An Overview of Geotagging

IRL dri

Geospatial location

Archival collections may include a huge amount of original documents which contain geospatial location (toponym) data. Scientists, researchers and archivists are aware of the value of this information, and various tools have been developed to extract this key data from archival records (Kemp, 2008). Geospatial identification metadata generally involves latitude and longitude coordinates, coded placenames and data sources, etc. (Hunter, 2012).

Tagging

Tagging, a natural result of social media and semantic web technologies, involves labelling content with arbitrarily selected tags. Geotags, which describe the geospatial information of the content, are one of the most common online tag types (Hu & Ge, 2008), (Intagorn , et al., 2010).

Geotagging

Geotagging refers to the process of adding geospatial identification metadata to various types of media such as e-books, narrative documents, web content, images and video and social media applications (Facebook, Twitter, Foursquare etc) (Hunter, 2012). With various social web tools and applications, users can add geospatial information to web content, photographs, audio and video. Using hardware integrated with Global Positioning System (GPS) receivers, this can be done automatically (Hu & Ge, 2008).

Social media tools with geo-tagging capabilities

Del.icio.us[1] — A social bookmarking site that allows users to save and tag web pages and resources.

CiteULike[2] — An online service to organise academic publications that allows users to tag academic papers and books.

Twitter[3] — A real-time micro-blogging platform/information network that allows users all over the world to share and discover what is happening. Users can publish their geotag while tweeting.

Flickr[4] — A photo-sharing service that allows users to store, tag and geotag their personal photos, maintain a network of contacts and tag others photos.

YouTube[5] — A video sharing system that allows users to upload video content and describe it with tags.

Last.fm[6] — A music information database that allows members to tag artists, albums, and songs.

Technorati[7] — A weblog aggregator and search tool that allows blog authors to tag their posts.

Four Square[8] — A cross between a friend-finder, a social city guide and a game that rewards you for doing interesting things based on the existing location data sent from the mobile phone.

In the geographical information systems domain, 'georeferencing' and 'geocoding' are sometimes used instead of the term geotagging. Georeferencing is defined as specifying the geographic location of an object, entity, phenomenon, image, concept, data, or information with universal parameters (direct georeferencing), code, or place (indirect georeferencing). Geocoding is defined as the process of finding the mathematical representation of a location. In addition, georegistration and rectification are sometimes used synonymously with georeferencing in the context of photographs, satellite images or scanned maps (Kemp, 2008).

[1] www.del.icio.us, last accessed 21 March 2016.
[2] www.citeulike.org, last accessed 21 March 2016.
[3] https://twitter.com, last accessed 21 March 2016.
[4] www.flickr.com, last accessed 21 March 2016.
[5] www.youtube.com, last accessed 21 March 2016.
[6] www.last.fm, last accessed 21 March 2016.
[7] www.technorati.com, last accessed 21 March 2016.
[8] www.foursquare.com, last accessed 21 March 2016.

Geotagging Approaches

Geotagging methods have been implemented in many different textual domains such as web pages, blogs, encyclopaedia articles, spreadsheets, on Twitter, the hidden Web, and news articles. Domains such as blogs and Twitter may pose additional challenges, such as having little or no formatting or grammatical requirements (Lieberman & Samet, 2012).

To annotate web content, various approaches using geospatial contextual information may be employed such as manual annotation by the author, annotation via location-aware devices, determining the location of the server, and automated annotation of existing content (geoparsing and geocoding) (Scharl, 2007).

The traditional approach to geotagging is to use trained human editors to read each document and manually assign geographic tags. This method is not scalable to the size of modern archives and is therefore not feasible. Automated software algorithms are replacing human indexers as the primary mechanism for geographic indexing of large text archives (Leetaru, 2012).

The most common approach, used by toolkits such as General Architecture for Text Engineering (GATE)[9], is to simply perform a keyword search for the names of countries, their capital, and their major cities. While very easy to implement and fast to run, this approach is problematic for a variety of reasons. One of the greatest limitations is that it is often assumed, especially in the case of news articles, that readers share common background knowledge about locations. Furthermore, even when a match is found, the resulting geographic index permits only country-level searches, while users are increasingly interested in the subnational information. Finally, not all country names are unambiguous (Leetaru, 2012).

