GIPSY:
Georeferenced Information Processing SYstem

Allison Gyle Woodruff
California Department of Water Resources
and
Computer Science Division
University of California, Berkeley
woodruff@cs.berkeley.edu

Christian Plaunt*
Library and Information Studies
University of California, Berkeley
chris@bliss.berkeley.edu

March 25, 1994

Abstract

In this paper we present an algorithm which automatically extracts geopositional coordinate index terms from text to support georeferenced document indexing and retrieval. Under this algorithm, words and phrases containing geographic place names or characteristics are extracted from a text document and used as input to database functions which use spatial reasoning to approximate statistically the geoposition being referenced in the text. We conclude with a discussion of preliminary results and future work.

*This research was sponsored by the University of California and Digital Equipment Corporation under Research Grant #1243. and a consortium of government and industrial partners.

Contents

List of Figures

List of Tables

1 Introduction

Georeferenced data is commonly characterized as having physical dimensions and a spatial location (Aronoff 1989), and a georeferenced document is indexed according to such data. Historically, documents have been indexed primarily by subject, author, title, and, to a lesser extent, by document type. However, a diverse and large group of information system users desire geographically-oriented access to document collections. These include:

- *natural resources managers* who would like to retrieve information pertinent to specific areas. For example, the preparation of an Environmental Impact Report (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.
- *tourists* who would like to locate brochures, guide books, and historical information germane to areas being considered for trips.

In searches such as those above, "relevant" documents may cross many subject areas and be of many different types (e.g. journals, newspapers, magazines, maps, computer files, etc.). This often makes a comprehensive search difficult, as Byron (1987) has observed. Geographic location is the shared attribute which can draw together these disparate subjects and document types to allow a cohesive search (Gardels 1993). Authors such as Byron (1987), Hill (1990), and RLG (1989) have therefore recognized the need for research, development, and implementation of geographic indexing.

The goal of the research reported here is to explore the utility and viability of indexing full-text documents using coordinates for spatial retrieval and display. Section 2 presents previous work in this area. Section 3 discusses design considerations and goals. Section 4 discusses the general approach to automated georeferencing favored by the authors. Section 5 presents a prototype system called GIPSY, an acronym for **G**eoreferenced **I**nformation **P**rocessing **SY**stem. GIPSY assigns to each text document zero or more index entries consisting of one or more polygons representing geographic locations. Section 6 details results. Section 7 discusses promising areas for further work, and Section 8 concludes our findings.

2 Previous Work in Georeferencing of Text Documents

Geographic indexes can be based on text terms or on coordinates and can be assigned by an indexer or derived from a text manually or automatically. In this section, indexing techniques and their application to georeferencing text documents will be discussed.

2.1 Text Indexing

Most operational and commercial text-based information systems have historically used manually assigned subject terms, author names and title words for indexing and retrieval, although some automatic keyword indexing techniques are beginning to make inroads in the commercial sector (Pritchard-Schoch 1993). This section details some of the previous work in these areas and their deficiencies in georeferencing text documents.

2.1.1 Library of Congress Subject Headings

Although the Library of Congress did not originally intend to develop a standardized subject protocol (Mischo 1982), the procedures it uses have had widespread impact on an international scale (Wellisch 1978). Therefore, perhaps the most significant example of manually-generated geographic subject description in library cataloging comes from the Library of Congress.

There are three types of geographic subject headings in Library of Congress Subject Headings (LCSH). The first consists of a topical subject heading followed by a geographic subdivision (e.g. ART – PARIS). The second consists of the 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) (Brinker 1962). These have several well-documented problems.

Inconsistent Subheadings

A text document is most often assigned only one of these three types of index terms. 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 library catalog user to locate reliably items of interest (Brinker 1962).

Scatter

Subdivision can also fragment references. Mischo (1982) 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." (p. 110)

Updating

Several authors have documented problems with terminology in LCSH. Many of the terms used are not kept up-to-date, as observed by Mischo (1982). Further, the meaning of geographic terms and place names are dependent upon the time period to which they refer. Wellisch (1978) discusses some of the problems which arise from attempts to use a single geographic term to refer to an area.

Subdivision Scoping

Problems also result from the hierarchical characteristics of geographic terminology. For example, in LCSH, the procedure for deciding whether islands should appear as high level entities, or as subdivisions, is apparently arbitrary (Wellisch 1978, pp. 164–165).

