In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health (HealthGIS'12)*, Redondo Beach, CA, USA, November 2012.

The Picture of Health: Map-Based, Collaborative Spatio-Temporal Disease Tracking*

Rongjian Lan Michael D. Lieberman Hanan Samet
Center for Automation Research, Institute for Advanced Computer Studies,
Department of Computer Science, University of Maryland
College Park, MD 20742
{rjlan, codepoet, hjs}@cs.umd.edu

ABSTRACT

Disease outbreaks are intimately tied to geographic locations and to times, and as a result, health-related GIS along with open, Web-based data sources are increasingly crucial for public health. One such data source, ProMED-mail, offers disease reports distributed as email postings, along with locations and times of relevance. Locations are specified in text rather than in geometry, which necessitates a method for mapping textual locations to their spatial representations, called geotagging. To address this need, the previously-developed STEWARD system is leveraged for disease detection and tracking by geotagging ProMED-mail postings. While STEWARD was previously used in a disease tracking role, improvements to STEWARD are described including an innovative time slider that allows powerful and intuitive spatio-textual querying. Many additional future improvements for STEWARD and related systems are also discussed.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Design, Performance

Keywords

Disease tracking, geotagging, GIS, spatio-temporal

1. INTRODUCTION

Disease detection and tracking plays an important role in today's globally connected society. Organizations such as the US Centers for Disease Control and Prevention and World Health Organization are intensely interested in monitoring infectious diseases around the world, especially since the Internet is increasingly accessible everywhere. In addition, the increasing prevalence of volunteered geographic information (VGI) [14] in media such as blogs and tweets can likewise be leveraged to effect greater tracking ability. Importantly, disease outbreaks have a strong geographic component, in that their origins and spread are intimately related to environmental conditions and movement patterns, which enhances the need for robust health-oriented geographic information systems (GIS). A number of Web-based services for disease tracking have likewise flourished, most notably *ProMED-mail* [16], henceforth referred to as *ProMED*, an online alert system intended to quickly disseminate news of the latest outbreaks to medical professionals and laymen around the world.

Our focus in this paper is to enable the simple and intuitive spatio-temporal querying and retrieval of relevant ProMED postings—that is, retrieval of postings that mention diseases, locations, and/or times of interest. We do so by leveraging the capabilities of STEWARD [19, 23], a system originally developed for *geotagging* documents from the hidden Web, i.e., mapping them to the geographic space (see also the related NewsStand system designed for news [33, 34, 36]). Geotagging consists of first finding in each document references to locations, called *toponyms*, and second, assigning each toponym a spatial interpretation in the form of lat/long values. This problem is challenging because toponyms exist in the space of natural language and hence exhibit ambiguity. Many toponyms are also names of other, non-location entities (e.g., "Paris, France", the location, versus "Paris Hilton", the person) [17], and further, many places have the same name (e.g., any of over 60 cities around the world named "Paris") [18, 21, 22, 29]. Likewise, many locations mentioned in ProMED postings will also be ambiguous, and it is here where STEWARD's utility is most apparent, in contrast to other disease tracking systems that are incapable of resolving such ambiguities. STEWARD's map-based Web interface also allows for powerful spatio-textual retrieval and display. In particular, it can be used for feature-based queries [1] (e.g., spatial data mining) as well as location-based queries (see also the related QUILT system [31, 35] and the SAND Browser [32]).

While STEWARD was previously used in an infectious disease tracking role [20], in this work we describe improvements to STEWARD that allow for intuitive and useful

*This work was supported in part by the National Science Foundation under Grants IIS-07-13501, IIS-08-12377, CCF-08-30618, IIS-09-48548, IIS-10-18475, and IIS-12-19023.

ACM SIGSPATIAL HealthGIS'12, November 6, 2012. Redondo Beach, CA, USA

spatio-temporal querying through the use of a *time slider*. This slider allows users to quickly and easily vary the temporal components of their queries by changing the time range and time size under consideration. This slider varies the display of relevant documents associated with locations on the map, by limiting the displayed locations to those associated with documents with times that fall within the time period covered by the slider. Further, by pressing a "play" button, the slider moves automatically across the entire time spanned by all result documents, and thus users can examine the changes in locations associated with a given disease or condition over time. In this way, disease detection and tracking amounts to observing these changes, which is made simple and intuitive by our improvements.

The rest of this paper is organized as follows. First, we outline related research systems and datasets (Section 2). Next, we describe our methods for retrieval and processing of ProMED postings, which leverage STEWARD's geotagging capabilities (Section 3). Then, we describe our Web interface, including the time slider improvements, and explain our implementation of temporal querying that enables the slider (Section 4) Finally, we offer concluding remarks and outline potential avenues for future research (Section 5).

2. RELATED WORK

A huge amount of resources have been devoted to the development of disease tracking and monitoring datasets, and associated strategies. In this section, we provide a brief survey of some of the systems and datasets that have been developed for this purpose. For more details see, e.g., Buckeridge et al. [2].