For more efficient geotagging, geospatial references within unstructured content are automatically determined and linked to geographical identifiers such as codes and coordinates. (Kemp, 2008), (Gelernter & Balaji, 2013). In some references, this approach is named as full text geocoding, or Geographic Information Retrieval (Leetaru, 2012). This geotagging approach includes two steps; geoparsing and geocoding. The geoparsing process (or toponym recognition) involves scanning and finding geographical entities in the text using Natural Language Processing techniques (Hecht & Gergle, 2011), (Scharl, 2007). Geoparsing is related to the Natural Language Processing task known as Named Entity Recognition (NER), dealing with the detection of general named entities (Nesi, et al., 2014).

Geocoding (or toponym resolution) refers to the process of selecting relevant geospatial (toponym) information from an external location knowledge base, namely a gazetteer (Lieberman & Samet, 2012), (Leetaru, 2012). The goal of the geoparsing process is to disambiguate toponyms from non-geographic named entities (solving "geo/non-geo" ambiguity). In other words, geoparsing resolves geo/non-geo ambiguity, whereas geocoding resolves geo/geo ambiguity (Hecht & Gergle, 2011).

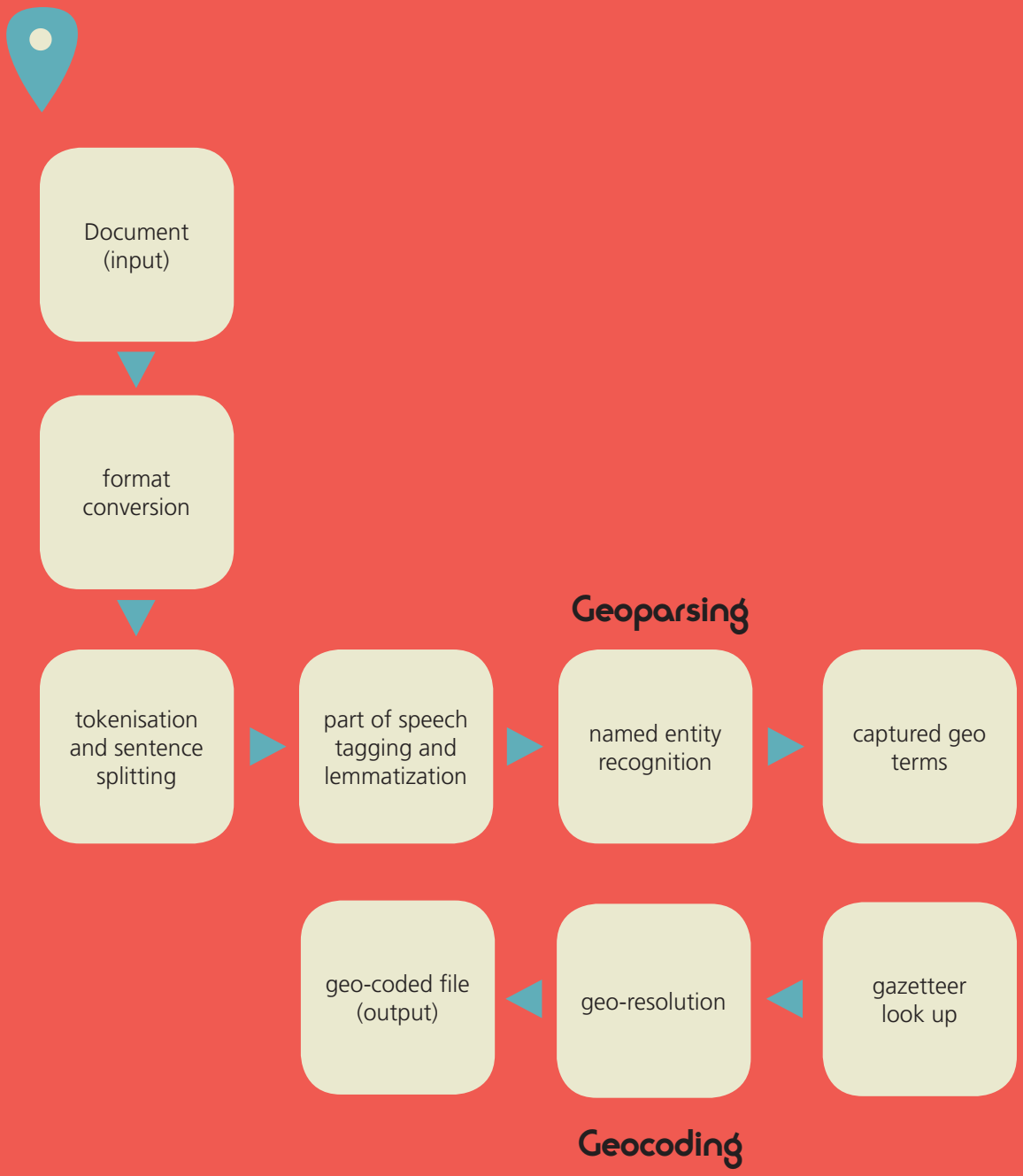
In the geoparsing process, the text is analysed and named entities are extracted using pattern matching and statistical based solutions. Other frequently used approaches are based on artificial neural networks, hidden Markov models, and maximum entropy models. Another family of methods which makes use of internal approaches are Natural Language Processing based solutions, such as Part-Of-Speech (POS) tagging. Typically, both of these methods are combined with an external resource using gazetteers or databases (Nesi, et al., 2014).

