GEOSPATIAL PUBLISHING
Creating and Managing Geo-Tagged Knowledge Repositories

Arno Scharl
Know-Center & Graz University of Technology,
Knowledge Management Institute; Inffeldgasse 21a, 8010 Graz, Austria

Keywords: Geospatial Web, Geo-Tagging, Content Production, Knowledge Acquisition.

Abstract: International media have recognized the potential of geo-browsers such as NASA World Wind and Google Earth, for example when Web and television coverage on hurricane "Katrina" used interactive geospatial projections to illustrate its path and the scale of destruction. Yet these early applications only hint at the true potential of geo-browsing technology to build and maintain virtual communities, and to revolutionize the production, distribution and consumption of media products. Investigating this potential, this paper reviews the literature on geospatial publishing with a special focus on extracting geospatial context from unstructured textual resources. A content analysis of online coverage based on a suite of text mining tools then sheds light on the popularity and adoption of geo-browsing platforms. While such platforms might help enrich a company's portfolio of media products, they also pose a threat for existing players through attracting new competitors; e.g., independent providers of geospatial metadata or location-based services.

1 INTRODUCTION

Contrary to early predictions of the Internet rendering geography irrelevant, the discipline is increasingly gaining importance. Geo-browsers facilitate the access to vast quantities of geo-referenced and time-stamped data. Keen competition between well-known software and media companies surrounds the provision of two-dimensional geospatial user interfaces. *Google Maps* (maps.google.com), *MapQuest* (www.mapquest.com), *MS Virtual Earth* (Windows Live Local; local.live.com), *Yahoo Local Maps* (maps.yahoo.com) and other online services are adding new functionality, data sources and interface options in rapid succession. These tools transmit cartographic data and visualize the context and geographic distribution of different types of location-based resources and services.

Three-dimensional geo-browsers combine satellite imagery with aerial photographs and Shuttle Radar Topography Mission (SRTM) elevation data. Using standardized services such as the bitmap-based WMS (Web Mapping Service) or the vector-based WFS (Web Feature Service) of the *Open Geospatial Consortium* (www.opengeospatial.org), image tiles and vector data including geo-positioning information are retrieved from a central server, arranged into a real-time mosaic, and mapped onto a three-dimensional representation of the globe. Altering the field-of-view angle allows to zoom in and out on Earth and increase or decrease the level of detail displayed. Users can seamlessly zoom from NASA Blue Marble data at 1-kilometer-per-pixel, for example, to the detailed mosaic of LandSat 7 data at 15-meters-per-pixel (Hogan & Kim, 2004). Adding the option to tilt the display relative to the spectator's point of view adds a third dimension, altitude. Layers built into the interface provide allow users to effortlessly switch between detailed views and highly aggregated representations.

Most providers of geo-browsing platforms offer Application Programming Interfaces (APIs) or XML scripting to facilitate building third-party online services on top of their platforms (Roush, 2005). Multiple layers of icons, paths and images can be projected via these services – referencing and scaling icons, for example, positioning them on the globe, and linking them to external knowledge repositories, (Web) documents, or photo collections. Latitude and longitude variables determine the symbols' position, while distance above surface values specify whether symbols hover above ground. A good example is the data from NASA's *Moderate Resolution Imaging Spectroradiometer* (MODIS), providing daily updated planetary imagery, documenting natural events such as fires and storms (Hogan & Kim, 2004).

Scharl A. (2006).
GEOSPATIAL PUBLISHING - Creating and Managing Geo-Tagged Knowledge Repositories.
In *Proceedings of the First International Conference on Software and Data Technologies*, pages 226-231
DOI: 10.5220/0001313302260231

Traditionally, the role of geography has been restricted to retrieving information more effectively and enhancing inference operations, but not for specification of queries and the presentation of results. Geo-browsers are about to address this shortcoming by redefining the look and feel of user interfaces, leveraging the knowledge about a user's location to unlock organized indices to the physical world (Kendall, 2005).