Disease surveillance systems can be classified in a variety of ways, but one simple measure is according to the types of data with which they make their predictions. In particular, datasets can be loosely characterized as pre-Web datasets, which predate the prevalence of Web-based communications, and complementary post-Web datasets.

Pre-Web datasets tend to consist of highly curated, verified, closed information that is limited in scope and size, which in turn limits their effectiveness in the context of a real-time surveillance system. Examples include information released by public governmental agencies devoted to health concerns. Such agencies release great quantities of public health information, and many public health services exist that are provided by these agencies, but we list only a few here. OutbreakNet [3], provided by the US Centers for Disease Control and Prevention, releases curated disease reports akin to press releases from a team of epidemiologists who investigate incidences of disease. The European Centre for Disease Prevention and Control maintains the Eurosurveillance journal [8], which publishes articles containing investigated reports of outbreaks. The Public Health Agency of Canada operates the Global Public Health Intelligence Network (GPHIN) [28], which collects news reports from around the world and employs human analysts to curate and verify reports, which are then released to subscribers—generally, various health organizations that then disseminate that information publicly in their own domains.

In addition to various public agencies, there exist a variety of private organizations that cater to specific member clients. One such example, the RODS network [37, 38], monitors member organizations' data, including hospital emergency department visits and retail store sales. The International Society of Travel Medicine's GeoSentinel network [10] collects data from faxable disease reports submitted by medical professionals, who report cases of patients with particular diseases visiting their practice. The ESSENCE II system [24] detects outbreaks using emergency department visits, as well as medication sales information, clinic visits, absentee records, and other sources, in a combined approach to disease detection.

Given the history and quantity of pre-Web data, many researchers have explored methods and techniques related to such data. Cooper et al. [6] model disease outbreaks as Bayesian networks and test a detection system based on these networks with emergency department records from several local hospitals. Likewise, Reis et al. [30] investigated methods for modeling simulated disease outbreaks as the result of bioterrorism using hospital data, and found that building such models using long-term as opposed to short-term data resulted in more accurate outbreak prediction. This indicates the importance of understanding the temporal components of disease outbreaks and calls for powerful and intuitive methods for temporal querying. Fienberg and Shmueli [9] note that the traditional forms of data used for biosurveillance, such as hospital visits, are inadequate for rapid outbreak detection, and outline various types of non-traditional data that may be used for detection, including school absence records, 911 call records, and their focus, grocery sales data. This finding alludes to the growing importance of Web-based datasets, such as the ProMED data that we use.

In contrast to pre-Web data, post-Web datasets consist primarily of open, minimally structured and verified information from a wide variety of nontraditional sources not specifically dedicated to health information, such as news reports, blogs, personal reports, and mailing lists such as ProMED (although ProMED is more curated than most sources). As a result, these datasets are much larger, and much noisier, than pre-Web datasets. Many systems have grown around post-Web data, and given the variety and magnitude of such data, they tend to act as aggregators in many respects. One well known system in this domain is HealthMap [4, 11], which gleans information from ProMED, GeoSentinel, Eurosurveillance, Google News, Moreover, and many other sources. The BioCaster system [5] gathers documents from several Web sources, including ProMED, and finds locations and other entities. Their main contribution is an expert-curated disease ontology that draws together disease information from multiple sources, and also has links to external disease databases. Also, the more recent GermTrax [12] employs collaborative disease tracking that relies primarily on disease reports from ordinary people who are sick. As a result, the system is intended for non-specific conditions such as "cold" or "flu". Grishman et al. [15] use ProMED and news data to search for outbreak events using a rule-based framework. Their browsing interface is rudimentary and presents only a tabular view of outbreaks, which limits spatial understanding.

All these systems, and others, suffer from the lack of a solid geotagging framework. They assign locations to disease reports using minimal metadata which does not correlate with the locations where diseases are occurring. Further, these systems lack an intuitive method for specifying temporal queries. The systems that do allow temporal querying tend to be limited to a single geographic area, or a single geographic resolution, with static displays of results. In contrast, STEWARD uses well-developed geotagging technology, in combination with simple and powerful temporal querying, to aid in the discovery and exploration of disease outbreaks.

Wilson and Brownstein [40] advocate the use of free Web-based data sources for rapid disease detection, as opposed to costly, closed traditional surveillance methods. In particular, they illustrate the utility of keyword searches in search engines and users' click streams. When many Web surfers from a single geographic region—determined by mapping their IP addresses to locations—all click on links related to the same disease, or search for the same disease keywords, it serves as an indication of prevalence of the disease within that region. Ginsberg et al. [13] describe one such implementation of this strategy for detecting outbreaks of influenza by observing patterns of influenza-related searches from particular geographic regions, determined by IP addresses. Areas with overlarge searches for "flu" and related keywords tended to experience more outbreaks of influenza. These techniques illustrate the growing importance of collaborative data collection and filtering, such as in the ProMED data which we use.

3. DATA PROCESSING

In this section, we describe the format of the ProMED data, and how we retrieve and process it for spatio-temporal querying and retrieval in STEWARD.

