

Media Watch on Climate Change: Building and Visualizing Contextualized Information Spaces

Arno Scharl², Albert Weichselbraun², Alexander Hubmann-Haidvogel³,
Hermann Stern⁴, Gerhard Wohlgenannt², and Dmytro Zibold²

¹ Department of New Media Technology
MODUL University Vienna, Austria
scharl@ecoresearch.net

² Research Institute for Computational Methods
Vienna University of Economics and Business Administration, Austria
{albert.weichselbraun, gerhard.wohlgenannt,
dmytro.zibold}@wu-wien.ac.at

³ Institute for Tourism and Leisure Studies
Vienna University of Economics and Business Administration, Austria
alexander.hubmann@wu-wien.ac.at

⁴ Knowledge Management Institute
Graz University of Technology, Austria
hermann.stern@tugraz.at

Abstract. This paper presents the 'Media Watch on Climate Change', an interactive Web portal that combines a portfolio of semantic services with a visual interface based on tightly coupled views. The interface enables users to access a repository of environmental knowledge, built by crawling about 300,000 news media articles in weekly intervals. Tagging services annotate each article to create a contextualized information space. A system component for automatically extending ontologies helps structure the knowledge repository, and feeds an ontology-based visualization module. The Web portal is intended to facilitate the access of complex datasets, help users navigate large collections of Web documents based on similarity, geospatial context and underlying concept, and provide annotations via standardized interfaces for third-party applications.

1 Motivation

The widespread acceptance of three-dimensional virtual globes such as Google Earth and NASA World Wind advances the vision of a *Geospatial Web*. The following hypothetical scenario outlines how an integration of geospatial and semantic data may radically change working environments, impact workflows within and across organizations, and enrich individual interactions.

“Kathryn O’Reilly is a freelance editor who sells her ability to gather, filter and prioritize electronic content. In a virtual world built on contextualized information spaces, Kathryn seamlessly switches between geographic and semantic topologies. She begins her typical working day floating in the virtual space above Earth, ready to navigate the globe and semantic structures via subtle movements of her eyes. From her elevated

position, Kathryn not only observes the rise and decay of topics, but also the unfolding of social structures based on the unique social networks of her friends and contacts. Across these networks she builds and shares her knowledge repository, and composes media products that are continuously being validated and enriched by the latest news feeds and third-party sources” [1, p.3f].

This scenario exemplifies the profound impact of the Geospatial Web on managing individual and organizational knowledge. The Geospatial Web will not only reveal the context and geographic distribution of location-based resources and services, but also catalyze virtual communities by matching people of similar interests, browsing behavior, or geographic location.

2 System Description

The vision of a Geospatial Web promotes the convergence of geographic information, Internet technology and social change. Taking a step towards this vision, the Media Watch on Climate Change uses automated content analysis to extract geospatial context and build a geotagged knowledge base. The interface provides various means to interactively access this knowledge base. It shows that geobrowsers are not only suited to explore geographic features, but can also render other types of imagery such as two-dimensional semantic maps or three-dimensional knowledge planets. Such maps are representations of semantic information spaces based on a landscape metaphor [2].

To increase awareness and the availability of environmental information, the IDIOM Media Watch on Climate Change provides a comprehensive and continuously updated account of media coverage on climate change and related issues. The portal aggregates, filters and visualizes environmental Web content from 150 Anglo-American news media sites. It is available at the following address:

<http://www.ecoresearch.net/iswc2007>

2.1 Annotation and Data Management

The knowledge base underlying the Media Watch on Climate Change is a contextualized information space. If authors and information providers do not provide contextual attributes, *annotation services* attempt to tag and classify new objects. If contextual attributes are provided, the annotation services validate the manual entries, add missing information, and suggest changes if conflicting information is found. The annotation services of the current prototype provide metadata along three dimensions: (i) *spatial* - distinguishing between source and target geography; (ii) *semantic* - assigning the most relevant concepts from a controlled vocabulary; and (iii) *temporal* - adding timestamps for the reported event, the initial publication and subsequent revisions. Each element of the contextualized information space can be organized, indexed, searched and navigated along these dimensions.