2 GEO-TAGGED KNOWLEDGE REPOSITORIES

Concentrated efforts are underway to geo-tag as much existing information as possible. Geo-tagging refers to the process of assigning geospatial context information, from specific point locations to arbitrarily shaped regions. Sources of geospatial context information for annotating Web resources include:

- Annotation by the authors (Daviel & Kaegi, 2003), manually or through location-aware devices such as GPS navigation systems, RFID-tagged products and cellular handsets (Francica, 2005). These devices geo-tag information automatically when it is being created.
- Determining the location of the server – e.g. by querying the *Whois* database for domain registrations, monitoring how Internet traffic is routed, or by analyzing the URL for additional cues (McCurley, 2001).
- Automated annotation of existing documents: The processes of recognizing geographic context and assigning spatial coordinates are commonly referred to as *geo-parsing* and *geo-coding*, respectively (McCurley, 2001).

Once geospatial context information becomes widely available, any point in space will be linked to a universe of commentary on its environmental, historical and cultural context, related community events and activities, as well as personal stories and preferences. With the widespread introduction of commercial applications such as location-based services and geospatial gaming environments, even locative spam will be a common phenomenon (Erle, Gibson, & Walsh, 2005). At present, however, many metadata initiatives still suffer from the chicken and egg problem of wishing that existing content was retrofitted with metadata (McCurley, 2001). Geo-tagging projects are no exception. Addressing this shortcoming, this paper focuses on the third category, the automated parsing and coding of existing resources (online news, for example, and other types of unstructured textual data found on the Web).

2.1 Geo-Parsing

All human artefacts have a location history, which commonly includes a creation location and current location (Spohrer, 1999). Depending on the availability of metadata, geospatial applications can map the whole life cycle of such artefacts. Electronic resources contain the required metadata as explicit or implicit geographic references. This includes references to physical features of the Earth's surface such as forests, lakes, rivers and mountains, and references to objects of the human-made environment such as cities, countries, roads and buildings (Jones, Alani, & Tudhope, 2001). Addresses, postal codes, descriptions of landmarks, and annotated hyperlinks also allow to pinpoint an exact location (Ding, Gravano, & Shivakumar, 2000; McCurley, 2001).

At least 20 percent of Web documents contain easily recognizable and unambiguous geographic identifiers (Delboni, Borges, & Laender, 2005). News articles are particularly rich in such identifiers, since they generally report on the location where an event took place, or where it was reported from (Morimoto, Aono, Houle, & McCurley, 2003) – a distinction also referred to as source versus target geography (Amitay, Har'El, Sivan, & Soffer, 2004). The BBC article "Vienna marking Mozart milestone" (Bell, 2006), for example, has a target geography of EUROPE/AUSTRIA/VIENNA, and a source geography of EUROPE/UNITED KINGDOM/LONDON. In addition to target and source geography, natural language processing also allows extracting the geographic scope (reach) of a Web resource in many cases (Wang, Xie, Wang, Lu, & Ma, 2005).

Identifying and ranking spatial references by semantically analyzing textual data is a subset of the more general problem of *named entity recognition*, which locates and interprets phrasal units such as the names of people, organizations, and places (Cowie & Lehnert, 1996). As with most named entity recognition tasks, false positives are inevitable – e.g., documents that quote addresses unrelated to the their actual content (Morimoto, Aono, Houle, & McCurley, 2003). Ambiguity, synonymy and changes in terminology over time further complicate the geo-parsing of documents (Amitay, Har'El, Sivan, & Soffer, 2004; Kienreich, Granitzer, & Lux, 2006; Larson, 1996). Identical lexical forms often refer to distinct places with the same name (VIENNA referring to the capital of Austria as well as a town in Northern Virginia, US), for example, or can have both geographic and non-geographic meanings – e.g., TURKEY (large gallinaceous bird; bi-continental country between Asia and Europe), MOBILE (capable of moving; city in Alabama, US) and READING (processing written linguistic messages; town in Massachusetts, US). The geo-parsing component

needs to correctly process references to identical or similar places that may be referred to by different names, may be at different levels of the administrative hierarchy, or nearby by some measure of proximity (Jones, Alani, & Tudhope, 2001).

2.2 Geo-Coding

Once a location has been identified, the content fragments can be assigned precise spatial coordinates – latitude, longitude and altitude – by querying a structured geographic index (gazetteer) for matching entries (Hill, Frew, & Zheng, 1999; Tochtermann, Riekert, Wiest, Seggelke, & Mohaupt-Jahr, 1997). Examples of public geographic indices are the *Geographic Names Information System (GNIS),* the *World Gazetteer*, the classifications of the *United Nations Group of Experts on Geographical Names,* the *Getty Thesaurus of Geographic Names,* and the *ISO 3166-1 Country Codes*.

