Environmental Research Programme
Of the Federal Ministry for Environment,
Nature Conservation and Nuclear Safety

Fundamentals of Environmental Policy, Environmental Strategies

Final Report 201 11 612

# Implementation of a Semantic Network Service (SNS) in the context of the German Environmental Information Network (*gein*®)

Summary

Thomas Bandholtz
SchlumbergerSema GmbH, Köln

ON BEHALF OF THE
UMWELTBUNDESAMT
(Federal Environmental Agency)
May 2003

Content

# 1 Introduction

In Germany, since 2000 the German Environmental Information Network (*gein*®, http://www.gein.de) has implemented a common content classification system as a first step of any further content-related integration of the different Internet information sources. Following this encouraging experience, the Research & Development project "Implementation of a Semantic Network Service (SNS) in the context of the German Environmental Information Network (*gein*®)" has been launched in 2001 to overcome some restrictions of the initial version of ontology management and automatic indexing by improvements such as:

- semantic integration of thesaurus, gazetteer, and chronology;
- resolving of homonym ambiguities by context analysis
- elaborated criteria for keyword ranking according to their significance in one document.
- distributed access to semantic methods by web services.

# 2 Integration of the gein® Taxonomy

The *gein*® Taxonomy has been developed since 1999 integrating and extending the major sources of the environmental domain.

*gein*® uses three semantic structures:

1. a thesaurus of currently 39,143 environmental terms (UmThes®),
2. a gazetteer including the intersections between 48,213 geographical objects of all kinds,
3. a chronology – the synopsis of historical and contemporary events that affected the environment.

UmThes® is a full-fledged Thesaurus supporting all the associations required by ISO 2788/5964 (Broader/Narrower; Synonym; Related; Component), and it contains the full word morphology. It is also used by several German-speaking authorities such as the German and Austrian Environmental Data Catalogue, and it is the German source of the GEneral Multilingual Environmental Thesaurus (GEMET, Batschi).

The gein® Gazetteer is based on the GN250 (by Federal Agency for Cartography and Geodesy), but it adds several layers relevant for the environment, and it contains all the spatial intersections as explicit relations in the data, ready-to-use in a rapid query.

The gein® Chronology has been started by GEIN from scratch, and it intends to become some kind of "temporal Thesaurus" of important environmental events – not to be misunderstood as a pin board for conference announcements.

## *2.1 Topic Maps*

All the three threads have been integrated into one semantic model by a Topic Map to build one ontology. So the term "Topic" refers to any of the ontology nodes regardless whether they have descended from the Thesaurus, the Gazetteer, or the Chronology. From the metadata point of view, "Topic" may be understood as equivalent for "keyword".

The Topic Map model does not only allow to integrate such different structures. Further on, a semantic net has been established by setting up cross-associations between thesaurus descriptors, events, and locations.

Topic Maps (ISO 13250) have proved to be the structure of choice whenever different thesauri, gazetteers, classifications, or even simple keyword lists have to be networked. Topic Maps are a rather new standard in the world of taxonomy and ontology, and they have emerged to become one of the most discussed contributions. At XML Europe 2002 (XMLe), there have been 10 contributions dedicated to Topic Maps, and the Knowledge Technology track has been dominated by this movement. Topic Maps have their roots in SGML, and have been standardized in parallel to and in permanent discussion with the Semantic Web and RDF process (Bandholtz, 1999). Today it looks like these two threads are going to find their deserved integration.

Basically, Topic Maps consist of three components: Topics, Associations, Occurrences.

Topics represent any kind of subject:

"In the most generic sense, a subject is anything whatsoever, regardless of whether it exists or has any other specific characteristics, about which anything whatsoever may be asserted by any means whatsoever." (ISO 13250)

So, a Topic may be a Thesaurus descriptor or synonym, a geographic object of a Gazetteer, an Event, a Person, an Organization, whatsoever. Distinct kinds of Topics are defined as Topic Types in a Topic Map application.