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

2.1.2 Keyword Indexing

Because of the volume of full-text documents available online and the cost of human indexing, automatic indexing is an increasingly desirable strategy (Salton 1989). Currently, automatic indexing is limited to the use of keywords and phrases extracted from the text. Retrieval of text documents is performed 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. In this model, georeferencing, if present at all, is merely a side-effect of the existing paradigm.

2.1.3 Deficiencies of Text-based Georeferencing

In addition to the issues detailed above, the use of textual terms to describe geographic location has many inherent difficulties. Even in cases in which a document is meticulously manually indexed, geographic index terms consisting of text strings have several well-documented problems, including:

- *lack of uniqueness* (Griffiths 1989; Holmes 1990)

  Confusion can result from a reference to Cambridge, which might refer to either Cambridge, England or Cambridge, Massachusetts (to mention only two possibilities).

- *spatial boundary changes* (Griffiths 1989)

  Political boundaries change over time, leading to confusion about the precise area being referred to.

- *name changes* (Farrar & Lerud 1982; Griffiths 1989)

  Geographic names change over time. This makes it difficult for a user to retrieve all information which has been written about an area during any lengthy time period.

- *spatial and naming variation* (Farrar & Lerud 1982)

  Terms vary not only over time, but in contemporary usage as well.

- *spelling variation* (Farrar & Lerud 1982; Griffiths 1989)

  Geographic names vary in spelling, not only over time, but by language as well.

- *neologisms* (Farrar & Lerud 1982)

  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.

2.2 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 (Tahirkheli 1989). These systems support range searching in which the user generally specifies 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 (RLG 1989; Holmes 1990). However, none of these systems has included automatic assignment of coordinates. The popularity of these systems has been limited by the lack of automatic indexing and by their primitive interface mechanisms.

3 Design Goals and Considerations

This section discusses the fundamental objectives of GIPSY. It additionally details technological motivations which make GIPSY's design viable. Finally, it presents GIPSY's georeferencing capabilities in the context of a complete information retrieval system using the TIOGA browser (Chen *et al.* 1992).

3.1 Primary Goals

Our system is motivated by the recognition of three separate but complementary goals. First, georeferencing should be based on a coordinate indexing system. Second, georeferencing should support a spatial display mechanism. Third, georeferencing should be automated.

3.1.1 Coordinate Indexing

An information retrieval system must share a common language with the user (Batty 1989). The lack of communication resulting from problems with text indexing has consequently made indexing and retrieval systems ineffective. Inconsistency and ambiguity in geographic headings in the Library of Congress Subject Headings (LCSH) list, for example, has been discussed at length in (Brinker 1962; Mischo 1982; Wellisch 1978). Further, an empirical investigation comparing the use of text representations and coordinate systems to represent geographic areas has shown the inadequacy of text terms for indexing and retrieval (Hill 1990). Because coordinates, unlike text, are unambiguous (Holmes 1990), they can be used to communicate uniquely locations to the information seeker. Further, many users require proximity queries which cannot be supported by textual terms. Coordinates, particularly those representing complex polygonal entities rather than points or boxes, can be used to support these nearness queries. Therefore, we argue that a georeferencing facility should make use of coordinate indexing.

3.1.2 Spatial Interface

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

A map-based graphical interface 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 (Furnas 1991). Because many 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 (Morris 1990).

Second, a graphical interface, such as a map, allows for a dense presentation of information (McCann *et al.* 1988). This addresses a problem which arises in the standard information retrieval system model. In this model, the user issues a query to the system, the system partitions the dataset into a "retrieved" set and a "not-retrieved" set, based on its judgment of the relevance of each document to the query, and presents to the user *only* the "retrieved" set. This partitioning (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 (1991) 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 (Korfhage 1986).

Third, in standard boolean information retrieval systems, the display is not ranked by (probable) relevance or utility; 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 at least one axis of relevance through visual presentation of proximity.

Therefore, we argue that the indexes provided by a georeferencing system should be usable in a graphical retrieval system.

3.1.3 Automated 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. The demands on resources are compounded by a desire, resulting in part from the hypertext paradigm, to index subcomponents of text using its structure (Hearst & Plaunt 1993) or its orthographic boundaries (e.g. paragraphs, sections). Second, it has long been recognized that manual indexing is inevitably inconsistent. Third, many other data types, such as satellite images and aerial photographs, exist primarily within a geographic space, and so naturally should be assigned geographic index terms. In the future, these data types may automatically be assigned geographic index terms through the use of global positioning systems (GPS) technologies (Holmes 1990). Therefore, automatic indexing of text documents in terms of geographic location will be necessary to integrate text with rapidly increasing amounts of georeferenced data of these other types.