[9] https://gate.ac.uk/, last accessed 21 March 2016.

Geotagging Software

There are a number of automatic geotagging tools (Leetaru, 2012). These tools have two components: geoparsing (Named Entity Recognition – NER) and georesolution components. The geoparsing component parses text and defines possible names of geographical places using Natural Language Processing techniques. The georesolution component looks up the location names in a gazetteer and resolves ambiguities to suggest the most likely interpretation for each location given its context (Gelernter & Balaji, 2013), (D'Ignazio, et al., 2014).

The simplified process is represented below (Grover, et al., 2010).



Geotagging tools

NewsStand's Geotagger[10]	This tool tracks RSS feeds from online news sources, retrieves articles, and then extracts geographic content from them (Teitler, et al., 2008). It also assigns geographic locations to clusters of news articles based on their content, and allows users to visually explore the news in an interactive map interface (Lieberman & Samet, 2012).
YahooBOSS Geo Services[11]	Yahoo BOSS Geo Services includes two tools, Placefinder and PlaceSpotter. These tools allow developers to enrich their applications with geographical information and make them more location aware. Yahoo PlaceFinder is a geocoding web service that helps developers make their applications location-aware by converting street addresses or placenames into geographic coordinates. Yahoo PlaceSpotter helps developers make their applications location-aware by identifying places in unstructured and "atomic" content such as feeds, web pages, news articles, and status updates, and returning geographic metadata for geographic indexing and markup.
MetaCarta[12]	MetaCarta is perhaps the best known geotagging tool and is widely used commercially, as well as by the US Government, but is focused primarily on contemporary geographies and is costly (Leetaru, 2012). MetaCarta identifies and leverages the geographic references in content. It searches and finds all of the information about a place. It also analyses content to identify geographic terms and references using Natural Language Processing. It accesses a digital gazetteer that includes over 225 million disambiguated placenames with unique latitude and longitude. Finally, it puts data on any map such as Google, Microsoft and ESRI - virtually any map server.
Thomson Reuter's OpenCalais[13]	A tool to tag the people, places, companies, facts and events in content using Natural Language Processing and machine learning algorithms.

[10] http://newsstand.umiacs.umd.edu/web/, last accessed 21 March 2016.
[11] https://developer.yahoo.com/boss/geo/, last accessed 21 March 2016.
[12] http://www.metacarta.com/, last accessed 21 March 2016.
[13] http://new.opencalais.com/, last accessed 21 March 2016.