3.1 Data Format

ProMED is an online alert system intended to act as a medical information clearinghouse, by quickly disseminating news of infectious disease outbreaks to medical professionals and other subscribers around the world. ProMED's editors monitor news media reports, official government reports, and online disease summaries to learn of new cases or updates to existing cases. Monitored diseases are limited to those of humans, animals, and plants. In addition, medical professionals sometimes send local reports of diseases, or commentary related to existing disease reports, directly to ProMED's editors. The editors vet and verify each report, and if they determine that the report is found to be accurate, such reports are republished to ProMED's subscribers on one or several mailing lists, organized by topic, such as animal or plant diseases, emerging disease reports, broad location-oriented posts (e.g., Africa, Latin America, Southeast Asia), and others. Prior to republication, the editors add metadata to each report, including the report's date and time, organisms of relevance (i.e., human, animal, plant), and crucially, the location or locations affected by the disease outbreak. They may also modify the report's text or add suitable commentary from contributors. In addition, once or twice daily, reports are synthesized into single digests for easier republication.

Figure 1 shows a typical ProMED post which illustrates its general structure. This post was released during the 2003 outbreak of severe acute respiratory syndrome (SARS) in southeast Asia and other parts of the world. It consists of a reposting of a World Health Organization (WHO) travel advisory, relating details of the disease and a warning for travelers leaving from or going to various locations affected by the outbreak. The post begins with ProMED metadata that is present in all such postings, namely the `Published Date`, `Subject`, and `Archive Number`, which related the date of posting, subject of the post, and a unique post identifier which can be used to retrieve the original post from ProMED's website. The `Subject` line contains additional metadata, starting with tags that identify the mailing lists of relevance—in this case, `PRO/ALL`, which indicates the posting's relevance to all of ProMED's mailing lists. The line also mentions the disease of relevance, namely `Severe acute respiratory syndrome`, and the general geographic location of relevance, which in this case is `Worldwide`. Thus, all of ProMED's posts feature geographic locations very prominently, which is not surprising given the heavily geographic nature of reports of disease outbreaks.

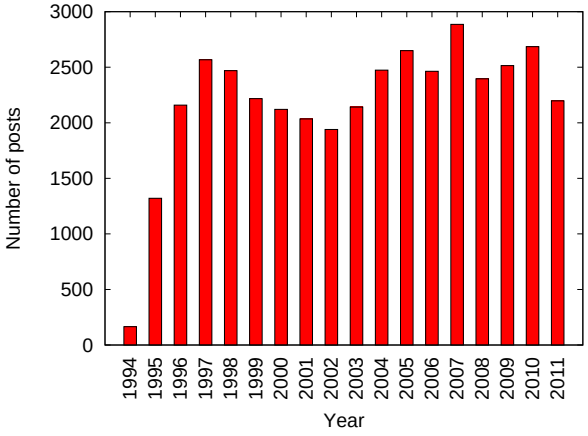
The rest of the post contains the actual content, which in this case is the full text of the corresponding WHO posting. This posting has a similar structure to news articles, starting with a *dateline* that contains the date and location of publication, namely `15 March 2003 | GENEVA`. This dateline has a clear format and both the date and location are readily machine-parseable. However, like the majority of such documents that describe ongoing events, the dates relevant to the events described in the content differ from the date of posting. Also, for digest and summary posts, the time need not have any relevance at all, since they act as groupings of older posts and will not represent the latest information. The very first sentence states that `during the past week` a large number of cases were reported, indicating the posting's relevance to this time range, as opposed to the posting time. Further, the dateline's location, `GENEVA`, in which the WHO is headquartered, is totally irrelevant to the locations of outbreaks of SARS, which are presented en masse in the post's body. Clearly, for this post and many others like it, naively tagging the post with locations and times mentioned in metadata, as opposed to content, will result in irrelevant results to spatio-temporal queries.