3.2 Design Considerations

Clearly, widespread 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 (Brinker 1962); the computational complexity of automatic indexing techniques was limited by cpu speed; the lack of authority control in automatic systems created consistency and update problems; and, since documents were not online, they could not easily be indexed through electronic means (Tenopir & Ro 1990).

These factors, and many others, are changing rapidly. Online indexing and access to documents is increasingly in demand as computers become more ubiquitous and as multi-media applications become more common. Further, 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 are being developed in the areas of tertiary storage, tertiary file systems, networking, scientific visualization, database management, and computer-facilitated collaboration (Chen *et al.* 1992). The latter two issues pertain directly to information retrieval.

To address the needs of Earth system scientists, a new browser paradigm has been proposed (Chen *et al.* 1992). In this system, called TIOGA, information will be displayed topologically according to continuous characteristics which are attributes of the data. For example, documents may 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 Project Sequoia 2000 browser is being designed not only to work with textual 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 data types, many of which will be entered into the database with explicit spatial (coordinate) location information. However, text documents will not have explicit spatial indexes; geographic index terms will need to be generated. Because of the large number of documents which will be stored in the text component of the Project Sequoia 2000 database, manual indexing of these documents will not be feasible. Therefore, a system must be designed which will automatically assign geographic location information to a text document based on its content.

In addition, Project Sequoia 2000 will incorporate at least two significant advantages over previous systems in its text processing and retrieval components. The first is the large amount of data supplied by the Earth system scientists. There will be terabytes 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 inferential reasoning (Lenat *et al.* 1990), previous natural language processing systems have not utilized such a volume of information. We believe the large scale of this raw data will make possible the kind of assigned inferential indexing GIPSY is trying to provide.

The second advantage provided by Project Sequoia 2000 will be the unique functionality offered by the easily extensible POSTGRES database management system (DBMS) (Stonebraker & Kemnitz 1991). POSTGRES will not only allow integrated access to any number of data types, but will also contain geographic information system (GIS) functionality, and hence, be able to support spatially-related queries.

4 Solution

This section describes the fundamental structure of GIPSY. A prototype which implements a subset of the features described in this section is presented in the Section 5.

GIPSY's basic strategy is to extract place names as well as more general geographic indicators from documents, and use the intersection of these referents 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 which are near both the habitat of eagles and an urban area. This method consists of three primary steps:

4.1 Step One: Parsing

The first step is to extract relevant key words and phrases. As an example of the desired output, consider a document which discusses below ground water storage in the San Joaquin Valley. This water enters the water table through the intentional flooding of a specific region, the Kern Fan Element, in Southern California. A document about the Kern Fan Element might contain the following content-bearing geographic words and phrases:

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 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

A parser for this strategy must "understand" how to identify geographic terminology of two types:

- terms which match objects or attributes in the data set. This step requires a large thesaurus, partly hand built and partly automatically generated.
- lexical constructs which contain spatial information, e.g. "contiguous", "adjacent", "south of", "between", etc. To implement this, a list of the most commonly occurring constructs must be created.

Once the geographic components of phrases have been identified, a certain amount of syntactic work will be required to omit the irrelevant 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.

4.2 Step Two: Locating pertinent data

The output of the parser is passed to a function which retrieves 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 information such as name, size, and location of cities or states; 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.

In this model, extracted keywords and phrases must be mapped into the terminology used to describe several types of data. A hierarchical taxonomy can not adequately represent geographic terms (Tahirkheli 1989). Consequently, our approach uses 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 cattails); and (4) quantification of terms (e.g. "brackish" is a degree of salinity). 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 occurring in phrases such as "adjacent to California" or "within a one hundred mile radius of Los Angeles".

One concern is whether sufficient data will exist for this system; areas which have very little data associated with them in the database might be discriminated against by the proposed indexing system. For example, if there were a location in the real world which matched a query exactly, but for which no data was stored in the system, it would not be found by the spatial location approximator. However, the system may tend to self-adjust: there will be a correlation between the areas about which documents are written and the areas for which data is available.