2.2 Interface Design and Map Synchronization

Figure 1 shows a screenshot of the current prototype. The content view in the upper left window contains the active document, including its mirror date and source/target geog-

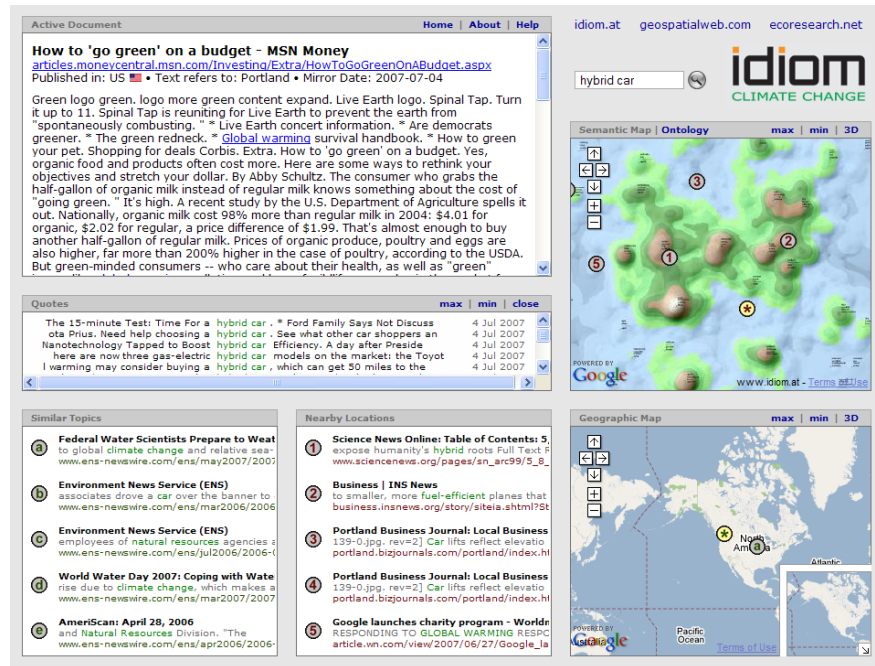


Fig. 1. Screenshot of the IDIOM Media Watch prototype

raphy. Below, just-in-time information retrieval agents list documents referring to similar topics and nearby locations. Clicking on the related references extends the quoted text, clicking on the circular marker on the left activates that particular document. The semantic and geographic maps facilitate access to the underlying knowledge base. The information landscape in the upper right window shows the semantic associations between documents. Peaks indicate clusters of documents about a popular topic, whereas valleys represent sparsely populated parts of the information space. The two links in the upper left corner of this window allow users to switch between the semantic map and a visual representation of an automatically generated domain ontology (see Section 2.5). Below, a geographic map shows the locations referred to in the listed documents. Clicking on the 'max' links increases the size of the maps.

Each of the '3D' links opens a new window. The semantic map links to a description of knowledge planets as a new interface metaphor. The link of the geographic map points towards a KML (Keyhole Markup Language) file that requires a geobrowser to display the 100 most relevant news items on a three-dimensional virtual globe.

Once users enter a search query, an additional window displays quotes including the target term as 'concordances' (centering the target term and showing its immediate context in the various documents). Besides entering query terms, users can click on any position in the maps (not only on the document markers) to retrieve articles related to that particular location, topic, or domain concept. The different views are therefore said to be 'tightly coupled': User actions in one window trigger an immediate update of all other displays.

2.3 Integration of Semantic Markup and Interoperability

Microformats like hCard, hCalendar, XFN and others provide straightforward techniques to embed semantically rich information into existing Web pages. Their markup is not as advanced and powerful as RDF, but aggregator services like Technorati have demonstrated how useful these tagging formats content have become. Within fifteen months, from November 2005 to April 2006, the number of tagged postings grew from zero to 100 million⁵.

Web2.0 sites like Dopplr.com⁶ or CouchSurfing.com⁷ started adapting microformats for building social networks using XFN and the hCard standard. “IBM is another early adaptor, enriching employee listings with hCards information. Unfolding the power of a decentralized and easy tagging architecture, Web developers and domain experts specify new domain-specific microformats in order to provide users with a steadily growing vocabulary for annotating Web content. Efforts like the W3C’s development of RDFa try to combine the advantages of micro-formats with the expressiveness of RDF.