3.2 Data Retrieval

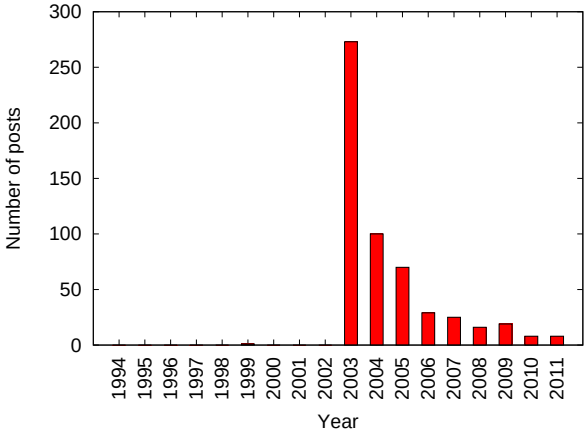
To retrieve an initial complement of data for our system, we crawled ProMED's website and downloaded its entire database of archived posts from 1994 to 2011, which numbered 39,420 posts in total. To give a general overview of the amount of data present, Figure 2 illustrates the volume of available ProMED posts, arranged by year. Examining Figure 2a, which shows the total number of ProMed posts per year, about 2,000–2,500 notices are consistently posted per year, or about six per day, which can be processed quickly. On the other hand, examining Figure 2b, showing counts of only SARS-related posts, a clear pattern of frequent posting appears in 2003, corresponding with the outbreak of SARS in that year, and tapers off in subsequent years, indicating a declining interest in such queries. Of course, these counts would be of little use to experts, who would be interested in much finer temporal ranges, and if we were to "zoom in"

```
Published Date: 2003-03-15 23:50:00
Subject: PRO/ALL> Severe acute respiratory syndrome - Worldwide:alert
Archive Number: 20030315.0637
[...]

World Health Organization issues emergency travel advisory
Severe Acute Respiratory Syndrome (SARS) Spreads Worldwide
----------------------

15 March 2003 | GENEVA -- During the past week, WHO has received reports of more
than 150 new suspected cases of Severe Acute Respiratory Syndrome (SARS), an
atypical pneumonia for which cause has not yet been determined. Reports to date
have been received from Canada, China, Hong Kong Special Administrative Region
of China, Indonesia, Philippines, Singapore, Thailand, and Viet Nam. Early
today, an ill passenger and companions who travelled from New York, United
States, and who landed in Frankfurt, Germany were removed from their flight and
taken to hospital isolation.
[...]
```

Figure 1: A typical ProMED posting, including metadata and body text.



(a) All posts



(b) SARS posts

Figure 2: Number of ProMED posts per year, considering (a) all posts and (b) only posts related to the keyword "SARS".

on particular years, we would see more variations throughout the year. However, we present these figures to give a high-level overview of the size and temporal aspects of our dataset.

3.3 Geotagging

After retrieving the ProMED data, we used STEWARD's existing infrastructure to associate each posting with the locations that it mentions and hence to which it is relevant. Lieberman et al. [19] provide details of such processing, but we provide a brief overview here.

The first stage of processing in STEWARD involves standardizing each document's format and extracting relevant metadata, including time of publication, title, and relevant keywords. STEWARD's purpose of geotagging documents from the hidden Web means this step may be involved for documents in general, which may include PDFs, MS Word, HTML, or other types of documents. However, for ProMED posts, there is not much work involved since the posts are in text format, and the date of publication and posting title are given with each post's metadata. Each posting is inserted into STEWARD's PostgreSQL database, with the posting's text indexed for full text searches using an *inverted index*.

Next, we use STEWARD's infrastructure to associate each posting with its locations, a process known as *geotagging*. As mentioned previously, each post's title contains general locations of relevance for the post, so one may be tempted to use these locations for geotagging. However, these locations are unsuitable for our purposes since they are of very coarse resolution, especially for widespread disease outbreaks (e.g., "worldwide"). Instead, STEWARD's geotagger finds references to locations, known as *toponyms*, in each posting's text, and further associates each toponym with its spatial interpretation in the form of lat/long values. These steps are difficult because toponyms exist in the space of human language, and as such exhibit the significant *ambiguity* that is characteristic of human language. In particular, many toponyms are also names of other entities (e.g., "Paris" may

refer to "Paris, France", a location, or "Paris Hilton", a person), and further, a given toponym may have many location interpretations (e.g., "Paris, France" versus "Paris, Texas" or any of over 60 other places named "Paris" all over the world).

STEWARD's geotagger uses many techniques to find toponyms and assign them lat/long values. To find toponyms, STEWARD uses a hybrid approach involving techniques from natural language processing (NLP), and in particular named-entity recognition (NER) and part-of-speech (POS) tagging. The NER task involves finding typed entities within free running text, including locations, but also variously persons, organizations, stock symbols, currencies, dates and times, genes and proteins, and others. Tools developed for NER can be leveraged for finding toponyms by simply retrieving the output location entities. Similarly, POS tagging involves assigning each word or token in a text with its corresponding part of speech. Names of locations, and other types of entities, tend to be proper nouns, so adjacent groups of proper nouns are taken as toponyms. Of course, some of these proper nouns will not be locations, but they will be filtered out in subsequent steps of processing.

After finding toponyms, STEWARD associates with each toponym one or more location interpretations from a location *gazetteer*, which is a database of locations and associated metadata. STEWARD uses a gazetteer constructed by merging two freely available gazetteers: the Geographic Names Information System (GNIS) [39], which contains US location interpretations, and the GeoNET Names Server (GNS) [27], which contains non-US locations. As of this writing, these gazetteers contain over 2.2 million and 5 million location interpretations, respectively, which means that STEWARD has excellent coverage of locations, at the cost of increased toponym ambiguity and hence geotagging difficulty. However, wide coverage is necessary for obtaining smaller locations of relevance that are crucial in disease detection and tracking.

Next, STEWARD resolves each toponym by assigning to it one of the interpretations associated with it. The resolution algorithm involves a variety of heuristic evidence applied through rules. For example, object/container pairs such as "Paris, Texas" indicate that the interpretation of "Paris" should be contained by the interpretation of "Texas". Additionally, STEWARD uses an algorithm termed *pair strength*, where pairs of toponym interpretations within the document are ranked according to their document distance, geographic distance, and population. The ranked pairs are then sorted, and toponyms are greedily resolved using the first pair in which they appear. For more details, see Lieberman et al. [19].

