EnviroInfo 2005 (Brno)
Informatics for Environmental Protection - Networking Environmental Information
 ISBN: 80-210-3780-6

Applications of Environmental Semantics

Maria Rüther[1], Thomas Bandholtz[2]

Abstract

This contribution covers a summary of the application experiences with semantic indexing and search services in the German Environmental Information Network (gein®) since 2000 as well as an outlook on the areas of further development an extended fields of application.

1. Introduction

Since 2000, the German Environmental Information Network (gein®)[3] has provided fully-automated, semantic indexing of documents in the Internet (Rüther 2000). In the following years an improved method has been implemented in the Semantic Network Service (SNS)[4] (Angrick 2002). SNS provides a controlled vocabulary organized in a "semantic network" based on interlinked thesaurus, gazetteer, and time table ("environmental chronicle") structures (see Figure 1). A complex text analysis algorithm determines the most significant keyword, location and time references of each document.

Access to the vocabulary and related methods is given by Web Services. The idea was to provide common vocabularies and semantics to the environmental community from a single authority (the Federal Environmental Agency) to support information sharing. As a first step gein® has dropped its build-in features in 2003 and relies on SNS Web Services since then.

In the following years the number of applications has been increasing slowly but steadily. Two portals of the Spatial Data Infrastructure, GeoPortal.Bund®[5] and GeoMIS.Thüringen[6] have integrated SNS Web Services in 2004. The Environmental Data Catalogue (UDK)[7] and the Environmental Portal of the State of Baden-Württemberg are going to follow in 2005/6.

2. Semantic Search and Indexing

gein® supports the semantic approach in parallel with conventional full text methods separately, so both can be compared. Unlikely the full text approach, semantic indexing does not register the complete set of character strings occurring in the document, but the most significant *topics*. The gein® index contains *identifiers* of these topics, so it is independent from the specific wording style, and even from the original language of the document.

[1] Federal Environmental Agency (UBA), Dessau, Germany. email: maria.ruether@uba.de, Internet: www.umweltbundesamt.de
[2] Consultant, Bonn, Germany. email: thomas@bandholtz.info, Internet: www.bandholtz.info
[3] www.gein.de
[4] www.semantic-network.de
[5] www.geoportal.bund.de
[6] www.geoportal-th.de
[7] www.umweltdatenkatalog.de

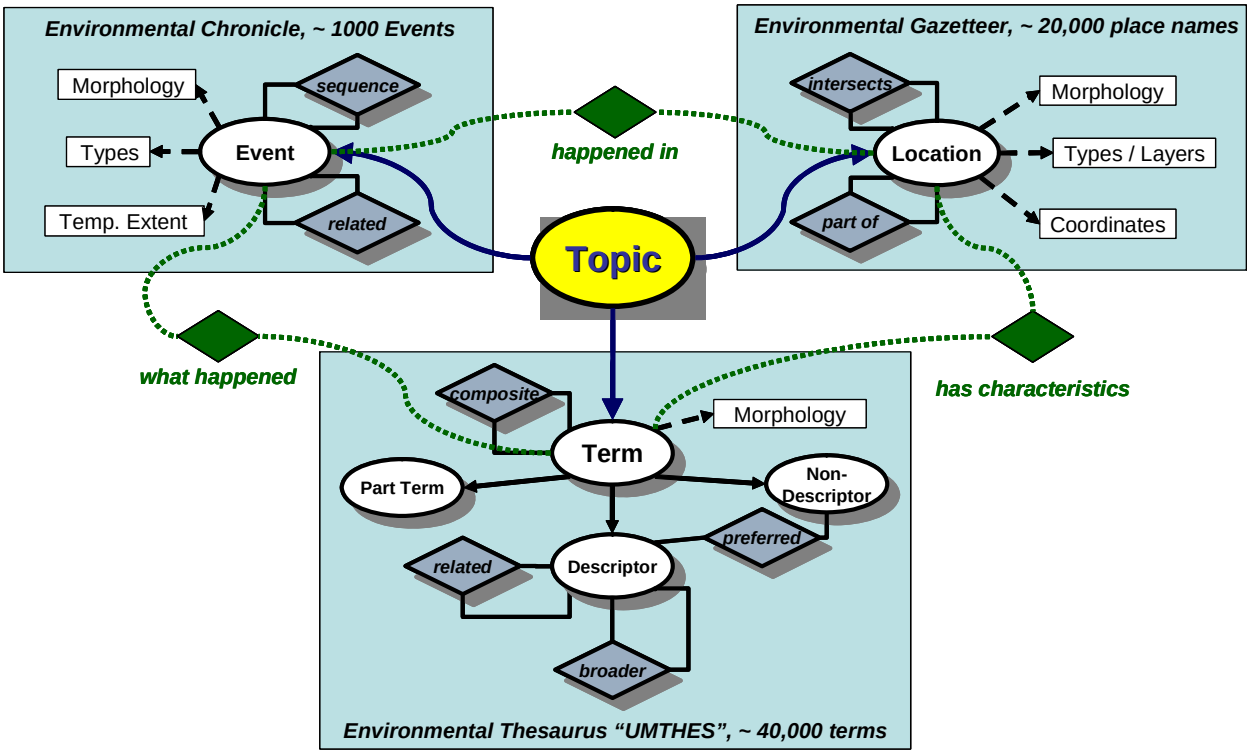

Fig. 1: Structure of the SNS Semantic Network

2.1 Search Method

This needs to be mapped by an appropriate search approach. Searching is *not* performed by exact mapping of certain character sequences to documents any more, but expands to mapping of meaning. For this reason, the initial search terms have to be processed before they are used. In general, they have to be mapped to the topics used in the index. Following the statistics, many of the initial search terms are misleading in one of following ways:

1. Terms are to general ("water", or even "environment") resulting in thousands of unspecific results. SNS services are used to propose more specific terms in this case.
2. Terms are taken from popular discussion ("forest dieback") while the official term may be less emotional ("forest damage"), or there are various terms in wide use as well ("decline", "disease", " deterioration"), so the results will be poor. SNS services are used to propose the *preferred term.*
3. Term spelling points to ambiguous meaning (*homographs*). This occurs very frequently when place names are considered as well. In this case, full text search results will hit the different meanings at random. SNS services are used to clarify the multiple meanings.

Doing so, the search condition is processed very similar to the indexed documents. Each term is not taken as a simple sequence of characters, but as a meaningful term. Multiple terms in the condition are proc-

essed like small documents. Composite terms can be understood in their composite meaning. Finally, complete sentences or paragraphs taken from articles can be used as a search condition.

In general, these methods improve the correspondence between the intention of the searcher and the thematic focus of the documents in the result list. However, the user has to become familiar with these methods. The gein® statistics indicate that the semantic search method is used less frequently than the simple full text method. This may be an issue of *application ergonomics*: while users are familiar with full text search engines such as Google, some of them do not understand the semantic features intuitively. Once they have entered some terms, they are expecting resulting documents immediately, but the semantic search feature responds with vocabulary propositions first (as exemplified above). In the current version the vocabulary is intentionally exposed to the user to allow some negotiation about the terms to be used, and to support a better understanding of the professional terminology. This turned out as a kind of barrier which needs to be addressed by a smooth and self-explaining user interface to be implemented in a more automated, "invisible" way in the future.

As a consequence of these experiences the next generation of gein® (Vögele 2005) is going to support the best of both worlds. The future version will continue with full text and semantic indexing, but there will be a closer integration of both search features.

Independent from ergonomics, there are two premises of *functional* quality:

1. The *vocabulary* needs to be maintained continuously to be up-to-date with the evolving terminology in practical use.
2. The *indexing algorithm* must be able to deal with the multitude of content types and writing styles found in the given corpus to allow the detection of the most significant topics covered in the document.

2.2 Maintenance of the Vocabulary

Maintenance of the *thesaurus* is assured by an experienced editorial group at the Federal Environmental Agency on an almost daily basis. There are statistics about "unknown" character strings in the corpus or the search conditions, but adding new terms to the semantic network is left to human intellectual filtering. Fully automated methods may serve for an initial setup, but the thesaurus of the Agency has been developed and maintained for almost two decades now, so this is a mature basis superior to automatic results.

Maintenance of the *geographic names* is based on the work of the Federal Agency for Cartography and Geodesy[8] (BKG) in the context of the Spatial Data Infrastructure (SDI). This basis has been extended by several environmental location types by the UBA. Till today there is no "official" national Gazetteer Service in Germany, and the maintenance of geographic names is still subject of discussion between the two agencies. We expect a Gazetteer Service to be established and maintained by the BKG in the near future. In the European context the EuroGeoNames[9] project (Sievers 2005) has been started last year and may lead to some additional influence, but it is too early to speculate about specific results.

The *Chronicle* is continuously maintained by the SNS project itself.

2.3 Indexing Methods

Automatic semantic indexing of unstructured data still remains a challenge. Especially when geographic place names are included, the ambiguity of single terms cannot be resolved with simple, straight-forward methods.

[8] www.bkg.bund.de

[9] http://www.eurogeographics.org/eng/03_projects_EuroGeoNames.asp

The original gein® ("GEIN 2000") had implemented two simple methods for the thesaurus and the place names side-by-side, based on traditional librarian methods. The results allowed a general acceptance of the automatic indexing approach, but they showed several weaknesses.

SNS developed a second generation of the indexing method 2002-2003 taking advantage of the integrated vocabulary in the semantic network. One of the major improvements related to resolving the ambiguity of homographs by context analysis.

Currently the project is on the way to improve the method by integration of advanced methods in cooperation with the Fraunhofer Institute for Media Communication (IMK)[10].

3. Publishing and Referencing Semantic Identifiers

While automatic semantic indexing of documents is a rather complex application, publishing semantic identifiers appears to be comparatively simple. From a technical point of view, this comes close to an online glossary, while the major purpose is providing a persistent Internet address (URL) for each term. Like always, things are getting complicated when they exceed the context of a single application.

The intention is that these URLs should be used as external identifiers from within data and metadata sets instead of re-typing a keyword term as an attribute value. The unique URL ensures a well-defined meaning. While many terms may have different meanings in different contexts, the URL of the semantic reference implies a single context (which is the context of the publishing authority).