While simple gazetteer lookup clearly benefits from being language-independent, more advanced algorithms consider lexical and structural linguistics clues, as well as contextual knowledge contained in the documents – e.g., dealing with ambiguity by removing stop-words, identifying references to people and organizations (Clough, 2005), and applying contextual rules such as "co-occurring place names indicate nearby locations". For each identified reference, this process assigns a probability *P(name, place)* that a given name refers to a particular place (Amitay, Har'El, Sivan, & Soffer, 2004). The interpretation with the highest probability is then assigned a canonical taxonomy node such as EUROPE/AUSTRIA/VIENNA (48°14' N; 16°20' E).

2.3 Managing Geospatial Context

Metadata frameworks often include geospatial attributes, e.g. the *Dublin Core Metadata Initiative's* "Coverage" tag (McCurley, 2001). The need for controlled vocabularies suggests that ontologies are going to play a key role in managing geospatial context information. While conflicting definitions of "ontology" abound (Guarino, 1997), most agree that the term refers to a designed artefact representing shared conceptualizations within a specific domain.

Geo-ontologies encode geographical terms and their semantic relationships – e.g. containment, overlap and adjacency (Tochtermann, Riekert, Wiest, Seggelke, & Mohaupt-Jahr, 1997). In the case of spatially aware search engines, for example, ontological knowledge supports query term expansion and disambiguation, relevance ranking and Web resource annotation (Abdelmoty, Smart, Jones, Fu, & Finch, 2005). Geo-ontologies can either be represented through generic markup languages such as the Web Ontology Language (OWL) endorsed by the World Wide Web Consortium (Horrocks, Patel-Schneider, & Harmelen, 2003; Smith, Welty, & McGuinness, 2004), or more specific approaches such as the Geographic Markup Language (GML) developed by the Open Geospatial Consortium (Lake, Burggraf, Trninic, & Rae, 2004).

3 GEOSPATIAL PUBLISHING

Technological convergence and the move towards digital media continue to drive today's newsrooms (Pavlik, 1998). While many innovations that gain ground in the media industry are largely invisible to the end user, geo-browsers impact the consumption of news media, change mainstream storytelling conventions, and provide new ways of selecting and filtering news stories.

3.1 Geospatial Literacy

International media have recognized the potential of geospatial interfaces, for example when Web and TV coverage on the hurricane "Katrina" used geo-browsers to illustrate its path and the scale of destruction. Such mainstream coverage is well suited to increase geospatial literacy, which today exists only among a small portion of highly educated people (Erle, Gibson, & Walsh, 2005). Geospatial literacy includes the ability to understand, create, and use spatial information and maps in navigating, in describing phenomena, in problem-solving, and in artistic expression (Liebhold, 2004).

In light of the explosive growth and diminished lifespan of information, geospatial literacy is becoming increasingly important, as the thought that needs to be followed in information discovery tasks is often spatial in nature (McCurley, 2001).

3.2 Content Production

Google's purchase of Keyhole and Microsoft's purchase of GeoTango demonstrate the perceived strategic potential of three-dimensional geographic mapping. Hybrid models of individual and collaborative content production are particularly suited for geo-browsers, which allow to seamlessly integrate and map *individual sources* (monographs, commentaries, blogs), *edited sources* (encyclopedias, conference proceedings, traditional newsrooms), *evolutionary sources* (Wiki applications, open-source project documentations), and *automated sources* (e.g. news aggregators, news summarizers).

Geo-browsing technology not only impacts the production of content, but also its distribution, packaging and consumption. When specifying preferences for personalized news services, for example, geo-browsers are effective tools to pinpoint locations and specify geographic areas to be covered by the news service. Such services require content that is correctly annotated along several dimensions:

- spatial (source and target geography),
- semantic (major topics covered, e.g. assigning terms from a controlled vocabulary),
- temporal (timestamp of the event reported, the initial publication of the article, as well as subsequent revisions).

Online news can be indexed, searched and navigated along these dimensions (McCurley, 2001). The geographical scope of an article, for example, allows filtering and prioritizing content in line with the user's area of interest (often different from his/her actual location).