4.3 Step Three: Performing the polygonal overlay

The final step is to perform a polygonal overlay of the data output by Step Two. The polygons and their associated weights are overlayed to identify areas which have a high probability of relevance based on their weights. 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 phrases and the real-world polygons they indicate. For the current study, California was chosen as the test area. The thesaurus contains references to three kinds of data about California.

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[1]
private	range	rapids	reserve	reservoir
ridge	school	sea	siding	site
slope	spring	stream	summit	swamp
toll	tower	trail	tunnel	valley
well	woods			

Table 1: GNIS Feature Types

5.1 Place Names

The first data set used is from the US Geological Survey's Geographic Names Information System (GNIS) (USGS 1985). 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
```

5.2 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 1.

5.3 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 (Anderson *et al.* 1976). This GIRAS data set contains over 60,000 entities in the Universal Transverse Mercator projection. These entities were processed to be explicitly represented

[1] Populated place.

as polygons in the Lambert-Azimuthal projection. The data set contains polygons associated with each of the land use types listed in Table 2. The majority of these polygons are represented by hundreds, and in some cases thousands, of points.

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 were not assigned synonyms or kind-of relationships, but were assigned 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, a list of "stop words", such as "of" and "the" (e.g. from Arch of the Navarro), were deleted from the 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).[2]

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 (Miller *et al.* 1990).

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.

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

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 Agricultural 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

Table 2: GIRAS Land Use Classifications

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>   := file name
<polygon>     := list of vertices
```

<char_str> is the natural language term which the parser searches for in documents, <level> indicates the "distance" the entry is removed from the original GIRAS or GNIS terms, <attr> is the "attribute"[3] of the term, source is the dataset (GNIS or GIRAS) or main thesaurus entry from which the current entry was derived, <coord> is either a GNIS point coordinate[4] or the name of a file containing the GIRAS polygon data, and <feature> is one of the authorized GNIS/GIRAS types.

For example, the following is a direct entry for the term "beach", which is a synonym[5] for the authorized GIRAS feature type "beaches" whose coordinates are collected in a file named "giras72":

```
beach:1:=:beaches:giras72
```

In contrast, indirect matches are removed from their sources by a greater distance:

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

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

Operations performed on the thesaurus implement the steps outlined in Section 4.

5.4 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*. The parser reads through a pre-determined segment of text one sentence at a time, turning it into a stream of tokens. For each non-stop word token in the stream, if that token begins one or more entries in the thesaurus, the longest exact matching entry is retrieved, and the stream pointer is advanced past the match. If the token is not found in the thesaurus, it is depluralized and rechecked in the thesaurus, e.g. *valleys* would fail to match, and would be transformed to valley, which would match tokens in the thesaurus.

[3] The "attributes" are defined in Table 1 on page 9.

[4] Points are converted to polygons later in the process.

[5] Synonyms are at a distance of 1 from their source.

5.5 Step Two: Locating Pertinent Data

The parser returns a list of matched tokens and entries, which are passed on for further processing.

First, 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).

Second, 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.

Third, the degree of correlation between a polygon and the document from which the natural language terms were derived is dependent on a number of factors, and the best measure of correlation remains an ongoing research question. Currently, the formula used to calculate the expected relevance of the polygon is to take the reciprocal of the following term: 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[6].

5.6 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 "weight" 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 1). 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 2). 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 3). 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 4). To minimize fragmentation in this case, polygons are sorted by size prior to being positioned in the skyline.

6 Results

Experimentation with GIPSY has begun, varying parameters to test both effectiveness and efficiency, including tuning various weighting factors, threshholds, and the size of the text segment being parsed. Experiments have also been run with and without the GNIS and the GIRAS data to determine their usefulness.

[6] 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." To mitigate this effect, tokens are checked against the standard dictionary, and if they are present, their weights are decreased.

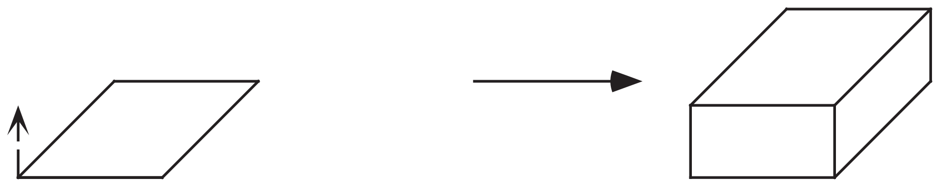

Figure 1: The "weight" of a polygon, indicated by the vertical arrow, is interpreted as "thickness".