Associations may interconnect Topics in some kind of semantic relation between Descriptors and Synonyms, Events and Locations, Descriptors and Events, whatsoever. Distinct kinds of Associations, bound to certain Topic Types as their members, are defined as Association Templates in a Topic Map application.

Occurrences are pieces of information that contribute to the definition of the Topic. Generally, an Occurrence may be seen as any kind of existing information about a Topic, but, as Occurrences are "groupings of addressable information objects around topics" (ISO 13250), this should not be misunderstood to be the general information index of a "corpora" like *gein*®. In SNS, the document index is separated from the Topic Map. Topics are used as classification properties in the document metadata, which rather means: "groupings of topics around addressable information objects" (Bandholtz, 2002).

The XML model of the SNS Topic Map is based on ISO/IEC 13250, but it does not use the (XTM) 1.0 DTD. XTM has been designed to be an interchange format, not a processing format (Bandholtz, 2002). We preferred to develop an XML Schema for Topic Maps after having reviewed the basic ISO SGML DTD, the XTM XML DTD, and Martin Bryan's early XML Schema proposal (Bryan).

### *2.2 The SNS Topic Map Model*

SNS has defined its own Topic types and Association templates to model the three components of the *gein*® taxonomy, as shown in Figure 1. The *Thesaurus* type and it's sub-types reproduce the classical thesaurus structure as defined in ISO 2788/5964. The *Location* type is the abstract parent of all the more specific location types such as cities, catchment areas, or national parks (not shown in the figure). Likewise, the *Event* type is parent of conferences, disasters, and so on. So far, the legacy taxonomy has been transferred into a common topic map without any semantic modification.

Beyond this, the three components have been interlinked by two new association types labeled *where*, and *what*. Both of the use *Event* as the integration point. The *Where*-association links between *Event* and *Location*, pointing out where an event has happened. The What-association links between *Event* and *Descriptor* to describe which topics have been touched by the event.

*Thesaurus*, *Location*, *Event* are abstract master types representing the three dimensions "topic-area-time". Being defined as abstract, these three cannot be used as types for real topics. But they can be used in an association template to express that any subtype of this

abstract type may be used in the association. In Figure 1 this is used e.g. for the association named “what?” to state that any *Event* can be linked to a *Descriptor* (but not to a *Synonym*).

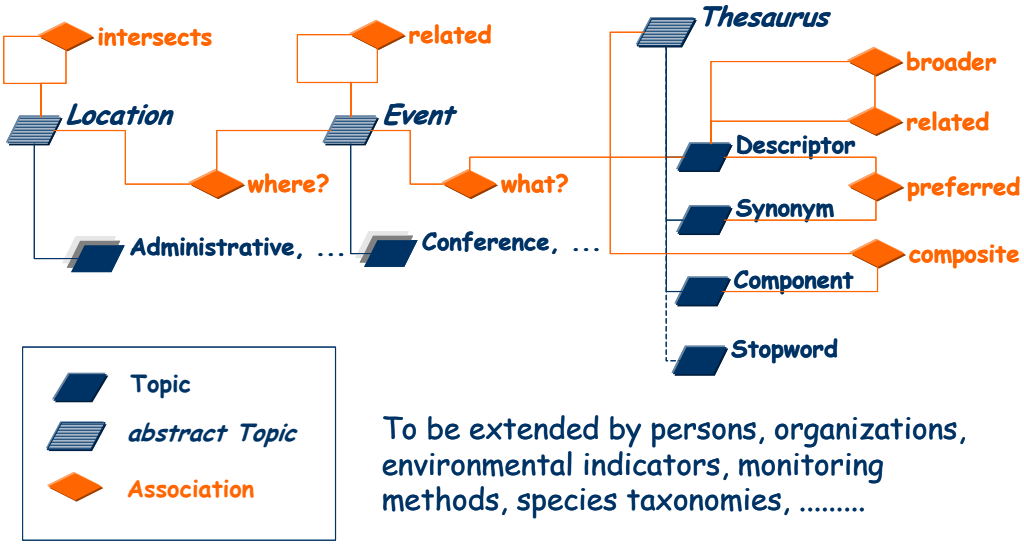


Figure 1 The SNS Topic Map Typology

This structure is intended to be extensible. There had been discussions about adding types like “persons”, organizations, environmental indicators, monitoring methods, or species taxonomies, but this could not be realized in the limited scope of the project.