3.3 Geospatial Media Coverage

Geo-informatics represents an established discipline that has created an industry with remarkable revenues (Wilk, 2005). Yet only with the launch of Google Maps, and its brother in crime, Google Earth, we've seen a dramatic increase in public awareness of the potential of geospatial technology (Francica, 2005). Spurred by space photography, global satellite positioning, mobile phones, adaptive search engines and new ways of annotating Web content, the "ancient art of cartography is now on the cutting edge" (Levy, 2004, 56).

Many current articles are shining a spotlight on geospatial technologies, describing trends in mobile services, investigating the emerging industry of local search, and reporting unidentified or unusual objects found on satellite images. In the past, the process of collecting and analyzing such articles was time consuming, expensive, and often yielded incomplete data. Nowadays articles are readily available online, allowing for inexpensive, fast and topical research.

As traditional media extend their dominant position to the online world, analyzing their Web sites reflects an important portion of Web content that the average Internet user accesses. On a macro-level, analysts gain insights into publicity through incidental news coverage by monitoring information flows within and across media (Scharl, Weichselbraun, & Liu, 2005). On a micro-level, documents retrieved from Web sites contain valuable information about trends and organizational strategies.

This study sampled 129 Web sites in quarterly intervals between May 2005 and January 2006, drawing upon the *Newslink.org*, *Kidon.com* and *ABYZNewsLinks.com* directories to compile a list of international media sites from seven English-speaking countries: United States, United Kingdom, Canada, Australia, South Africa, New Zealand and Ireland. A Web crawler mirrored the Web sites by following their hierarchical structure until reaching 50 megabytes of textual data, a limit that helped reduce the dilution of top-level information by content in lower hierarchical levels (Scharl, 2000). Updates and revisions of news articles often result in multiple versions of the same content (Kutz & Herring, 2005). The system therefore identified and removed redundant segments such as headlines and news summaries, whose appearance on multiple pages would otherwise distort frequency counts.

Media attention was calculated as the relative number of references to a technology or product, measured in occurrences per million tokens. A pattern matching algorithm processed a list of regular expressions, considering common term inflections while excluding ambiguous expressions.

Figure 1 summarizes the number of occurrences identified through these regular expressions. Between Q2/2005 and Q1/2006, coverage on 2D and 3D platforms increased significantly by more than 300 and 1,100 percent, respectively (Wilcoxon Signed Ranks; $p<0.05$). While in Q2/2005, coverage on 2D platforms exceeded coverage on their 3D counterparts (Mann-Whitney; $p<0.05$), Q1/2006 showed a different picture. There was no significant difference between the categories, although 3D platforms took a slight lead with an average relative frequency of exactly one occurrence per million tokens. With 83 percent share of coverage, Google Earth has been the primary driver behind the observable increase in popularity. This represents a remarkable feat with a product only launched in June 2005, not receiving any mentions in Q2/2005. As of January 2006, MapQuest still dominated the 2D category with 46 percent of total coverage, while Google Maps and Google Local were catching up rapidly with a share of 44 percent (in the second quarter of 2005, MapQuest had received nearly twice as many mentions).

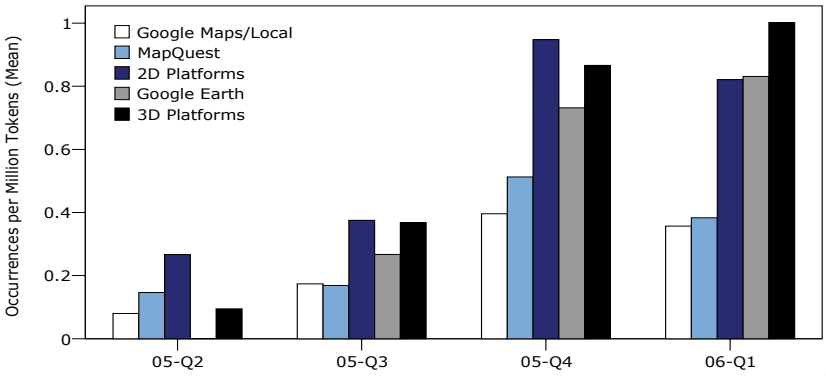

Figure 1: Media Coverage of Geospatial Platforms (Q2/2005 – Q1/2006).