‘Geo’ microformats describe the target geography of the IDIOM Media Watch’s documents by encoding geographical information. ‘Rel-tag’ markup extends existing hyperlinks, marking the destination as a system-designated tag. Data from the just-in-time information retrieval agents is annotated using the ‘hReview’ microformat, marking it as a review and thus giving it additional information such as a relevance rating and a review date. Browser add-ons such as the “Operator” and “Tails Export” Firefox extensions provide a user-friendly interface for importing such entries into existing applications. Geo-metadata from the IDIOM Media Watch prototype can be presented as a link to mapping applications such as Yahoo! or Google Maps, ‘rel-tag’ markup refers the user to community sites such as del.icio.us⁸ and flickr⁹. RSS feeds provide annotated document feeds to topics and keywords promoting interoperability with third party applications. The feeds are available at the following address:

`http://www.ecoresearch.net/iswc2007/tags/{keyword|topic}`

2.4 Knowledge Planets

Generated by orthographically projecting and tiling semantic maps (= visual representations of semantic information spaces based on a landscape metaphor), knowledge planets allow visualizing massive amounts of textual data. At the time of map generation, the planet’s topology is determined by the content of the knowledge base. The peaks of the virtual landscape indicate abundant coverage on a particular topic, whereas valleys represent sparsely populated parts of the information space.

VisIslands, a thematic mapping algorithm similar to SPIRE’s Themescape [3] and its commercial successor Cartia/Aureka, supports dynamic document clustering [4, 5]. Initially, the document set is pre-clustered using hierarchical agglomerative clustering

⁵ http://www.digital-web.com/articles/the_big_picture_on_microformats/

⁶ <http://dopplr.com/>

⁷ <http://www.couchsurfing.com/>

⁸ <http://del.icio.us/>

⁹ <http://www.flickr.com/>

[6], randomly distributing the cluster centroids in the viewing rectangle. The documents belonging to each cluster, as determined by the pre-clustering, are then placed in circles around each centroid. The arrangement is fine-tuned using a linear iteration force-directed placement algorithm adapted from Chalmers [7]. The result resembles a contour map of islands. Fortunately, algorithms based on force models easily generalize to the knowledge planets' spherical geometries. The IDIOM research project (www.idiom.at) extends and refines the VisIslands thematic mapping component to improve throughput and scalability, generate layered thematic maps, and provide a Web Map Service (WMS) that serves these maps as image tiles for various geobrowsing platforms.

2.5 Ontology Extension

The need for controlled vocabularies and shared meaning suggests that ontologies - shared conceptualizations within a specific domain [8] - are going to play a key role in managing context information. Geo-ontologies encode geographical terms and semantic relationships such as containment, overlap and adjacency. Domain ontologies are used to organize topics and to navigate through knowledge repositories. Spatially aware search engines use ontological knowledge for query term expansion and disambiguation, relevance ranking and Web resource annotation [9]. This project does not intend to generate a universally accepted domain ontology, but to accurately represent knowledge contained within a specific corpus. Changes in sample composition or trends in media coverage inevitably affect the generated ontology. This represents a promising avenue for comparative studies and future research on ontology evolution.

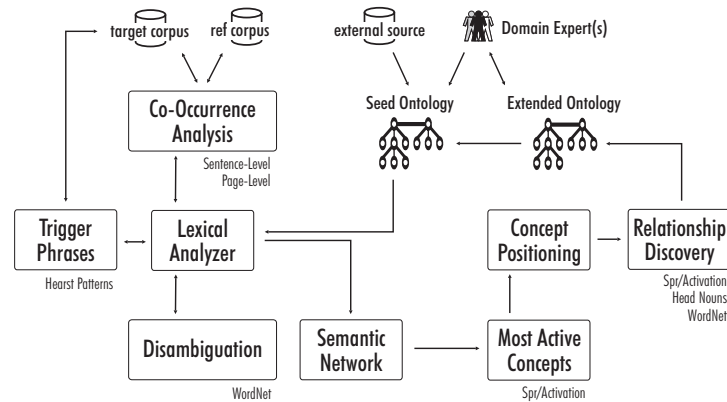


Fig. 2. Ontology Extension System Architecture

Due to the sheer volume of data, the use of ontologies in large information spaces is difficult without an ontology extension system that suggests concepts and relations automatically. Figure 2 presents a conceptual view of the ontology extension architecture underlying the Media Watch on Climate Change.