Anyway, the implemented typology of SNS can be understood as a core typology that may be extended by types like these, and by additional associations such as organizations associated with events, and similar, without any change to the given topic map itself.

# 3 Functional Services

The informational model described so far has been implemented on top of a generic Topic map Engine provided by SchlumbergerSema.

The SNS project has established what we call an “eXtensible Topic Map Engine Architecture” (see Figure 2). This architecture is based on XML formats for requests and resulting Topic Map fragments. It provides basic services that may be utilized by processes that are out of scope of the SNS project itself. The major use case has been *gein*®, and there has been a productive integration with *gein*® in the end of the project. However, this is not the final result. It is intended that these services can be crucial for future applications of the semantic web even if these processes cannot be determined today.

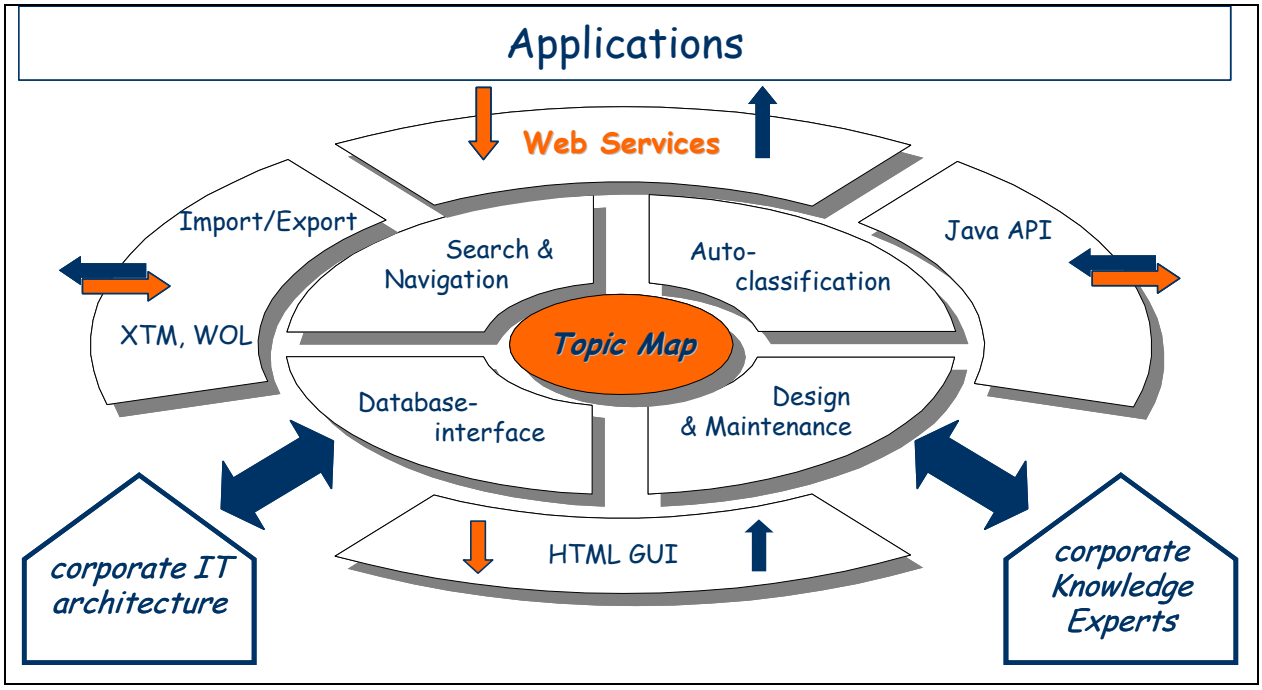


Figure 2: eXtensible Topic Map Engine Architecture

As part of the SNS project, the following 4 functional services have been implemented so far:

- Topic Map Browser and Navigator
- Editorial Staff Services
- Enhanced Automatic Indexing
- Semantic Web Services

### *3.1 Topic Map Browser and Navigator*

Though this was not the main target, SNS includes a stand-alone user interface independent from any application. This was built to provide search facilities, basic visualization, and interactive navigation in the map.