4 CONCLUSION AND OUTLOOK

By integrating traditional cartographic geodata with geo-tagged hypermedia, the *Geospatial Web* "may ultimately be the big disruptive innovation of the coming decade" (Erle, Gibson, & Walsh, 2005, xxv). As such, it will serve as a catalyst of social change and enabler of a broad range of as yet unforeseen applications.

The introduction of geo-browsing platforms such as Google Earth and NASA World Wind has popularized the process of "annotating the Planet" (Udell, 2005). This paper presented the underlying technology, methods to "geo-enable" existing knowledge repositories through parsing and coding geospatial references, and geospatial applications in a media context. A quarterly snapshot of international media coverage revealed the increasing popularity of geospatial products and technologies, particularly as far as three-dimensional platforms are concerned.

Science and technology's accelerated advancement demands constant media innovation, from idea to utility (Stapleton & Hughes, 2006). In this competitive environment, geography is emerging as the fundamental principle for structuring the Web (Roush, 2005), yielding the world's knowledge through the lens of location (Levy, 2004, 58). Geotagging aka adding location metadata to existing databases and using geo-browsing platforms and location-based services to access the vast amounts of information stored in these databases weds physical and virtual spaces, deepening our experiences of these spaces and incorporating them into our everyday lives (Roush, 2005).

ACKNOWLEDGEMENTS

The Know-Center is funded by the Austrian Competence Center program Kplus under the auspices of the Austrian Ministry of Transport, Innovation and Technology (www.ffg.at), and by the State of Styria.

REFERENCES

Abdelmoty, A. I., Smart, P. D., Jones, C. B., Fu, G., & Finch, D. (2005). A Critical Evaluation of Ontology Languages for Geographic Information Retrieval on the Internet. *Journal of Visual Languages and Computing, 16*(4), 331-358.

Amitay, E., Har'El, N., Sivan, R., & Soffer, A. (2004). *Web-a-Where: Geotagging Web Content.* Paper presented at the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK.

Bell, B. (2006). Vienna Marking Mozart Milestone. news.bbc.co.uk/2/hi/entertainment/4654880.stm.

Clough, P. (2005). *Extracting Metadata for Spatially-Aware Information Retrieval on the Internet.* Paper presented at the 2nd International Workshop on Geographic Information Retrieval (GIR-2005), Bremen, Germany.

Cowie, J., & Lehnert, W. (1996). Information Extraction. *Communications of the ACM, 39*(1), 80-91.

Daviel, A., & Kaegi, F. A. (2003). *Geographic Registration of HTML Documents (IETF Internet-Draft, July 2003)*. Sterling: Internet Engineering Task Force.

Delboni, T. M., Borges, K. A. V., & Laender, A. H. F. (2005). *Geographic Web Search based on Positioning Expressions.* Paper presented at the 2nd International Workshop on Geographic Information Retrieval (GIR-2005), Bremen, Germany.

Ding, J., Gravano, L., & Shivakumar, N. (2000). *Computing Geographical Scopes of Web Resources.* Paper

presented at the 26th International Conference on VLDB, Cairo, Egypty.

Erle, S., Gibson, R., & Walsh, J. (2005). *Mapping Hacks - Tips & Tools for Electronic Cartography*. Sebastopol: O'Reilly.

Francica, J. (2005). Struggling for Relevance in the Era of the Google Phenomenon. *Directions Magazine, Dec 1, 2005,* http://www.directionsmag.com/editorials.php?article_id=2035.

Guarino, N. (1997). Understanding, Building and Using Ontologies. *International Journal of Human-Computer Studies, 46*(2-3), 293-310.

Hill, L. L., Frew, J., & Zheng, Q. (1999). Geographic Names - The Implementation of a Gazetteer in a Georeferenced Digital Library. *D-Lib Magazine, 5*(1), http://www.dlib.org/.

Hogan, P., & Kim, R. (2004). *NASA Planetary Visualization Tool.* Paper presented at the American Geophysical Union Fall Meeting, San Francisco, USA.

Horrocks, I., Patel-Schneider, P. F., & Harmelen, F. v. (2003). From SHIQ and RDF to OWL: The Making of a Web Ontology Language. *Journal of Web Semantics, 1*(1), 7-26.

