been computed, they can be displayed in the TIOGA browser system being developed as part of Project Sequoia 2000. Simultaneous display of surface plots generated by GIPSY will allow the user to identify the documents which have the highest probability of referring to specific areas of interest.

It would be useful to compare GIPSY techniques to text indexes. The only related work in this area is a disseration which consists primarily of a benchmarking effort [Hill90]. The text corpus and the techniques used in Hill's study do not seem easily duplicable. Because GIPSY is not directly comparable to any existing system, benchmarking should be done after GIPSY is more mature. At that time, it is expected that GIPSY will have much higher precision and recall than keyword indexing techniques and manual subject indexing. GIPSY may also outperform many manual coordinate indexing systems which use bounding boxes and points, rather than polygons, for indexing.

Future work also involves the exploration of the potential of applying GIPSY techniques to several issues other than the assignation of polygonal indexes to text documents. For example, a desirable byproduct of the proposed design is that natural language user queries relating to all database types could be parsed in the same way as documents. This could be useful in cases in which the user does not know the exact spatial location of the geographic entity being referred to. In this instance, the user could enter a natural language description of the area in a dialog box. The area of interest could be calculated as above and then highlighted.

Another desirable side-effect is that the dictionary and thesaurus work in the geographic discriminator will facilitate the development of subject searches. Natural language subject searches can use the thesaurus terms. Graphical subject searches could consist of navigation of trees representing the hierarchical taxonomies of plants, animals, etc.

The model proposed above also extends nicely to indexing of dimensions other than geographic location. In fact, the more dimensions incorporated, the more accurate, albeit compute-intensive, the indexing. For example, if time were used to constrain a certain index, the query passed to the database function would be more specific, e.g. "Locate water bodies in which the saline level is high in the summer".

8. Concluding Comments

It has been shown that users have a need to access information according to geographic attributes of documents. It has also been demonstrated that users would like to retrieve information not only by explicit spatial locations, i.e. by place names, but also by more descriptive geographic characteristics. However, current systems do not adequately support these types of indexing. This inadequacy can be alleviated by a system designed as proposed in section 4. A prototype implementation has been presented and discussed in sections 5 and 6 to illustrate the viability of using spatial reasoning and a large geographic knowledge base to automatically assign polygonal indexes to text documents.

Appendix 1a: Agriculture, Economics

- Assess the likely short-run and long-run financial impacts of a range of alternative future water supply levels and quality standards on agriculture and related industries. This includes assessing the short-run responses of water districts and agriculturalists to recent water scarcities and identifying the likely long-run responses to continued constraints. The opportunities and constraints for adjustments to alternative supply and quality levels will be analyzed and the broader economic/financial impacts of adjustments on the agricultural sector and the State economy determined.
- (1) Evaluate the effects of newly developing technology on the State's agriculture, e.g. wind and solar energy and their economic feasibility in agriculture. (2) Evaluate fossil energy sources and future availability and future price. (3) Design a model that assesses the effect of varying supply and price of one farm commodity on related commodities.
- New [contract date: 1979] energy technologies are evolving, some of which have large water requirements for cooling or resource processing. It is imperative for water and energy planning at both the State and national level to assess the future water demands by energy technologies and possible impediments to energy development posed by limited water supplies.
- Perform a cross section survey of CA growers that will provide a statistically valid sample of regional crop production costs. The sample frame will be stratified by scale of farm production, crop, and county. [DWR makes direct use of farm production data in various analysis, including water allocation studies, project evaluations, and drainage considerations.]

Appendix 1b: Water Quality

- Geotechnical investigation and analysis of possible seepage from the proposed Los Banos Grandes Reservoir and possible impact of such seepage on existing shallow ground water lands in nearby farm areas. [Legislation authorizing Los Banos Grandes Reservoir directed DWR to consider possible impact of this project on high ground water and drainage problem areas on nearby irrigated farm lands.]
- Chemical analysis of soil samples for pesticides, pesticide degradation products, and petroleum hydrocarbons. [It is necessary to evaluate the possibility of soil contamination by toxic organic chemicals of land acquired for the development of the Kern Water Bank, a ground water storage project.]
- Review operation history at oil fields and related activities to identify possible sources of toxic organic chemical, heavy metal, and other toxic element contamination. The possible sources shall include, but not necessarily be limited to: (1) spills; (2) waste disposal practices; (3) improperly abandoned wells; (4) mud pits associated with drilling activities; (5) leaking storage facilities and pipelines; and (6) failure of production and injection wells allowing contamination of fresh water. [Above information to be obtained by interviews and review of the pertinent records, literature, and aerial photography.]
- Identify specific sites where significant toxic organic, heavy metal, and other toxic element contamination is likely to occur, and (1) prepare map showing location of all waste disposal areas; (2) prepare map showing location of all abandoned and/or destroyed wells; and (3) prepare map showing location of all present or former mud pits associated with drilling operations.
- Review agricultural land use history and historical pesticide usage in the study area.