There are HTML displays for Topics and Associations. There are HTML forms to specify search conditions, and there are browsing features for result sets and alphabetic lists. Each Topic and Association appears as a Hyperlink in this interface, so the user can navigate rather intuitively by following the Associations between Topics.

All this requires a simple Web Browser and Internet connectivity at the client end, no additional Software or “Plug-In”.

To visualize the semantic network more adequately, we added some kind of a pilot of *interactive graphics* (see Figure 3), a more intuitive display of nodes and arcs. Each node represents a Topic. By clicking on it, you can display all the Topic’s properties, and you can make it the new center of the whole picture, and so navigate the Topic Map right in a graphic display.

These interactive graphics are implemented in Scalable Vector Graphics (SVG), and so they require a Plug-In which is offered for free by the major players of the graphical industry, such as Adobe, Corel, etc..

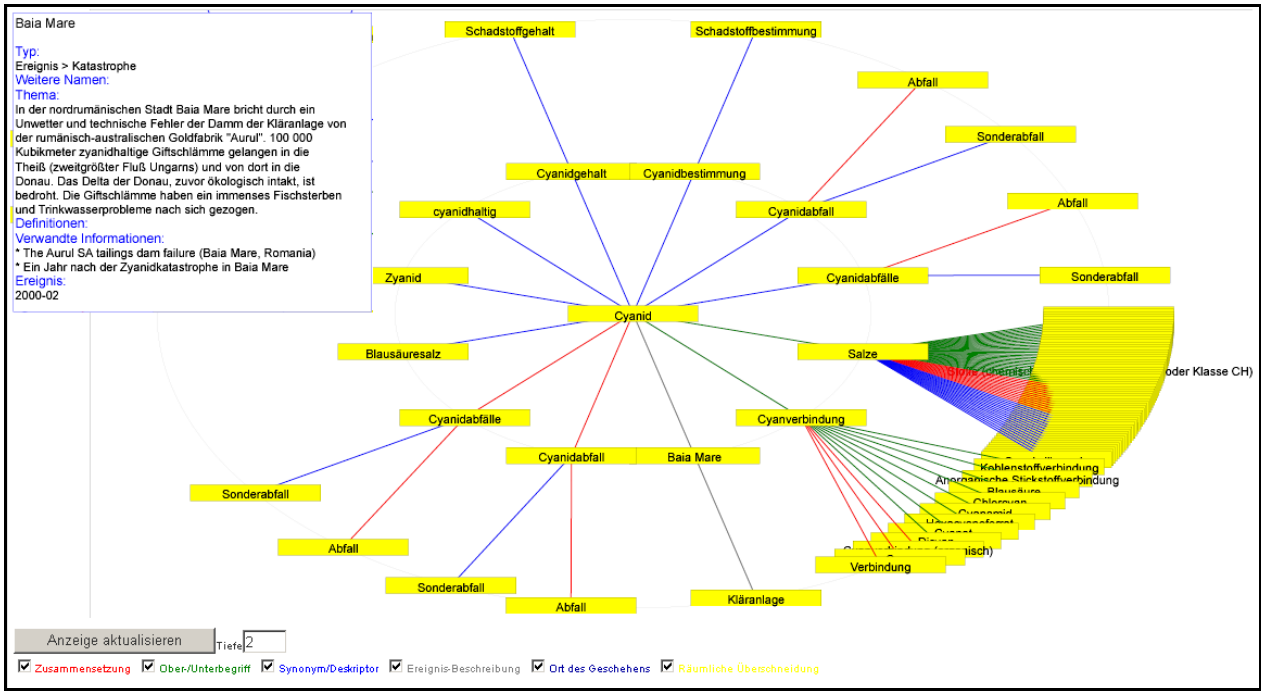


Figure 3: Interactive Graphics

### *3.2 Editorial Staff Services*

For editorial purposes, there is an editor's version of the HTML user interface which allows to add, modify, and delete Topics and Associations freely. In SNS, this interface is used to continue the chronology, and to interlink it with Thesaurus and Gazetteer.