At this point, each document is associated with a set of toponyms and their location interpretations (i.e., lat/long values). The final stage of processing involves finding an overall geographic focus of each document by ranking the locations present in it. Locations are ranked using a combination of document frequency and document position—toponyms mentioned earlier are presumed to be more important to the content as a whole than those mentioned later. By ranking locations in this manner, spatial components of queries to STEWARD return more relevant postings.

4. WEB INTERFACE

In this section we describe STEWARD's Web interface, as enhanced for intuitive and powerful temporal querying via its time slider, as well as details of the time slider's implementation that enables interactive temporal querying.

4.1 Visual Elements

Figure 3 shows STEWARD's Web interface, along with our new extension to allow spatio-temporal querying capabilities. Here, the user has entered a query for "SARS", and the results are shown both in the textual pane at bottom left and the spatial pane at bottom right. The textual pane contains information about individual search results, with each entry corresponding to a ProMED posting. The primary locations associated with the postings are shown in the map pane, with each marker corresponding to a posting. In this case, all ProMED reports containing "SARS" are returned, which number 856 in our database.

In addition to the keyword search, the user has performed additional temporal filtering on the results by selecting the "Temporal" tab at the top, which enables the display of a time slider that grants intuitive access to STEWARD's temporal querying capability. Users move the slider along the timeline to control the temporal range in which query results are shown in the map pane. Further, users can drag either of the slider's endpoints to adjust the size of the slider's range, which correspondingly changes the time range of results displayed in the map. In Figure 3, the user has selected to only show notices that were posted in 2004. In addition, the timeline's endpoints change dynamically to match query results by being updated to the time span of the returned data. In this case, ProMED postings relevant to "SARS" span the years of 1997–2012. We used ArcGIS's public API [7] to facilitate the implementation of our slider.

In addition to moving the slider manually, users can click the "play" button to the left of the timeline to cause the time slider to automatically move across the timeline. The slider can also be moved stepwise forward and backward through time intervals by clicking the "step forward" and "step back" buttons to the right of the timeline. These features enable an intuitive and compelling display of the data's evolution over time, which allows the discovery of temporal trends and outliers. Figure 4 illustrates one such discovery of the worldwide outbreak of SARS using these features of the slider. ProMED postings relevant to "SARS" and posted in 2001, 2002, and 2003 are displayed when the slider moves over the corresponding times in the time scale. Bursts of ProMED postings related to "SARS" are limited in number, between 2001–2002, expand greatly in 2002–2003, and explosively appear in 2003–2004, reflecting the number of SARS cases and their news coverage during those time periods. The map also illustrates the outbreak's evolving geographic nature, with early cases concentrated in Southeast Asia, and later cases having worldwide distribution. While these examples are simplistic in scope, they serve as a powerful illustration of the power and potential of dynamic spatio-temporal querying.

4.2 Temporal Querying

Here, we provide some brief details of our implementation of temporal querying to support the above interface devel-