Drupal Geoparser[14]	The geoparser for the Drupal content management system which provides a common interface for geoparsing services.
Edina Unlock Text[15]	Unlock Text is a Natural Language Processing geoparser system developed by the Language Technology group at the School of Informatics at the University of Edinburgh. This tool can search text hosted on the web in .TXT or HTML formats for references to locations. These locations are then returned ready for use in a results page, web map or any other application.
Rosoka Geospatial Analysis Products[16]	Rosoka NLP Geo is a complete Natural Language Processing and Geospatial Analysis software product bundle that includes the Rosoka NLP product suite and the GeoGravy product suite. Rosoka GeoGravy is a stand-alone geospatial tagging service and gazetteer. Rosoka Geotagger provides the most likely disambiguated geotag based on the context of how a place name is used in a document regardless of language.
CLAVIN (Cartographic Location and Vicinity Indexer)[17]	An open source software package for document geotagging and geoparsing. It extracts location names from unstructured text and resolves them against a gazetteer to produce data-rich geographic entities. CLAVIN uses intelligent heuristics based on the context of the document and also employs fuzzy search to handle incorrectly-spelled location names, and it recognises alternative names as referring to the same geographic entity.
Geotxt OpenSextant[18]	GeoTxt detects location names mentioned in text and assigns geographic coordinates to those locations. Currently GeoTxt development is focused particularly on support for processing short microblog posts.
(The Open Spatial Extraction and Tagging)[19]	OpenSextant provides an unstructured textual data geotagging and geocoding capability. This open source software was developed by American governmental agencies.
NetOwl Extractor[20]	NetOwl Extractor offers named entity extraction, relationship and event extraction, geotagging and sentiment analysis in multiple languages using computational linguistics and Natural Language Processing technologies.

[14] https://www.drupal.org/project/geoparser, last accessed 21 March 2016.
[15] http://edina.ac.uk/unlock/texts/, last accessed 21 March 2016.
[16] http://www.rosoka.com/content/geospatial-analysis, last accessed 21 March 2016.
[17] https://clavin.bericotechnologies.com/, last accessed 21 March 2016.
[18] http://www.geotxt.org/, last accessed 21 March 2016.
[19] http://www.opensextant.org, last accessed 21 March 2016.
[20] https://www.netowl.com/entity-extraction/, last accessed 21 March 2016.

Gazetteers and Thesauri

A gazetteer is a list of names of geographic places or features, together with their geographic locations and other descriptive information. A thesaurus is a "vocabulary of a controlled indexing language, formally organised so that a priori relationships between concepts (for example "broader" and "narrower") are made explicit." A thesaurus usually provides a preferred term, synonym or quasi-synonym of a preferred term and relationships between the terms (Brauner, et al., 2006).

Digital gazetteers, readily available on the web, have been developed to support geospatial information systems (Breitman, et al., 2007). Well known gazetteers and thesauri for geographical places are presented below.

Gazetteers and thesauri for geographical places

Geoscience Australia[21]	In this country-specific gazetteer, the Australian Government provides a database of 322,000 locations in Australia (Australia Search) (Leetaru, 2012).
The Canadian Atlas Online gazetteer[22]	Similar to Geoscience Australia, in this country-specific gazetteer, Canada's government offers 350,000 locations including 200 Canadian cities, towns and villages. It is also possible to browse through maps that pinpoint each location using the interactive mapping tool (Leetaru, 2012).
The Geographic Names Information System (GNIS)[23]	The Geographic Names Information System (GNIS) is the Federal and national standard for geographic nomenclature in the USA. The U.S. Geological Survey developed the GNIS in support of the U.S. Board on Geographic Names as the official repository of domestic geographic names data, the official vehicle for geographic names used by all departments of the Federal Government, and the source for applying geographic names to Federal electronic and printed products.
Eurogeonames[24]	Eurogeonames provides names for Europe, built jointly by European mapping agencies. Smart software links all national databases with their different feature categories, name models and terminology together and makes them accessible as a virtual names server for all of Europe.

[21] http://www.ga.gov.au/place-names/index.xhtml, last accessed 21 March 2016.
[22] http://www.canadiangeographic.ca/atlas/gazetteer.aspx?lang=En, last accessed 21 March 2016.
[23] http://geonames.usgs.gov/domestic/index.html, last accessed 21 March 2016.
[24] http://www.eurogeographics.org/eurogeonames, last accessed 21 March 2016.