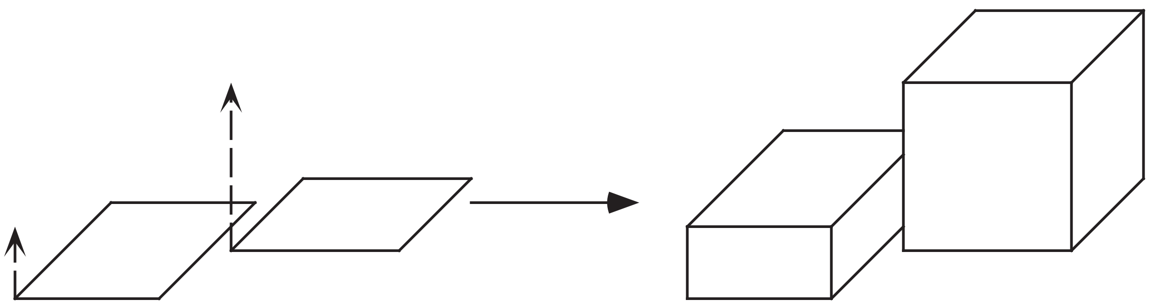

Figure 2: Two adjacent polygons do not affect each other, each is merely assigned its appropriate "thickness".

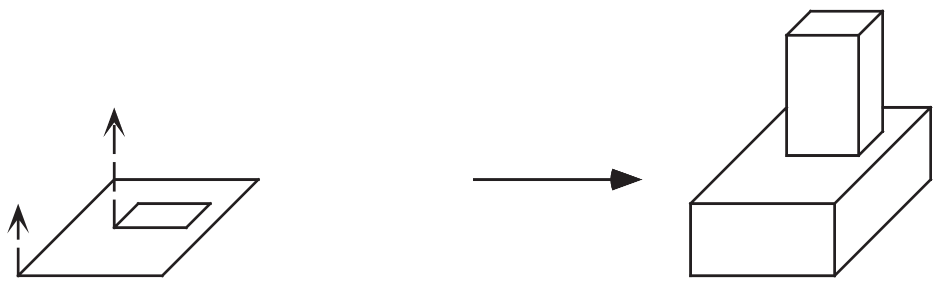

Figure 3: When one polygon subsumes another, their "thicknesses" in the area of overlap are summed.

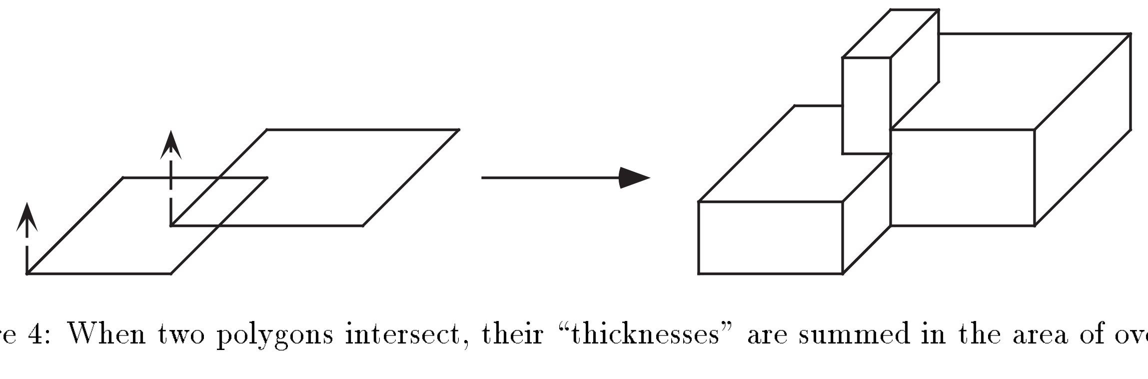

Figure 4: When two polygons intersect, their "thicknesses" are summed in the area of overlap.

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

Not surprisingly, there are many examples in which exact place name hits were correct indicators of the location being discussed in a paragraph. For example, GIPSY has processed the following text which centers on Santa Barbara County:

> 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 down-sized 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.

The resulting surface plot appears in Figure 5. The figure contains a gridded representation of the state of California, in which California is elevated to distinguish it from the base of the grid. The northern part of the state is on the lefthand side of the image. The towers stacked on California represent polygons in the skyline generated by GIPSY's interpretation of the text. The largest towers occur in the area referred to by the text.