Appendix 1c: Hydrology

- Provide computer operations to study the impact of deepening the Sacramento River Deep Water Ship Channel on State Water Project operations.
- Estimate the hydrologic conditions of the Sacramento-San Joaquin Delta that would occur in the absence of the Central Valley Project and the State Water Project.
- Implement a computer model which simulates the Central Valley Project and State Water Project system operation of reservoirs, river flows, and conveyance facilities.
- Develop an artistic representation of flow patterns in the Sacramento-San Joaquin Delta. Investigate and develop methods to visually describe: (1) various hydrologic conditions in the Delta including reverse flow conditions in the lower San Joaquin River due to State Water Project operations and (2) salinity intrusion and salinity gradients in the Delta which occur with varying flow patterns including reverse flow conditions in the lower San Joaquin River.
- Determine the hydrogeologic characteristics and properties of the Indian Wells Valley ground water system.
- Estimate the annual unimpaired flows during the past 400 years of the Santa Ynez River at Bradbury Dam, and make hydrological analyses of this time series. Such an estimate of unimpaired flows is to be made by reconstruction of flows prior to 1918 from tree ring data, and unimpaired flows of the Santa Ynez River after 1918 to be provided by the State.
- Investigate the use of weather modification techniques to augment snowpack and streamflow in the Sierra Nevada. The plan will be sufficiently specific to guide an actual seeding operation in the Feather River Basin without further elaboration. The plan will specify types of clouds to be seeded, seeding agents to be used and their rates of application, locations for ground-based generators, aircraft seeding tracks (if applicable), suspension criteria, and a proposed method of evaluation.

Appendix 1d: Environmental

- Review the literature on the environmental effects of shrubland management. [There are differences of opinion within the scientific community as to the effects of vegetation management on water and sediment yield, etc. A review of the literature is needed to evaluate the extent of existing knowledge and design a research program to fill in the gaps found to be essential to valid definition of the effects of shrubland management.]
- Provide an environmental evaluation of the effects of weather modification and vegetation management in the Santa Ynez Watershed above the Cachuma Reservoir. [These contracts relied on previous studies and the readily available literature, as well as on new studies, e.g. an analysis of existing vegetation composition in the watershed, and the development of vegetation composition scenarios with and without enhanced controlled burning for vegetation management.]
- Estimate the areal extent of each of seven management categories within three administrative areas in three watersheds. The seven management areas are: geologic hazard, riparian zone, reservoir buffer, chamise climax, north-facing slope ecotype, oak savanna, and other wildlands except barren.
- Prepare a report analyzing the degree of success of an experimental transplanting program for a state protected plant.
- Document and characterize long-term trends in the abundance and composition of key biological components of SF bay during the next 10-20 years. The study will use long term monitoring of selected biological components (plankton, benthic infauna, and macroalgea) of the Bay ecosystem to identify and evaluate environmental stress.
- Conduct a literature review of past relevant fish and wildlife studies in the Cottonwood Creek study area. Perform field surveys to determine bird and mammal populations. [DWR will use the results of the fish and wildlife studies to ascertain whether a proposed project is environmentally feasible.]
- Examine ways to mitigate the impacts of the Los Banos Grandes project on (1) the San Joaquin kit fox, Swainson's hawk, bald eagle, peregrine falcon, and greater sandhill crane, and (2) State/Federal candidate mammal and bird species. Also, address the relationship of the population of kit fox at Los Banos Grandes with the kit fox population as a whole.