As most of the integrated terminology (UmThes® and GN250) is maintained outside of the SNS engine today, there are basic import interfaces for both of these packages. However, it is intended to have the Gazetteer maintained directly in the same Topic Map model in the future, so the picture may change.

For SNS editors, the concept includes basic statistic services about the usage of the Topics, and about unknown words in documents or search conditions. Both will be helpful to keep the terminology up-to-date, and may be extended to a more automatic behavior in the future.

### *3.3 Enhanced Automatic Indexing*

It had been one of the major objectives to improve the accuracy of the automatic classification and indexing of environmental information as implemented by *gein*® .

This has been achieved in several parallel ways, such as:

- avoiding misunderstandings of geographical names,
- generally better resolving of homographs,
- more careful detection of composites,
- improved ranking of occurring topics to determine the most significant ones for the document in whole.

Except from some minor enhancements in document parsing and the internal document model, these improvements have become possible due to the integration of all the taxonomy into one structure (the topic map) and by establishing cross links between thesaurus, gazetteer, and chronology – the associations reported in chapter 2.2.

The analysis of associations between the topics apparently detected in the document proved to be the most important tool to clarify the accurate meaning. Although there may occur cascading misunderstandings theoretically, as misleading topics may be associated with other misleading topics, the practice showed that this risk can be disregarded. Association analysis proved to resolve ambiguities between descriptors and geographic names, and even between different geographic names that are homographs with each other.

### *3.4 Semantic Web Services*

Since 1999, the *gein*® Broker has been hosting all the taxonomy. It has been used for the classification of currently 200,000 static Web pages published by 89 information providers. There have been several requests by the information providers to be enabled to use the same taxonomy and auto-classification methods for their own purposes as well. Thinking about the effort to prepare a compact module to be distributed for implementation in 89 different technical environments, we preferred to think about a centralized service that can be accessed online by all of them. This service is designed to support the basic needs for content classification, such as:

1. Topic matching – searching for any kind of occurrence of a given character string in the Topic characteristics,
2. Topic navigation – moving from topic to topic using associations as stepping stones,
3. Auto-Classification – using a new topic-map-based linguistic analysis to extract the most characteristic Topics of a given document.

gein® looks back to very positive experiences with distributed queries using XML packed into HTTP requests, so we considered to consult the W3C Web Services Activity (http://www.w3.org/2002/ws/). According to the Web Service Description Usage Scenarios (W3C, 2002), SNS represents a “UC0015 Request-Response” scenario defined as: “Ability to describe an operation of a web service that responds with an output message or a fault based on at least one or more input messages received.”

Examples of messages and responses may be:

SNS-client: “My user is searching for information using the string “....”. Is there any Topic matching with a string like that?”

SNS: “I have X Topics, here I give you their definitions in the XML structure that we have agreed on.”

SNS-client: “I see. Please give me all the associations of Topic XYZ.”

SNS: “Topic XYZ is narrower term of Topic ABC, it is related to the Calendar Topic DEF, etc, etc.”

SNS-client: “Please give me the auto-classification results for the document at http://www.abc.xyz/any.foo. Please find the 10 most significant topics!”

SNS: “The 10 most significant topics for this document are ....”

Additionally, there may be messages according to the “UC0025 Event notification” scenario described as: “An application subscribes to notifications of certain named events from an event source. When such events occur, notifications are sent back to the originating

application (first party notification) or to another application (third party notification).” (W3C, 2002)

In this scenario, SNS is the “event source”, sending “named events” (such as a Topic Map update has happened) to the SNS-clients that have subscribed to this services before.