The GEOnet Names Server (GNS)[25]	Provides access to the National Geospatial Intelligence Agency (NGA) and the U.S. BGN database of foreign geographic names, containing about 4 million features with 5.5 million names (Brauner, et al., 2006). The Alexandria Digital Library (ADL) Gazetteer[25]
The Alexandria Digital Library (ADL) Gazetteer[26]	The ADL Project is a research project for modelling, educational applications, and software components. The ADL Project also developed HTML clients to access the ADL collections and gazetteer. The ADL Gazetteer prototype to evaluate digital library architectures and gazetteer applications has approximately 5.9 million geospatial names. It combines data from the GNIS (U.S. names) and the GNPS (non-U.S. names) (Breitman, et al., 2007).
Geonames[27]	A geographical database containing millions of geographical names. This database is maintained by a small community of experts from several different countries (Intagorn , et al., 2010). The data are accessible through a user interface, as well as through web services that offer geocoding based on geographical names or postal codes (Breitman, et al., 2007).
The Getty Thesaurus of Geospatial Names (TGN)[28]	A structured, world-coverage vocabulary of 1.3 million names, coordinates and other information for around 892,000 geospatial places. For each placename, it contains a unique ID, a set of place types taken from the Art and Architecture Thesaurus, alternative versions of the name, a containing administrative region, and a footprint in the form of a point in latitude and longitude. The temporal coverage of the TGN ranges from prehistory to the present and the scope is global (Breitman, et al., 2007).
The Fuzzy Gazetteer (European Commission/ JRC Digital Map Archive)[29]	FuzzyG 1.0 is the result of research collaboration between Hof University (Germany) and the European Joint Research Centre. Its purpose is to better inform humanitarian and foreign affairs decision makers on the landscape and environment of places in the world. FuzzyG searches for place names worldwide and can handle variations in spelling.
UN/EC Common Gazetteer Search (JRC Fuzzy Gazetteer)[30]	The UN Gazetteer was developed by the UN Cartographic Section (UNCS). It employs "fuzzy logic" to find place name locations worldwide. It accomplishes this by searching with a place name's phonetic spelling (how it sounds) and searching through a database that contains over 8 million entries. Many placenames have multiple spellings. A search for a location may not return a result even though a different spelling is in the database. The database of place names would be incorporated into the UN map geo-database for additional utilisation as well.
Global Gazetteer (Version 2.3)[31]	A non-comprehensive directory, it includes elevation and airport information for a limited number of places.

[25] http://geonames.nga.mil/gns/html/, last accessed 21 March 2016.
[26] http://legacy.alexandria.ucsb.edu/, last accessed 21 March 2016.
[27] http://www.geonames.org/, last accessed 21 March 2016.
[28] http://www.getty.edu/research/tools/vocabularies/tgn/about.html, last accessed 21 March 2016.
[29] http://isodp.hof-university.de/fuzzyg/query/, last accessed 21 March 2016.
[30] http://dma.jrc.it/services/fuzzyg/, last accessed 21 March 2016.
[31] www.fallingrain.com/world, last accessed 21 March 2016.

GeoTagging in Archives

Place and time are valuable concepts for researchers and the use of spatial and temporal filters to study datasets (including narratives) is increasingly important. In most archival data, there is a significant amount of geographic information in natural language. This kind of information is geographically relevant, but not geographically discoverable (Hecht & Gergle, 2011).

Geotagging (geoparsing and geocoding) is an efficient technology that has been applied to archival data in various projects. Digitising archival records and then automatically geotagging them may be a useful way to improve geographical search capacities and extract relevant information. Accurate, automatic, geographical referencing allows archival researchers to discover information relating to time and location as they search through records. If digitised data were geotagged, this would ensure a holistic understanding of the data by adding to the pool of distributed resources.

There are various examples of projects which geotag archives. Edinburgh Geoparser, for example, is used to process historical text in various projects e.g. Trading Consequences, Google Ancient Places (GAP) and The Digital Exposure of English Place-names (DEEP) project. In Trading Consequences, a Digging into Data II Project (CIINN01), the aim was to increase historians' understanding of the economic and environmental consequences of commodity trading in the British Empire in the 19th century. In this project, researchers analysed large quantities of digitised historical text from major British and Canadian datasets using text mining methods. The Edinburgh Geoparser was combined with the GeoNames gazetteer as part of the text mining component to determine locations in the archives. The goal of the GAP project was to geotag English translations of Greek and Roman classical texts. In the DEEP project, a detailed historical gazetteer for most of the counties in England was developed (Alex, et al., 2015).

The Edinburgh Geoparser was also used in the GeoDigRef and the Embedding GeoCrosswalk projects to geotag digitised historical collections. These collections were separately digitised, processed and geocoded using GeoNames gazetteer or the Ordnance Survey-derived GeoCrossWalk gazetteer (Grover, et al., 2010).