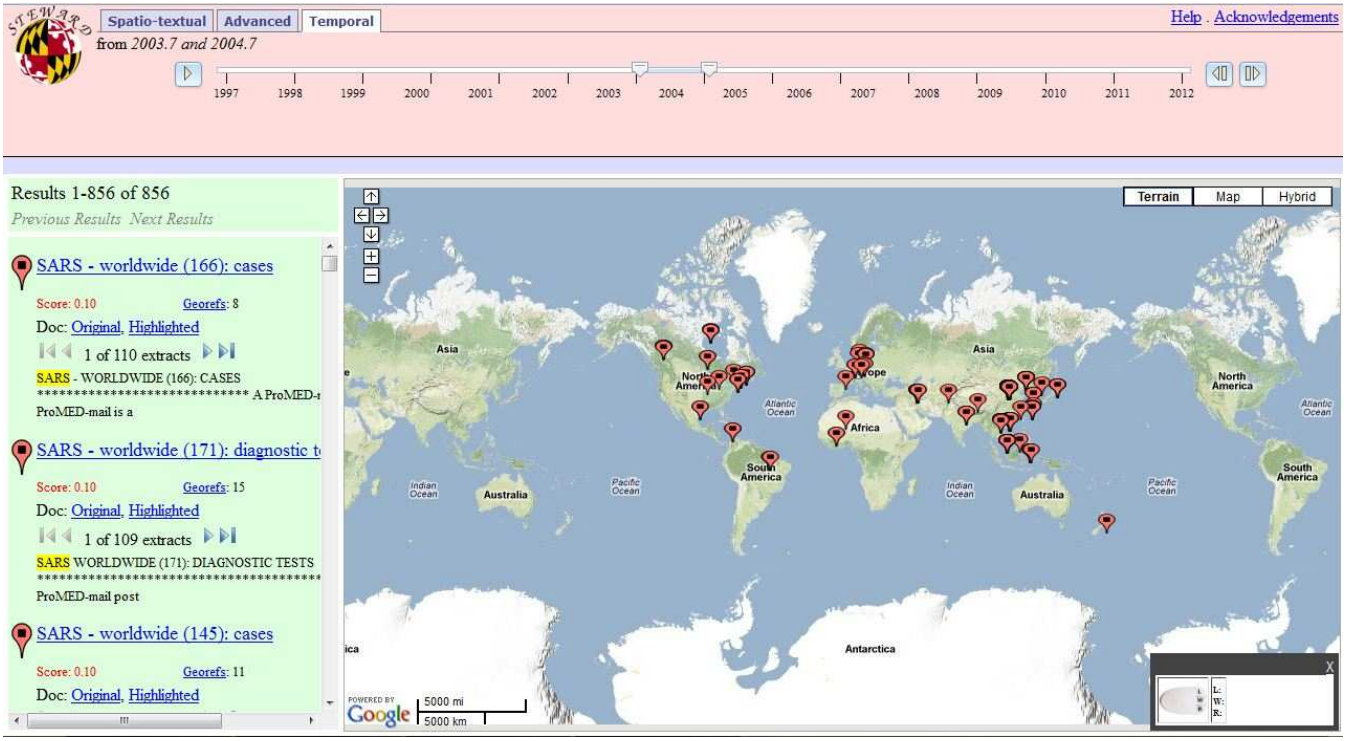


Figure 3: STEWARD's map-based Web interface after a query for "SARS". All relevant ProMED postings appear in the left textual pane, while the locations obtained by geotagging the postings appear as markers in the right map pane. Mapped locations are restricted to the range covered by the time slider at the top.

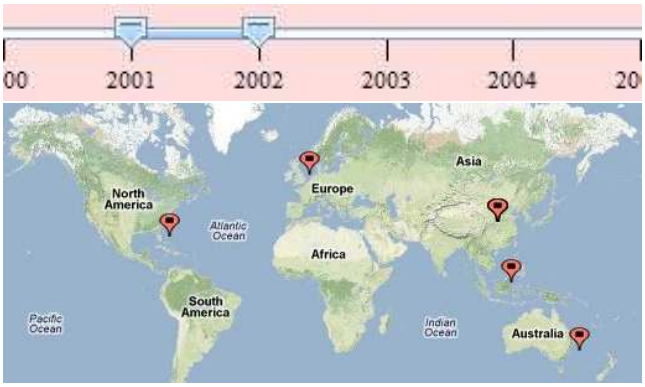
opments. Temporal aspects of queries are fully implemented in the client, in that the underlying database of STEWARD is not involved in the queries. To execute spatio-temporal queries, the keyword and spatial components of the query are first executed in STEWARD's database, and relevant documents R are retrieved. STEWARD's interface also supports top-k querying of the database, where the first k ranked results of R are returned rather than all of them. Users subsequently specify temporal parameters using the time slider described in the previous section, and these parameters determine which documents in R have their corresponding locations mapped in the Web interface. This determination is made by the client when rendering results. Thus, the temporal query component currently functions as a client-side postprocessing filter, although it would be better to integrate such queries more closely with the spatio-textual components.

Having a dynamic and configurable time slider demands a correspondingly fast implementation of temporal querying. However, for queries with relevance to a large portion of documents, a large number of result documents R are returned from the initial spatio-textual query. Initial versions of temporal querying implemented in STEWARD suffered from interactivity problems with large query results, due to the screen's limited update rate, as well as the Web browser's slow script processing. Therefore, prior to temporal querying via the slider, we perform simple indexing of result documents by placing them in a list and sorting them by time, using an in-place quicksort implementation. Temporal range querying then reduces to binary search in the sorted list.

Related to interactivity is the issue of rendering temporal query updates. When users move the slider, markers that fall outside the range are removed, and those that are now relevant are added. This could be implemented by clearing the map of all markers after every update, and then reading only the markers that are relevant to the time range. Unfortunately, this incurs unacceptable performance penalties and noticeably reduces interactivity, especially when using the automatic slider functionality. Instead, we leave in place those markers that are unaffected by the query update, i.e., those that remain in the range after the update.

5. CONCLUSION

STEWARD's geotagging capability, in combination with its use of a time slider, enables intuitive and powerful temporal querying for disease outbreaks. However, as a prototype system, the work completed thus far is preliminary and a number of improvements would increase its querying and retrieval capability, and are of general interest. First, our system could be augmented to consider additional data sources for processing. ProMED itself features postings in language other than English, such as Portuguese, Spanish, and Russian, and we could leverage machine translation capabilities to find relevant disease reports in these languages. In addition, PubMed data [26], which contains case studies of disease incidences as well as more recent disease reports, would be a valuable source of data for temporal queries. We have retrieved all PubMed documents from 2011, consisting of 885,316 documents, and are currently processing them for inclusion in our retrieval system. We could also potentially use tweets from individuals with particular sicknesses.