This allows "different data sources committing to the same ontology for shared meaning" (OWL 2002). The World Wide Web architecture ensures that URL will be globally unique, while persistency needs to be maintained by the publishers. The same vision has been promoted by the "Published Subjects" Technical Committee at OASIS since 2001 in order to "identify subjects of discourse" unambiguously (OASIS 2003). Semantic referencing is also recommended by several important standards such as ISO 19139 (the XML serialization of OGC Catalog Services), by the Universal Description, Discovery and Integration (UDDI) concept of "value sets", or by the "Registry of Observables" in the OGC Sensor Web, and of course by the W3C Semantic Web.

gein® has been aware of the Semantic Web research right from the beginning (Bandholtz 1999), but these developments have not been ready for production at that time. SNS finally is publishing semantic identifiers (URL) for any of its topics since 2003. However, applications correspond to this feature very hesitatingly. The gein® index contains SNS topic IDs consequently, but all the other applications mentioned still rely on names in text fields. (UDK saves identifiers of administrative areas only). Discussions show that there is no common understanding about how URL references could be processed by applications in this case.

To solve this issue, providers of terminology references need to agree to a common pattern of a machine readable terminology presentation behind URL identifiers. This is the basic requirement of applications that want to traverse the boarders between their own data and the terminology services. From today's point of view, the Web Ontology Language (OWL) might be the answer.

In the environmental domain, the Ecoinformatics Initiative[11] (organized by UNEP, USEPA, EEA, GBIF, JRC, NBII, USGS) has established a thesaurus and terminology activity ("Ecoterm") in 2004 to "bring together the major providers of terminologies to discuss how new technologies are being applied and how these valuable resources can be integrated using the Web as a platform for sharing". Looking for a common interface pattern, the Ecoterm community has acknowledged the work of the W3C Semantic Web Best Practices and Deployment Working Group (SWBPD). One of the first results is the SKOS Core Vocabulary Specification as a common basis for "the RDF description of language-oriented knowledge

[10] www.imk.fraunhofer.de

[11] http://ecoinfo.eionet.eu.int/

organisation systems such as thesauri, glossaries, controlled vocabularies, taxonomies and classification schemes" (SKOS 2005). The second meeting of Ecoterm in 2005 has demonstrated a growing acceptance of SKOS and OWL representations by the terminology providers. Developments still are in the state of drafts and prototyping, but the awareness has reached a critical mass, and solutions will follow.

4. Summary

Semantic indexing and search methods of SNS have proved as a valuable alternative to common full text methods in the German Environmental Information Network (gein®). These methods generally improve the correspondence between the intention of the searcher and the thematic focus of the documents in the result list. However, there have been some weaknesses in the earlier state of development.

Starting with 2000, there have been several enhancements that need to be carried on continuously. Co-operative development is needed to reach a new level of quality and user acceptance. The main issues are:

- maintenance of the vocabulary and the semantic network in SNS,
- improvement of the text analysis and indexing algorithm in SNS,
- ergonomic design of the user assistance navigational model in the applications.

Several further applications have started to integrate SNS services or are currently considering so. Compared to search and indexing, the underlying principle of semantic references in data and metadata has been less successful on the application side. However, this is more and more propagated in common information standards of the Spatial Data Infrastructure and the Semantic Web. In the environmental community, the international Ecoterm working group is fostering this development.

Bibliography

Angrick, M. (et al) (2002): Semantic Network Services – Sorting Signal from Noise. In: Environmental Communication in the Information Society. Proceedings of the 16th Conference Environmental Informatics", Vienna, Austria, 2002

Bandholtz, T. (1999): GEIN 2000 and beyond: Information about the environment in the „semantic web“. In: Environmental Markup Language (EML). Proceedings of Workshop 1, Berlin 1999.

Gruber, T. R. (1993) Toward principles for the design of ontologies used for knowledge sharing. Padua workshop on Formal Ontology, March 1993.

OASIS (2003): Published Subjects: Introduction and Basic Requirements. OASIS Published Subjects Technical Committee Recommendation, 2003-06-24. http://www.oasis-open.org/committees/download.php/3050/pubsubj-pt1-1.02-cs.pdf

OWL (2002) Requirements for a Web Ontology Language. W3C Working Draft 08 July 2002. http://www.w3.org/TR/webont-req/

Rüther, M.: (et al) (2000): The German Environmental Information Network (GEIN). Environmental Informatics, Bonn 2000. http://enviroinfo.isep.at/Umweltinformatik2000.htm

Sievers, J. (2005): EuroGeoNames (EGN) - Integration of geographical names data into the European SDI. Joint Workshop "NMCA's and the Internet II". Electronic Delivery and Feature Serving. Frankfurt (DE) 23-25 February 2005

SKOS (2005): SKOS Core Guide. W3C Working Draft 10 May 2005. http://www.w3.org/TR/swbp-skos-core-guide

Vögele (2005): A New and Flexible Architecture for the German Environmental Information Network. EnviroInfo 2005, Brno (CZ)