Appendix 1e: Cultural

- Survey the literature and other information sources with respect to potential reservoir sites and submit a report setting forth the presently known aspects of cultural heritage which are significant in the reservoir areas. [Study includes researching available published and unpublished archaeological, historical, and ethnographic references dealing with the proposed reservoirs outlined above and the areas immediately surrounding them. Reports dealing with similar reservoir projects will also be used for comparative data and to outline the next phase of cultural resource investigation necessary in connection with advanced project planning.]
- Conduct an archaeological survey of the Los Banos Grandes project areas. In addition to providing a complete inventory of the project areas' cultural resources and evaluating their significance, the contractor will also provide a recommendation of measures to mitigate any impacts by the project on these resources.

Appendix 1f: Project

- Conduct a feasibility investigation on enlarging Shasta Lake for the purpose of increasing water and energy supplies.
- Apply value engineering in a broadly defined sense to review and evaluate project design proposals and cost estimates to determine whether it will be possible to develop a facility less expensive than the proposed Auburn Dam which will accomplish essentially the same purposes for water supply, power, flood control, and corollary recreation and instream benefits.
- The study will be a feasibility investigation of the engineering, economic, institutional, legal, and environmental considerations of using the Oroville watershed for precipitation enhancement programs. A reconnaissance-level study was judged unnecessary due to the information available from other studies along the Sierra Nevada range. Work includes review of precipitation enhancement projects conducted by PG&E, the Bureau of Reclamation, and others.
- Determine, based on geological engineering, economic, institutional, legal, and environment considerations, the feasibility of using the Chino Ground Water Basin for the underground storage of water.
- Provide a findings report which investigates the physical, economic, and environmental effects of a controllable, submerged, mechanical barrier in the vicinity of Carquinez Strait and develop operational criteria for a model of the barrier to be used in conjunction with the Corps of Engineers' Bay-Delta model. [A number of interests within the water community felt it was appropriate at this time to review past studies and identify environmental and technical concerns on this issue.]
- Provide consulting services for the development of conceptual designs and a model testing program for large water control otructures to be located in the Sacramento-San Joaquin Delta. [Study includes detailed review of engineering literature to locate texts and journal articles describing similar types of structures. Computerized searches will be made at University of California libraries, through the Archives of the Water Resources Center at Berkeley, and other libraries where appropriate.]
- Determine implementation procedures, necessary improvements, project costs, revenues, funding, and a benefit/cost ratio for the wildlife management plan which is being considered as an alternative to an overland water facility for Sherman Island. (1) implementation and project management procedures for a wildlife management plan (e.g. Define who shall possess the legal ownership of land titles.) (2) description of needed and existing improvements (e.g. island levees, roads, pumps) (3) consumptive water use differences between marsh and riparian habitat, and historic crop patterns (4) water quality and quantity needed to maintain wildlife habitat (5) description of water rights (6) water and marsh management strategies (7) description of existing wildlife populations and Department of Fish and Game future objectives (8) waterfowl disease potential and probable control measures (9) mosquito abatement considerations (10) inventory of existing vegetation (11) soils map (12) history of past land use (13) topography (estimate of expected survey costs) (14) potential archaeological considerations (15) estimated capital costs (16) predicted funding needs (17) benefit/cost ratio.

Acknowledgements

Thanks are owed to Michael Berger for providing the transcripts from Melvyl. Thanks are also due to Ron Maeder for work on polygonal functions, Chris Plaunt for work on the parser and thesaurus, and Matt Hamre of the California Department of Water Resources for assistance with operations on the thesaurus. Thanks also to Gary Darling, Ronald Braun, and Marti Hearst for helpful discussions and comments on earlier drafts of this paper. Finally, a debt of gratitude is due to my thesis committee, Matt Farrens, Dick Walters, and Mike Stonebraker, for many helpful suggestions and comments.

References

[AHRW76] J. R. Anderson, E. E. Hardy, J. T. Roach and R. E. Witmer, "A Land Use and Land Cover Classification System for Use with Remote Sensor Data", *Geological Survey Professional Paper,* vol. 964 (1976).

[Aven85] B. Aveney, "Facing Our Technology", *Journal of Library Automation,* vol. 8 (June 1985), pp. 85-86.

[Batt89] D. Batty, "Thesaurus Construction and Maintenance: a survival kit", *Database,* vol. 12, no. 1 (February 1989), pp. 13-20.

[Berg92] M. Berger, *Personal communication*(1992).