The XML structure and syntax of this communication are well defined using XML Schema and Web Service Description (W3C, 2001), being registered in an Universal Description, Discovery, and Integration (UDDI) registry. Details on these documents are beyond the scope of this paper.

Conforming with these standardization processes provides a general frame for interoperability of SNS. Not only the *gein*® information providers, but any potentially interested system all over the world will be able to understand immediately the technical means of integration (given an awareness of global standardization in this field).

# 4 Getting integrated

SNS has not been intended to be a stand-alone system. Its semantic model and functional services are provided to be integrated in any kind of information system maintained or planned by environmental authorities in Germany.

As a preliminarily result of the project, the semantic features of *gein*® (as developed in 1999 and 2000) have been replaced by SNS services completely.

At first, this is more than a prototype proof-of-concept, as *gein*® is the productive public portal of the German environmental authorities at the federal and state level.

Further on, the *gein*® integration is a demonstration of quality, performance, and usability of the SNS services in the community of environmental information systems.

In this field there are several integration options, targeted to different users in different application areas.

1. UDK (German Catalog of Data Sources): The (VV UDK/GEIN), an administrative agreement of the Federal and *Länder* authorities in Germany about an integrated future maintenance and development of UDK and gein®, became effective on January, 1st, 2003. One of the key subjects currently in discussion is, how to integrate the taxonomy features of both systems. Currently, both use UmThes®, but in different implementations. SNS is intended to become the common basis of both systems in the future.
2. gein® Information Providers: the (currently 89) contributing organizations are invited to integrate SNS by Web Services for any kind of information activities. Some of them intend to implement a local version of SNS themselves. Finally, there may be a network of cascading Topic Maps parts of which may be consulted depending on whatsoever localized or thematic focus of an application.
3. GeoMIS.Bund: the “Metainformation-System for geodata of the Federation” of (IMAGI) within their “national Geo data infrastructure”, a prototype of which has been developed by SchlumbergerSema in early 2002, has considered to make use of SNS by integrating an early prototype interface of the SNS Gazetteer. Currently there are negotiations with the Interministerial Committee for Geoinformation ( IMAGI ) about the future support and extended application.

4. Europe: The eEIONET community discusses “environmental web services e.g. Reportnet, country networks, and metadata, as well as terminology/ontology issues” on a European level (EIONET). As the relation between GEMET and UmThes® is very close, and as SNS already is working bi-lingual (German/English), it is a candidate to be extended to a European Scope (gazetteer) and to the full multilingual context of currently 19 GEMET languages. This has been proposed in (Bandholtz, 2001) on the EIONET work conference in Thun (CH), which led to an Expression of Interest (EoI) within the 6th Framework Programme of the European Commission.

# 5 Conclusion

SNS has successfully integrated and developed the existing *gein*® taxonomy into a service-oriented, active ontology system to be utilized by the environmental information domain by standard interfaces over the Internet. The first application that integrates SNS is *gein*® itself which now benefits from an enhanced terminology management and auto-classification. SNS is ready implemented to be used as well by national and regional systems such as the UDK, or any environmental information system. SNS includes basic interfaces for the terminology maintenance.

SNS has proved the usability of the Topic Map semantic model as well as the realism of the web service architecture which separates functional services from processes. Having this architecture, SNS is ready to play a ground role in the further development of public environmental information in the next future.

# 6 Bibliography

Angrick M., Bös R., Rüther M., Bandholtz T.: Semantic Network Services (SNS). EnviroInfo 2002. http://enviroinfo.isep.at

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. Metropolis. Marburg, Germany 2000

Bandholtz, T. / Bös, R. / Rüther, M. (2000): The German Environmental Information Network (GEIN). In Cremers, A.; Greve, K. (eds.): Computer Science for Environmental Protection '00. 2000.