Jones, C. B., Alani, H., & Tudhope, D. (2001). Geographical Information Retrieval with Ontologies of Place. In D. R. Montello (Ed.), *International Conference on Spatial Information Theory: Foundations of Geographic Information Science (= Lecture Notes in Computer Science, vol. 2205)* (pp. 322-335). Berlin: Springer.

Kendall, J. E. (2005). Satellite Mapping and Its Potential in Ecommerce: Why We Need Directions to Follow Our New Maps. *Decision Line, 36*(5), 11-14.

Kienreich, W., Granitzer, M., & Lux, M. (2006). *Geospatial Anchoring of Encyclopedia Articles.* Paper presented at the 10th International Conference on Information Visualisation (iV-06), London, UK.

Kutz, D., & Herring, S. C. (2005). *Micro-longitudinal Analysis of Web News Updates.* Paper presented at the 38th Hawaii International Conference on System Sciences (HICSS-38), Hawaii, USA.

Lake, R., Burggraf, D., Trninic, M., & Rae, L. (2004). *Geography Mark-Up Language: Foundation for the Geo-Web*. New York: John Wiley & Sons.

Larson, R. R. (1996). Geographic Information Retrieval and Spatial Browsing. In L. Smith & M. Gluck (Eds.), *GIS and Libraries: Patrons, Maps and Spatial Information* (pp. 81-124). Urbana-Champaign: University of Illinois.

Levy, S. (2004). Making the Ultimate Map. *Newsweek, 143*(23), 56-58.

Liebhold, M. (2004). *Infrastructure for the New Geography*. Menlo Park: Institute for the Future.

McCurley, K. S. (2001). *Geospatial Mapping and Navigation of the Web.* Paper presented at the 10th International World Wide Web Conference, Hong Kong, China.

Morimoto, Y., Aono, M., Houle, M. E., & McCurley, K. S. (2003). *Extracting Spatial Knowledge from the Web.* Paper presented at the Symposium on Applications and the Internet (SAINT-2003), Orlando, USA.

Pavlik, J. V. (1998). *New Media Technology - Cultural and Commercial Perspectives*. Needham Heights: Allyn & Bacon.

Roush, W. (2005). Killer Maps. *Technology Review, 108*(10), 54-60.

Scharl, A. (2000). *Evolutionary Web Development*. London: Springer. http://webdev.wu-wien.ac.at/.

Scharl, A., Weichselbraun, A., & Liu, W. (2005). *An Ontology-based Architecture for Tracking Information across Interactive Electronic Environments.* Paper presented at the 39th Hawaii International Conference on System Sciences (HICSS-39), Kauai, USA.

Smith, M. K., Welty, C., & McGuinness, D. (2004). Web Ontology Language (OWL) Guide Version 1.0. from http://www.w3.org/TR/2004/REC-owl-guide-20040210/

Spohrer, J. (1999). Information in Places. *IBM Systems Journal, 38*(4), 602-628.

Stapleton, C., & Hughes, C. E. (2006). Believing is Seeing: Cultivating Radical Media Innovations. *IEEE Computer Graphics and Applications, 26*(1), 88-93.

Tochtermann, K., Riekert, W.-F., Wiest, G., Seggelke, J., & Mohaupt-Jahr, B. (1997). *Using Semantic, Geographical, and Temporal Relationships to Enhance Search and Retrieval in Digital Catalogs.* Paper presented at the 1st European Conference on Research and Advanced Technology for Digital Libraries (= Lecture Notes in Computer Science, Vol 1324), Pisa, Italy.

Udell, J. (2005). Annotating the Planet with Google Maps. *InfoWorld, March 04, 2005,* http://www.infoworld.com/article/05/03/04/10OPstrategic_01.html.

Wang, C., Xie, X., Wang, L., Lu, Y., & Ma, W.-Y. (2005). *Detecting Geographic Locations from Web Resources.* Paper presented at the 2nd International Workshop on Geographic Information Retrieval (GIR-2005), Bremen, Germany.

Wilk, C. (2005). Welt in Händen: Arbeiten mit Google Earth und World Wind. *iX - Magazin für professionelle Informationstechnik, 12/05*, 50-62.