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.[7]

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

[7] 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.

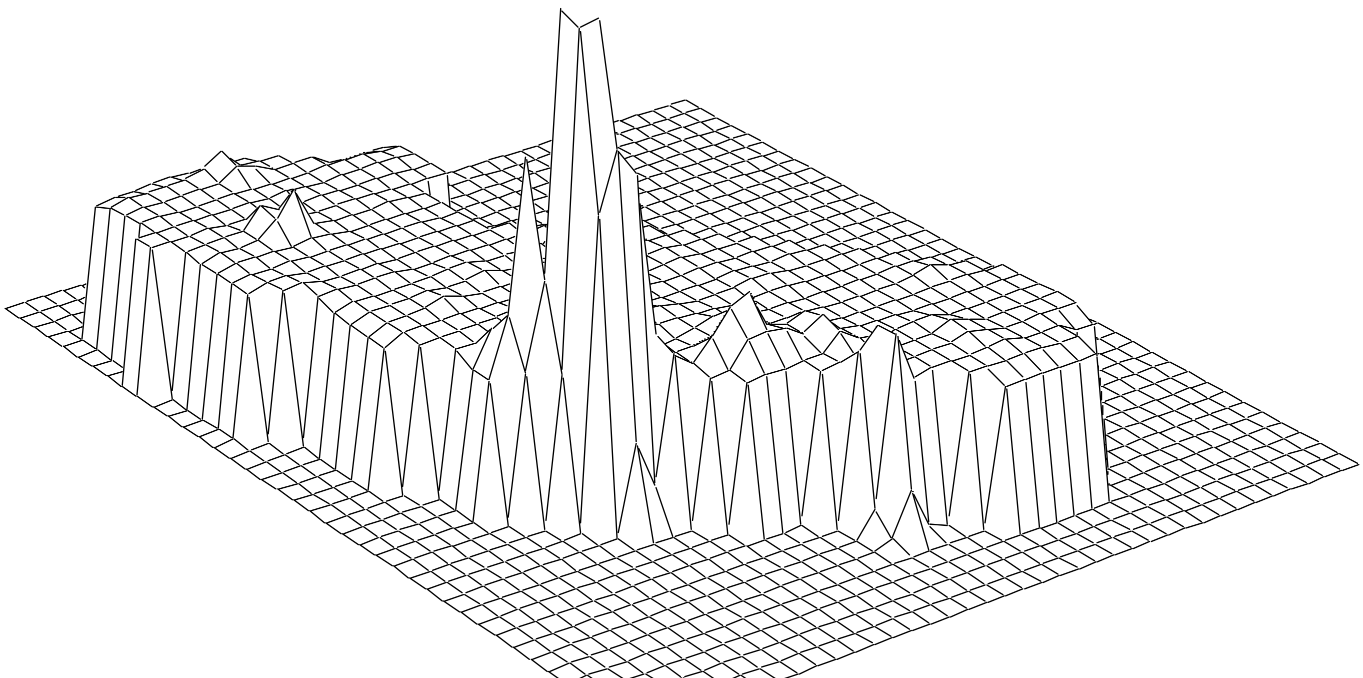

Figure 5: Surface plot produced from the State Water Project text which talks about Santa Barbara County, San Luis Obispo, and the Santa Ynez Valley area at some length.

> 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 6. Note the density of reference in the north part of the state (at the top of the diagram), 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).

Weighting is crucial

Varying virtually any parameter had significant impact on results. There are many complex trade-offs which should be further explored.

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 advanced 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*. Further, all parsing was case-insensitive, which had several benefits, but also created noise.

Polygon manipulation is resource-intensive

Programming GIPSY to stack polygons consumed a significant amount of time, and the resulting code is by far the most CPU-intensive part of GIPSY. In retrospect, we believe that prototyping should have been done by mapping the polygons onto a regular grid and manipulating the associated grid cell numbers.

Data quality is a limiting factor

Because the GNIS data was point data, it was necessary for us to devise an arbitrary method expand the points into polygons. More explicit polygonal data for the place names would improve the specificity of the technique.

The surface plots can be processed for use in a retrieval system

The surface plots generated can be processed and used for browsing and retrieval. For example, the two-dimensional base of a polygon with a thickness above a certain threshhold can be assigned as a coordinate index to a document. These two-dimensional polygons can then be displayed as icons on a map.