A small set of terms obtained from domain experts or known ontology repositories is first selected as a seed ontology. The seed ontology terms are then fed into the analyzer, which is distributing the input to different plugins that provide evidence sources, suggesting promising new concepts. Currently three evidence sources are considered:

1. *Co-occurrence analysis* [10] at both the sentence and the document level uses a threshold based on the co-occurrence significance to select 20 terms on the sentence and 20 terms on the document level.
2. *Trigger Phrases* [11] matching a fragment of text that indicates a particular relation between terms (e.g. parent-child relation).
3. The *WordNet* lexical dictionary [12], which is used as additional evidence source after disambiguating the seed ontology's concepts using a vector space model.

The generated terms are then connected with a seed ontology via directed weighted links. Once the network is established, a spreading activation algorithm identifies the most relevant terms and suggests their incorporation into the seed ontology as described in Wei et al. [13]. Then the following heuristic is used to determine their relation to the concepts in the seed ontology: (i) WordNet, head noun analysis and additional rounds of spreading activation determine the new concepts' position within the ontology, (ii) Subsumption analysis together with WordNet and head noun analysis identify the semantic relation's type. For terms not confirmed automatically, domain experts are consulted. Optionally another iteration of spreading activation over newly acquired terms is triggered to integrate additional concepts into the ontology. In Figure 3, the concepts' shading differentiates the seed ontology terms (= lighter shading) from those added in several iterations of automated ontology learning.

2.6 Ontology Visualization

Ontologies retrieved from the ontology extension process are serialized in OWL and then fed into the visualization framework. Transforming relations annotated using reification and all other information encoded in the serialized version of the ontology into the graphviz¹⁰ description language yields an SVG-file with a graphical representation of the ontology. The framework extracts the concepts' positions from the SVG files, which provides the mapping required for interactions with the ontology view.

Applying the iterative ontology extension process to 50 Web sites mirrored in July 2007 and comprising approximately 80,000 documents yielded the Web portal's domain ontology (see Figure 3). Arrows indicate confirmed hierarchical relations. The dotted lines connect semantically related terms whose exact type of relation could not be determined automatically. For these non-hierarchical relations, the (*r*) values indicate their strength based on the link assignment's spreading activation level. High values suggest a strong relation between the concepts, a value of 8 being the maximum due to the specific setup of the spreading activation network.

The current navigational aid contains unlabeled associations, only indicating the strength of the relation but not its specific type. The IDIOM research project is currently extending the underlying algorithm to detect a much broader set of relation types

¹⁰ <http://www.graphviz.org/>

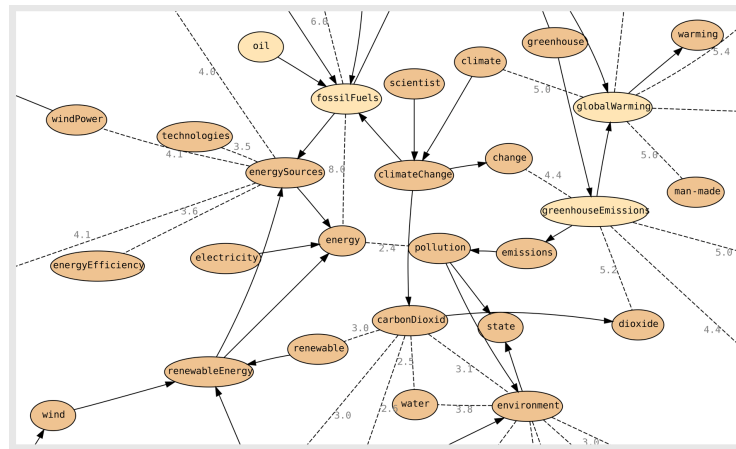


Fig. 3. The extended ontology used in the IDIOM Media Watch prototype

3 Conclusions

Current applications only hint at the true potential of geospatial technology to build and maintain virtual communities and to revolutionize the production, distribution and consumption of media products. Automatically annotating content from heterogeneous sources creates knowledge repositories spanning multiple dimensions such as space, time, and semantics.