(a) 2001

(b) 2002

(c) 2003

Figure 4: Origin and spread of the 2003 SARS outbreak. Markers correspond to geotagged ProMED postings. As the time slider is moved across times of interest, the locations affected by the outbreak are easily apparent.

Because each tweet contains a large amount of metadata, including GPS values and timestamp, this would provide a source of timestamped geographic information. Of course, filters would have to be developed to filter out the vast majority of tweets which do not concern diseases, and likewise to determine the veracity of disease-related tweets.

Additional work could be done related to information extraction within ProMED postings or other unstructured text documents. Currently STEWARD's geotagger provides a robust mechanism for associating documents with the locations contained within them. However, it has a limited facility for recognizing dates and times, and hence is currently limited to temporal retrieval using the timestamp associated with each ProMED posting. Recognizing dates and times in the text, in addition to locations, would enhance STEWARD's temporal querying capability for such data. These dates and times could be explicit, such as "June 1", but could also include relative times such as "last week" or "yesterday" which would be normalized to the dates to which they correspond. In addition, many ProMED postings consist of daily post digests, or summaries of multiple posts about the same disease. These posts could be more effectively processed by developing techniques for detecting such divisions within the document, and processing it with these distinctions in mind.

STEWARD's Web interface could be enhanced in several ways to improve its usability with temporal querying. Currently, all documents within the time slider's range are displayed on the map. For times with disease outbreaks, this results in the map being completely filled with markers that overlap significantly, which in turn impedes understanding. Instead, we could rank and display documents on the map using a mix of *importance*, measured by the prevalence of disease at the location, and *geographic spread*, where we ensure good coverage of the map. This feature would be useful both from a usability perspective, since users will not be overwhelmed with too many markers on the map, and from a database performance perspective, since fewer markers need to be retrieved. Additionally, the timeline could augment users' exploratory capability by examining the distribution of documents throughout the timeline and suggesting times of interest for users to query. This would be accomplished by clustering documents in the time dimension, and searching for clusters and outliers.

Finally, the temporal aspects of queries could be more tightly integrated into STEWARD's database. Currently, the ProMED dataset is relatively small with an infrequent publication rate, but to apply our processing to other, much larger datasets such as tweets, faster indexing and retrieval methods are required. In particular, spatio-temporal indexes [25] could be integrated or developed that combine times and locations, which would leverage the database's query planner to more efficiently perform spatio-temporal queries. Development of such techniques, in combination with the ever-increasing availability of public health reports and data, ensures that health-related GIS plays a central role in maintaining and improving societal health.

References

[1] W. G. Aref and H. Samet. Efficient processing of window queries in the pyramid data structure. In *PODS'90: Proceedings of the 9th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 265–272, Nashville, TN, Apr. 1990.

[2] D. L. Buckeridge, H. Burkom, M. Campbell, W. R. Hogan, and A. W. Moore. Algorithms for rapid outbreak detection: a research synthesis. *Journal of Biomedical Informatics*, 38(2):99–113, Apr. 2005.

[3] Centers for Disease Control and Prevention. CDC - OutbreakNet Team, Sept. 2012. URL `http://www.cdc.gov/outbreaknet`.

[4] Children's Hospital Boston. HealthMap | Global Health, Local Knowledge, Sept. 2012. URL `http://www.healthmap.org`.

[5] N. Collier, S. Doan, A. Kawazoe, R. M. Goodwin, M. Conway, Y. Tateno, Q.-H. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, M. Shigematsu, and K. Taniguchi. BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics*, 24(24):2940–2941, Oct. 2008.

[6] G. F. Cooper, D. Dash, J. Levander, W.-K. Wong, W. R. Hogan, and M. M. Wagner. Bayesian biosurveillance of disease outbreaks. In *UAI'04: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 94–103, Banff, Canada, July 2004.

[7] Esri. ArcGIS Online, Sept. 2012. URL `http://www.arcgis.com`.

[8] European Centre for Disease Prevention and Control. Eurosurveillance, Sept. 2012. URL `http://www.eurosurveillance.org`.

[9] S. E. Fienberg and G. Shmueli. Statistical issues and challenges associated with rapid detection of bioterrorist attacks. *Statistics in Medicine*, 24(4):513–529, Feb. 2005.

[10] D. O. Freedman, P. E. Kozarsky, L. H. Weld, and M. S. Cetron. Geosentinel: The global emerging infections sentinel network of the international society of travel medicine. *Journal of Travel Medicine*, 6(2):94–98, June 1999.

[11] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein. HealthMap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *JAMIA: Journal of the American Medical Informatics Association*, 15(2):150–157, Mar. 2008.

[12] GermTrax. Who is sick? track the spread of sickness and disease at germtrax, Sept. 2012. URL `http://www.germtrax.com`.

[13] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457 (7232):1012–1014, Feb. 2009.

[14] M. F. Goodchild. Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4):211–221, Aug. 2007.

[15] R. Grishman, S. Huttunen, and R. Yangarber. Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, 35(4):236–246, Aug. 2002.

[16] International Society for Infectious Diseases. ProMED-mail, Sept. 2012. URL `http://www.promedmail.org`.