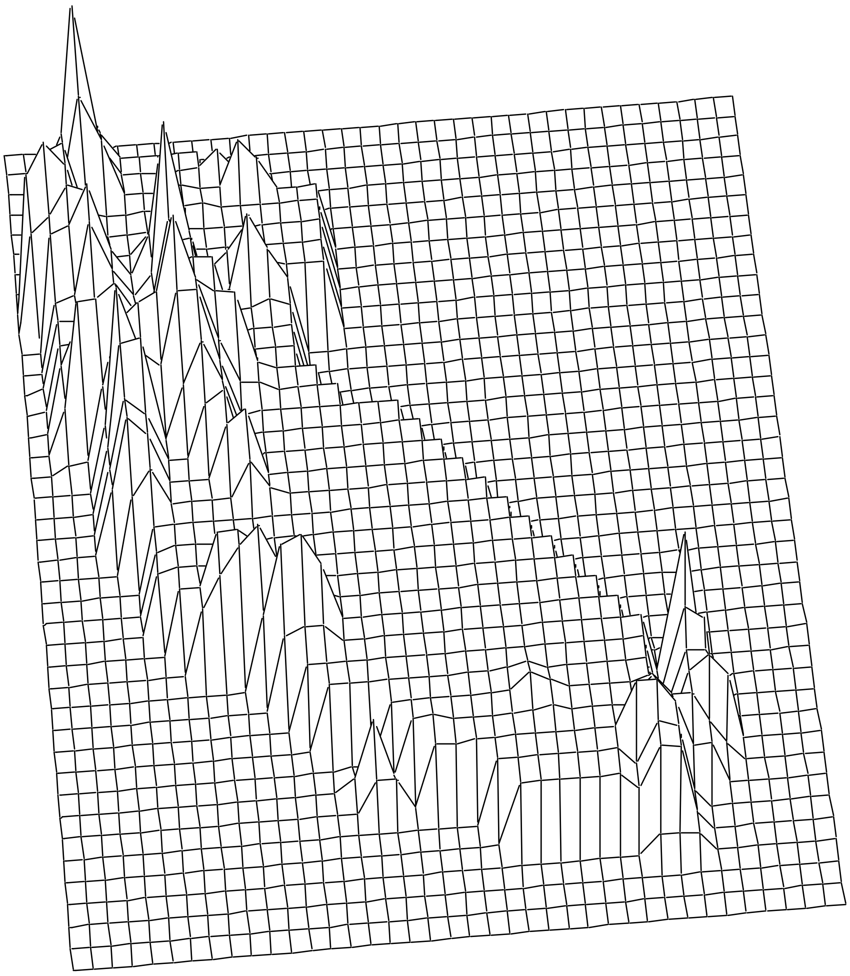

Figure 6: The surface plot generated by GIPSY for the "proposed extraction" text *without* normalizing for the number of polygons found. Note the density of reference in the north part of the state (at the top of the diagram), in which there are many canals, and the low reference to the south-eastern part of the state (desert area, in which there are few canals, other than in the Colorado River area).

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

Some of the natural extensions are discussed below.

7 Future Work

Promising research includes several extensions to the existing GIPSY implementation. First, it would be tremendously useful to develop a benchmark to compare GIPSY techniques to more traditional text-based indexing. A benchmark could also be used to tune system parameters. The most closely related work in this area is a dissertation comparing text-based and spatially-based indexing and retrieval (Hill 1990). The text corpus and the techniques used in this study do not seem easily duplicable. However, although benchmarking is a daunting task, evaluation is extremely significant. Consequently, future work should include the development of a benchmark. The most meaningful evaluation would probably be a comparison of manually-assigned coordinate indexes of documents to those generated by GIPSY.

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. Adjacency indicators such as "South of" or "between" should 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.

Non-text data in documents could also be used for indexing. For example, if a document contained a map of its study area, this map could be matched against a large knowledge base to approximate its coordinates.

Future work also involves the exploration of the potential of applying GIPSY techniques to several issues other than the assignation of polygonal index terms 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 referenced. 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.

Finally, the polygonal indexes can be used in a retrieval system. One straightforward method is to assign the polygons with extremely high relevance values as indexes of the document, and display a collection of these polygons from multiple documents as icons on a map. Work is currently in progress to use the TIOGA browser in this manner.