Geotagging allows researchers to compare historical archival datasets and to extract information effectively. During this process, researchers may encounter some significant problems. The available geotagging tools, gazetteers and approaches may be insufficient in terms of the project goals and the characteristics of the studied dataset. For example, the meaning of placenames may have changed over time and it might be highly unlikely that they are properly grounded to the correct coordinates. Finally, new tools, algorithms and approaches may need to be developed for the particular specifications of a project.

References

Alex, B., Byrne, K., Grover , C. & Tobin, R., 2015 . Adapting the Edinburgh Geoparser for Historical Georeferencing. International Journal of Humanities and Arts Computing, 9(1), pp. 15-35.

Brauner, D. F., Casanova, M. A., Breitman, K. K. & Leme, L. A., 2006. Using Gazetteers to Annotate Geographic Catalog Entries.. Paphos, Cyprus, ICEIS 2006 - Proceedings of the Eighth International Conference on Enterprise Information Systems: Databases and Information Systems Integration.

Breitman, K. K., Casanova, M. A. & Truszkowski, W., 2007. Semantic Web, Concepts, Technologies and Applications. 1st ed. Londan, UK: Springer-Verlag London.

D'Ignazio, C., Bhargava, R., Zuckerman, E. & Beck, L., 2014. CLIFF-CLAVIN, Determining Geographic Focus for News Articles. s.l., Extended abstract published in the proceedings of NewsKDD .

Gelernter, J. & Balaji, S., 2013. An algorithm for local geoparsing of microtext. GeoInformatica, 17(4), pp. 635-667.

Grover, C. et al., 2010. Use of the Edinburgh geoparser for georeferencing digitised historical collections. Philosophical Transactions of the Royal Society, 368(1925), pp. 3875-3889.

Grover, G. & Tobin, R., 2014. A Gazetteer and Georeferencing for Historical English Documents. Gothenburg, Sweden, Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014.

Hecht, B. & Gergle, D., 2011. Chapter 19: A Beginner's Guide to Geographic Virtual Communities Research. In: B. K. Daniel, ed. Handbook of Research on Methods and Techniques for Studying Virtual Communities: Paradigms and Phenomena, Vol.1. New York, USA: Information Science Reference (an imprint of IGI Global), pp. 333-347.

Hunter, D. S., 2012. eBook Geotagging: Linking Literature and Location. [Online] Available at: http://www.gis.smumn.edu/GradProjects/HunterD.pdf [Accessed 04 08 2015].

Hu, Y.-H. & Ge, L., 2008. Chapter 11: GeoTagMapper: An Online Map-based Geographic Information Retrieval System for Geo-Tagged Web Content. In: M. P. Peterson, ed. International Perspectives on Maps and the Internet Lecture Notes in Geoinformation and Cartography. Berlin, Germany: Springer Berlin Heidelberg, pp. 153-164.

Intagorn , S., Plangprasopchok , A. & Lerman, K., 2010. Harvesting geospatial knowledge from social metadata. Seattle, USA, Proceedings of the 7th International ISCRAM Conference.

Kemp, K. K., 2008. Encyclopedia of Geographic Information Science. 2nd ed. Thousand Oaks, CA: SAGE Publications, Inc..

Leetaru, K. H., 2012. Fulltext Geocoding Versus Spatial Metadata for Large Text Archives: Towards a Geographically Enriched Wikipedia. D-Lib Magazine, 18(9/10).

Lieberman, M. D. & Samet, H., 2012. Adaptive Context Features for Toponym Resolution in Streaming News. Portland, OR, USA, Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information.

Nesi, P., Pantaleo, G. & Tenti, M., 2014. Ge(o)Lo(cator): Geographic Information Extraction from Unstructured Text Data and Web Documents. Corfu, Greece, 9th International Workshop on Semantic and Social Media Adaptation and Personalization.

Scharl, A., 2007. Towards the Geospatial Web: Media Platforms for Managing Geotagged Knowledge Repositories. In: A. Scharl & K. Tochtermann, eds. The Geospatial Web - How Geo-Browsers, Social Software and the Web 2.0 are Shaping the Network Society. London, UK: Springer, pp. 3-14.

Teitler, B. E. et al., 2008. NewsStand, A New View on News. Irvine, CA, Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.