[Brin62] B. Brinker, "The Geographical Approach to Materials in the Library of Congress Subject Headings", *Library Resources and Technical Services,* vol. 6, no. 1 (1962).

[Byro87] J. Byron, "Topographical indexing", *The Indexer,* vol. 15, no. 4 (October 1987), pp. 211-214.

[ChLS92] J. Chen, R. Larson and M. Stonebraker, "The Sequoia 2000 Object Browser", *Proceedings of Compcon*(February 1992).

[FaLe82] R. K. Farrar and J. V. Lerud, "Online Searching Using Geographic Coordinates", *Second International Conference on Geological Information*(1982), pp. 109-116.

[Furn91] G. W. Furnas, "New Graphic Reasoning Models for Understanding Graphical Interfaces", *Human Factors in Computing Systems: Reaching Through Technology CHI '91 Conference Proceedings*, New Orleans, Louisiana (April-May 1991), pp. 71-78.

[Gard93] K. Gardels, *Interview in Sequoia 2000 Video*, California Department of Water Resources (1993).

[Grif89] A. Griffiths, “SAGIS: a proposal for a Sardinian geographical information system and an assessment of alternative implementation strategies.”, *Journal of Information Science,* vol. 15 (1989), pp. 261-267.

[Hear92] M. Hearst, *Personal communication*(1992).

[Hill90] L. L. Hill, “Access to Geographic Concepts in Online Bibliographic Files: Effectiveness of Current Practices and the Potential of a Graphic Interface”, *Dissertation*, University of Pittsburg (1990).

[Holm90] D. O. Holmes, “Computers and Geographic Information Access”, *Meridian,* vol. 4 (1990), pp. 37-49.

[Korf86] R. R. Korfhage, “Browser: A Concept for Visual Navigation of a Database”, *Proceedings of the IEEE Computer Society Workshop on Visual Languages*, Dallas, Texas (June 1986), pp. 143-148.

[Korf91] R. R. Korfhage, “To See or Not to See: Is That the Query?”, *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, Chicago, Illinois (1991), pp. 134-141.

[Lars91] R. Larson, “Classification Clustering, Probabilistic Information Retrieval, and the Online Catalog”, *Library Quarterly,* vol. 61, no. 2 (April 1991), pp. 133-173.

[LGPP90] D. B. Lenat, R. V. Guha, K. Pittman, D. Pratt and M. Shepard, “Cyc: Toward Programs with Common Sense”, *Communications of the ACM*(August 1990), pp. 30-49.

[McTT88] C. A. McCann, M. M. Taylor and M. I. Tuori, “ISIS: the interactive spatial information system.”, *International Journal of Man-machine Studies,* vol. 28 (1988), pp. 101-138.

[Misc82] W. Mischo, “Library of Congress Subject Headings: A Review of the Problems, and Prospects for Improved Subject Access”, *Cataloging and Classification Quarterly,* vol. 1, no. 2/3 (1982), pp. 105-124.

[Morr88] B. Morris, “CARTO-NET: Graphic retrieval and Management in an Automated Map Library”, *Special Libraries Association, Geography and Map Division, Bulletin,* vol. 152 (June 1988), pp. 19-35.

[MuEa87] J. G. Mulvihill and J. T. Eaglesfield, “Improving Subject Access in Geoscience Library Catalogs”, *Proceedings of the Twentieth Meeting of the Geoscience Information Society,* vol. 16 (1987), pp. 141-157.

[RLG89] RLG, “RLG Enters New Sphere with Geoinformation Project”, *The Research Libraries Group News,* vol. 19 (Spring 1989).

[StKe91] M. Stonebraker and G. Kemnitz, “The POSTGRES Next Generation Database Management System”, *Communications of the ACM*(October 1991), pp. 78-92.

[Tahi89] S. N. Tahirkheli, “Thesaurus Problems and Solutions: The Language of Geology Develops Steadily”, *Proceedings of the Twenty-Second Meeting of the Geoscience Information Society,* vol. 18 (1989), pp. 89-94.

[USGS87] USGS, “Geographic Names Information System”, *Data Users Guide,* vol. 6 (1987).

[Well78] H. H. Wellisch, “Poland is Not Yet Defeated, or: Should Catalogers Rewrite History? With a Discourse on When is an Island not an Island?”, *Library Resources and Technical Services,* vol. 22 (Spring 1978), pp. 158-167.