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 above. A prototype implementation has been presented and discussed to illustrate the viability of using spatial reasoning and a large geographic knowledge base to automatically assign polygonal index terms to text documents.

References

Anderson, J. R., E. E. Hardy, J. T. Roach, & R. E. Witmer (1976). A land use and land cover classification system for use with remote sensor data. *Geological Survey Professional Paper.*

Aronoff, S. (1989). *Geographic Information Systems: A Management Perspective.* WDL Publication, Ottawa.

Batty, D. (1989). Thesaurus construction and maintenance: A survival kit. *Database*, 12:13–20.

Brinker, B. (1962). Geographic approach to materials in the Library of Congress subject headings. *Library Research and Technical Services*, 6(1):49–64.

Byron, J. (1987). Topographical indexing. *The Indexer*, 15(4):211–214.

Chen, J., R. R. Larson, & M. Stonebraker (1992). Sequoia 2000 object browser. In *Digest of Papers. COMPCON Spring 1992. Thirty-Seventh IEEE Computer Society International Conference, San Francisco, 24–28 Feb. 1992*, pages 389–394, Los Alamitos, CA. Comput. Soc. Press.

Farrar, R. K. & J. V. Lerud (1982). Online searching using geographic coordinates. In *Second International Conference on Geological Information*, pages 109–116.

Furnas, G. W. (1991). New graphic reasoning models for understanding graphical interfaces. In *Human Factors in Computing Systems: Reaching Through Technology, CHI '91 Conference Proceedings, New Orleans, April–May 1991*, pages 71–78.

Gardels, K. (1993). Interview in video *Sequoia 2000: Computer Science Applied to Global Change Research.* California Department of Water Resources, Sacramento.

Griffiths, A. (1989). SAGIS: A proposal for a Sardinian geographical information system and an assessment of alternative implementation strategies. *Journal of Information Science*, 15:261–267.

Hearst, M. A. & C. Plaunt (1993). Subtopic structuring for full-length document access. In *16th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, June 1993*, pages 59–68, New York. Association for Computing Machinery.

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

Holmes, D. O. (1990). Computers and geographic information access. *Meridian*, 4:37–49.

Korfhage, R. R. (1986). Browser: A concept for visual navigation of a database. In *Proceedings of the IEEE Computer Society Workshop on Visual Languages, Dallas, June 1986*, pages 143–148, Washington, DC. IEEE Comput. Soc. Press.

Lenat, D. B., R. V. Guha, K. Pittman, D. Pratt, & M. Shepard (1990). Cyc: Towards programs with common sense. *Communications of the ACM*, 33(8):30–49.

McCann, C. A., M. M. Taylor, & M. I. Tuori (1988). ISIS: The interactive spatial information system. *International Journal of Man-Machine Studies*, 28:101–138.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, & K. Miller (1990). Five papers on WordNet. *CSL Report 43*. Cognitive Science Laboratory, Princeton University.

Mischo, W. (1982). Library of Congress subject headings: A review of the problems, and prospects for improved subject access. *Cataloging and Classification Quarterly*, 1(2/3):105–124.

Morris, B. (1988). CARTO-NET: Graphic retrieval and management in an automated map library. *Special Libraries Association, Geography and Map Division Bulletin*, 152:19–35.

Pritchard-Schoch, T. (1993). Natural language comes of age (Westlaw's WIN). *Online*, 17(3):33–8, 40, 42–3.

RLG (1989). Research Libraries Group enters new sphere with georeferencing project. *Research Libraries Group News*, 19.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA.

Stonebraker, M. & G. Kemnitz (1991). The POSTGRES next generation database management system. *Communications of the ACM*, 34(10):78–92.

Tahirkheli, S. N. (1989). Thesaurus problems and solutions: The language of geology develops steadily. In *Proceedings of the Twenty-Second Meeting of the Geoscience Information Society*, pages 89–94.

Tenopir, C. & J. S. Ro (1990). *Full Text Databases*. New Directions in Information Management, Number 21. Greenwood Press, New York.

United States Department of the Interior, U.S. Geological Survey (1985). Geographic Names Information System: Data Users Guide 6. Reston, Va.: The Survey, 1985.

Wellisch, H. H. (1978). Poland is not yet defeated, or: Should catalogers rewrite history. *Library Resources and Technical Services*, 22:158–167.