Virtual globes improve the accessibility and transparency of such complex repositories. Their introduction has popularized the process of “annotating the Planet” [14]. This process yields *geospatially* referenced information, enabling virtual globes to map annotated content units from various sources. But geospatial interfaces can also serve as a generic image rendering engine to project other types of imagery.

The Media Watch on Climate Change demonstrates how geospatial interfaces can be used to visualize layered representations of thematic maps and domain ontologies as a frontend for semantic services.

Acknowledgment

The Media Watch on Climate Change has been developed as part of IDIOM (Information Diffusion across Interactive Online Media; www.idiom.at), a research project funded by the Austrian Ministry of Transport, Innovation & Technology (BMVIT) and the Austrian Research Promotion Agency (FFG) within the strategic objective FIT-IT (www.fit-it.at).

References

1. Scharl, A.: Towards the geospatial web: Media platforms for managing geotagged knowledge repositories. In Scharl, A., Tochtermann, K., eds.: *The Geospatial Web - How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society*. Springer, London (2007) 3–14
2. Chalmers, M.: Using a landscape metaphor to represent a corpus of documents. In Frank, A.U., Campari, I., eds.: *Spatial Information Theory: A Theoretical Basis for GIS (Lecture Notes in Computer Science, Vol 716)*. Springer, Berlin (1993) 377–390
3. Wise, J.A.: The ecological approach to text visualization. *Journal of the American Society for Science* **50**(9) (1999) 814–835
4. Andrews, K., Guetl, C., Moser, J., Sabol, V., Lackner, W.: Search result visualisation with xfind. In Kapetanios, E., Hinterberger, H., eds.: *Second International Workshop on User Interfaces to Data Intensive Systems (UIDIS 2001)*, Zurich, Switzerland, IEEE Press (2001) 50–58
5. Sabol, V., Kienreich, W., Granitzer, M., Becker, J., Tochtermann, K., Andrews, K.: Applications of a lightweight, web-based retrieval, clustering, and visualisation framework. In Karagiannis, D., Reimer, U., eds.: *4th International Conference on Practical Aspects of Knowledge Management (Lecture Notes in Computer Science, Vol 2569)*. Springer, Berlin (2002) 359–368
6. Jain, Anil, K., Murty, M.N., Flynn, P.J.: Data clustering: A review. *ACM Computing Surveys* **31**(3) (1999) 264–323
7. Chalmers, M.: A linear iteration time layout algorithm for visualising high-dimensional data. In: *7th Conference on Visualization*, San Francisco, USA, IEEE Computer Society (1996) 127–132
8. Gahleitner, E., Behrendt, W., Palkoska, J., Weippl, E.: Knowledge sharing and reuse: On cooperatively creating dynamic ontologies. In: *16th ACM Conference on Hypertext and Hypermedia (Hypertext-2005)*, Salzburg, Austria, ACM Press (2005)
9. Abdelmoty, A.I., Smart, P.D., Jones, C.B., Fu, G., Finch, D.: A critical evaluation of ontology languages for geographic information retrieval on the internet. *Journal of Visual Languages and Computing* **16**(4) (2005) 331–358
10. Roussinov, D., Zhao, J.L.: Automatic discovery of similarity relationships through web mining. *Decision Support Systems* **35** (2003) 149–166
11. Joho, H., Sanderson, M., Beaulieu, M.: A study of user interaction with a concept-based interactive query expansion support tool. In: *Advances in Information Retrieval, 26th European Conference on Information Retrieval*. (2004) 42–56
12. Fellbaum, C.: Wordnet an electronic lexical database. *Computational Linguistics* **25**(2) (1998) 292–296
13. Liu, W., Weichselbraun, A., Scharl, A., Chang, E.: Semi-automatic ontology extension using spreading activation. *Journal of Universal Knowledge Management* **0**(1) (2005) 50–58
14. Udell, J.: Annotating the planet with google maps. *InfoWorld* **March 04** (2005) www.infoworld.com/article/05/03/04/10OPstrategic_1.html.