[17] M. D. Lieberman and H. Samet. Multifaceted toponym recognition for streaming news. In *SIGIR'11: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 843–852, Beijing, China, July 2011.

[18] M. D. Lieberman and H. Samet. Adaptive context features for toponym resolution in streaming news. In *SIGIR'12: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 731–740, Portland, OR, Aug. 2012.

[19] M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling. STEWARD: Architecture of a spatio-textual search engine. In *GIS'07: Proceedings of the 15th ACM International Symposium on Geographic Information Systems*, pages 186–193, Seattle, WA, Nov. 2007.

[20] M. D. Lieberman, J. Sankaranarayanan, H. Samet, and J. Sperling. Augmenting spatio-textual search with an infectious disease ontology. In *IIMAS'08: Proceedings of the Workshop on Information Integration Methods, Architectures, and Systems*, pages 266–269, Cancún, Mexico, Apr. 2008.

[21] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging: Using proximity, sibling, and prominence clues to understand comma groups. In *GIR'10: Proceedings of the 6th Workshop on Geographic Information Retrieval*, Zurich, Switzerland, Feb. 2010.

[22] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *ICDE'10: Proceedings of the 26th International Conference on Data Engineering*, pages 201–212, Long Beach, CA, Mar. 2010.

[23] M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling. STEWARD – Spatio-Textual Extraction on the Web Aiding Retrieval of Documents, Sept. 2012. URL `http://steward.umiacs.umd.edu`.

[24] J. S. Lombardo, H. Burkom, and J. Pavlin. ESSENCE II and the framework for evaluating syndromic surveillance systems. *MMWR: Morbidity and Mortality Weekly Report*, 53(suppl):159–165, Sept. 2004.

[25] M. F. Mokbel, T. M. Ghanem, and W. G. Aref. Spatio-temporal access methods. *IEEE Data Engineering Bulletin*, 26(2):40–49, June 2003.

[26] National Center for Biotechnology Information. Home - PubMed - NCBI, Sept. 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed`.

[27] National Geospatial-Intelligence Agency. NGA: GNS Home, Sept. 2012. URL `http://earth-info.nga.mil`.

[28] Public Health Agency of Canada. The Global Public Health Intelligence Network (GPHIN) - Public Health Agency of Canada, Sept. 2012. URL `http://www.phac-aspc.gc.ca/gphin`.

[29] G. Quercini, H. Samet, J. Sankaranarayanan, and M. D. Lieberman. Determining the spatial reader scopes of news sources using local lexicons. In *GIS'10: Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 43–52, San Jose, CA, Nov. 2010.

[30] B. Y. Reis, M. Pagano, and K. D. Mandl. Using temporal context to improve biosurveillance. *PNAS: Proceedings of the National Academy of Sciences of the United States of America*, 100(4):1961–1965, Feb. 2003.

[31] H. Samet, A. Rosenfeld, C. A. Shaffer, and R. E. Webber. A geographic information system using quadtrees. *Pattern Recognition*, 17(6):647–656, November/December 1984.

[32] H. Samet, H. Alborzi, F. Brabec, C. Esperança, G. R. Hjaltason, F. Morgan, and E. Tanin. Use of the SAND spatial browser for digital government applications. *Communications of the ACM*, 46(1):63–66, Jan. 2003.

[33] H. Samet, M. D. Adelfio, B. C. Fruin, M. D. Lieberman, and B. E. Teitler. Porting a web-based mapping application to a smartphone app. In *GIS'11: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 525–528, Chicago, Nov. 2011.

[34] H. Samet, B. E. Teitler, M. D. Adelfio, and M. D. Lieberman. Adapting a map query interface for a gesturing touch screen interface. In *WWW'11: Proceedings of the 20th International World Wide Web Conference*, pages 257–260, Hyderabad, India, Mar. 2011.

[35] C. A. Shaffer, H. Samet, and R. C. Nelson. QUILT: a geographic information system based on quadtrees. *International Journal of Geographical Information Systems*, 4(2):103–131, April–June 1990. Also University of Maryland Computer Science Technical Report TR–1885.1, July 1987.

[36] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. NewsStand: A new view on news. In *GIS'08: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 144–153, Irvine, CA, Nov. 2008.

[37] F.-C. Tsui, J. U. Espino, V. M. Dato, P. H. Gesteland, J. Hutman, and M. M. Wagner. Technical description of RODS: A real-time public health surveillance system. *JAMIA: Journal of the American Medical Informatics Association*, 10(5):399–408, Sept. 2003.

[38] University of Pittsburgh. Real-time Outbreak and Disease Surveillance Laboratory - Home, Sept. 2012. URL `http://www.rods.pitt.edu`.

[39] U.S. Geological Survey. U.S. Board on Geographic Names (BGN), Sept. 2012. URL `http://geonames.usgs.gov`.

[40] K. Wilson and J. S. Brownstein. Early detection of disease outbreaks using the internet. *CMAJ: Canadian Medical Association Journal*, 180(8):829–831, Apr. 2009.