Bandholtz, T. : Semantic Network Services (SNS) - a webservice for environmental taxonomy. CDS and e-EIONET Work Conference 2001: Building environmental web services. http://www.eionet.eu.int/events/cdsittag2001/abstracts/band.htm

Bandholtz, T. : A Taxi in Knowledge Land. XMLeurope 2002, Barcelona. http://62.231.133.220/idea-eks-nav/papers/03-05-03/03-05-03.html

Bandholtz, T: Taxonomie und Topic Maps. Vom Stichwortverzeichnis zur Wissensnavigation. In: XML & Web Services Magazin, Heft 2, 2002. http://www.xmlmagazin.de/itr/ausgaben/show.php3?id=103&nodeid=67

Bandholtz, T,_ Semantic Network Services (SNS). Open Forum 2003 on Metadata Registries. January 20-24, 2003. Santa Fe, New Mexico, USA. http://metadata-stds.org/OpenForum2003/

Bandholtz, T,_ Semantic Network Services (SNS). Web Services for the Integrated Enterprise. OMG's 2nd Workshop On Web Services. Modeling, Architectures, Infrastructures And Standards. February 10-13, 2003. München/Philadelphia. http://www.omg.org/news/meetings/webservices2003europe/index.htm

Batschi, W.D. Felluga, B., Legat, R., Plini, P., Stallbaumer, H., Zirm, K.: „SuperThes“: A New Software for Construction, Maintenance and Visualisation of Multilingual Thesauri. In: EnviroInfo 2002, Vienna 2002.

Bryan, Martin : XML Schema for ISO 13250 Topic Maps. Proposed Syntax 29th January 2001. http://www.diffuse.org/TopicMaps/schema.html

EIONET: European Environment Information and Observation Network. eEIONET Work Conference from 26-28 September 2002 in Vienna. Released: 2002/07/03. http://eea.eionet.eu.int/Best_Practice/eEIONET2002

EoI: European Environmental Topic Map Engine with Multilingual Auto-Classification (EETM). Expression of Interest to the 6th Framework Programme of the European Commission. June 2002. http://www.jiscmail.ac.uk/files/DC-ENVIRONMENT/EoI_Bandholtz.doc

FP6KT: Knowledge Technologies Workshop. Luxembourg, 15 May 2002. Summary report. (ISTweb | KA3 | IAF) http://www.cordis.lu/ist/ka3/iaf/iaf_workshop.htm.

IMAGI: Interministerial Committee for Geoinformation ( IMAGI ). http://www.imagi.de/welcome_eng.htm

ISO 13250: Topic Maps (ISO/IEC FCD 13250:2000). Prepared by: ISO/IEC JTC1/SC34 - Document Description and Processing Languages. http://www.ornl.gov/sgml/sc34/document/0058.htm

IST: Knowledge Technologies within the IST 2002-2006 programme. Some reflections. Internal report on Knowledge Technologies (aka IRG Report). Luxembourg 2002.

SNS: Federal Environmental Agency of Germany, SchlumbergerSema: Semantic Network Services. Research Project UFOPLAN-Ref. No. 20111612, promoted by BMU/UBA, Germany

UDDI: UDDI Version 3.0 Specification. Open Draft, Dated 03 July 2002. http://www.uddi.org/pubs/UDDI-V3.00-Open-Draft-20020703.htm

VV UDK/GEIN: Verwaltungsvereinbarung zwischen Bund und Ländern über den gemeinsamen Betrieb und die gemeinsame Entwicklung und Pflege des Metainformationssystems Umwelt - Datenkatalog UDK und des Umweltinformationsnetzes Deutschland GEIN.

W3C (2001): Web Services Activity: Web Services Description Language (WSDL) 1.1. W3C Note 15 March 2001. http://www.w3.org/TR/wsdl

W3C (2002): Web Service Description Usage Scenarios. W3C Working Draft 4 June 2002. http://www.w3.org/TR/2002/WD-ws-desc-usecases-20020604/

XMLe: XML Europe 2002. Down to business: Getting serious about XML. Barcelona, 20-23 May 2002. Proceedings. http://www.idealliance.org/papers/xmle02/

XTM: Members of the TopicMaps.Org Authoring Group: XML Topic Maps (XTM) 1.0. TopicMaps.Org Specification. http://www.topicmaps.org/xtm/index.html