

A Taxi in Knowledge Land

A Use Case that Combines Topic Maps and Web Services in a Public Portal

Thomas **Bandholtz** <TBandholtz@slb.com>

Abstract

Bandholtz reports on a R&D project with a focus on implementing topic map functionality as a web service. The primary usage is within a public portal (German Environmental Information Network), but the service is provided for general use in the scope of environmental protection. The topic map contains a thesaurus and a gazetteer. The report gives an overview on architecture and methods, and on the tools employed, and on standardization issues.

Table of Contents

1. Intro	1
2. The German Environmental Information Network (GEIN)	2
2.1. XML-Based Internet Information Brokerage	3
2.2. Automated Knowledge Management	3
3. A Topic Map for GEIN	4
4. Enhanced Auto-Classification	5
5. Semantic Web Services	7
6. Knowledge Taxi and the XML Topic Map Engine	7
6.1. Topic Map Design and Editing	7
6.2. Storage and Retrieval	8
6.3. Methods and User Interface	8
6.4. Interfaces	9
6.5. Basic Classification Support	9
6.6. Index of Classified Documents	10
7. XML Topic Map Schema	10
7.1. DTD vs. Schema	10
7.2. Using Inheritance to Define Topic Types	11
7.3. Defining Association Templates	11
7.4. Extensions of the ISO 13250 Standard ?	12
8. Interoperability	13
8.1. Interchange Formats	13
8.2. Published Subject Identifiers	14
9. Conclusion	15
Bibliography	16

1. Intro

Knowledge Land is the place where we all live. But most of the time we don't care – may be even don't know. And when we try to access some of the Knowledge around, we are missing some kind of map to make our way.

If you live in an unknown city, it will be comfortable to use a taxi now and then. This is what we are trying to develop under the code name *KnowledgeTaxi*.

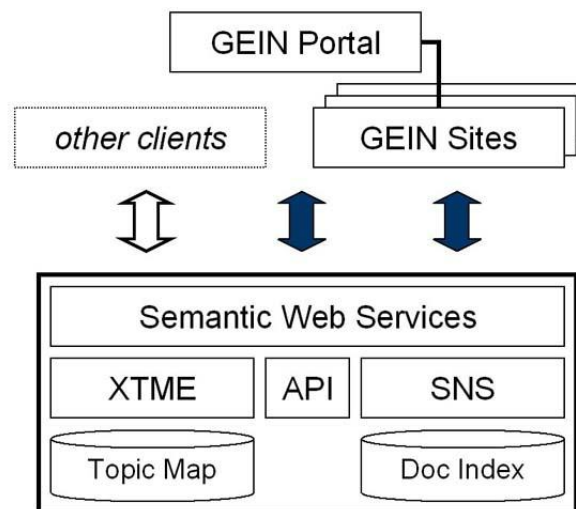
I will discuss a use case here that combines Topic Maps and Web Services in a life project, accomplished by a governmental authority in Germany.

All this is being implemented using XML as the anywhere data structure. And as this is a dedicated XML conference, I presume that you have a basic understanding of [XML Schema], and [Web Services]. If not, you might have difficulties to understand some of my lingo now and then, but I will try to explain when these items come up.

As you are interested in XML and Knowledge management, you should also have heard about Topic Maps. Basically, a Topic Map [ISO 13250] is a network of *topics* ("terms" pointing to subjects), *associations* between topics, and of *occurrences* of these topics in information objects (documents or databases).

The general architecture of the use case scenario is shown in Figure 1.

Figure 1. General Architecture



The use case project is a follow-up of the [GEIN] project that I will introduce below. The main issue is that there are some semantic Web Services that will be accessed by the GEIN Portal and the GEIN Sites (that are represented by the portal) equally. Further on, there may be "other clients" to use the same Web Services. The Web Services themselves are based on the XML Topic Map Engine [XTME] module, and on some functional extensions implemented by the Semantic Network Services [SNS] R&D project. While XTME supports core functionality like searching the Topic Map, or accessing topics and associations by reference, SNS is dedicated to more complex methods like auto-classification of documents, or searching for documents that have been classified before.

How can this work like a taxi? Because the SNS auto-classification also works for your query input. So you can give a very vague description of what you are searching for, as "Have to catch my plane at seven" and the taxi driver will understand that you are trying to be at the airport in half an hour.

After having found some documents using a query like "What really happened since XML Europe 2001?" (documents that have been published later than the conference date, and covering the major conference topics) you may point to any of these documents and say: "Tell me more about this!", and the Service will read the classification (or auto classify any new document) and present some topics that you should use to find "more about this".

Also, you can navigate through the map, starting with one of the topics, see the topic characteristics, take a look at some associations, and proceed to any topic associated to your current topic.

So topics appear like *named locations* in Knowledge Land, and the "taxi driver" will know anything about them.

2. The German Environmental Information Network (GEIN)

One of the first governmental Web Services in Germany is currently developed in an R&D project, [SNS] by the Federal Environmental Agency and SchlumbergerSema - based on the experiences with [GEIN]. SNS will result in

"... a set of semantic Web Services that make thesaurus-based content analysis accessible by any [...] Web Server via HTTP. GEIN should be on the way to become an interactive environmental topic map for public use." [\[BA 2001\]](#)

Let's take a short look at the project history. In 1998 the European Union released the Aarhus convention on "access to information, public participation in decision-making and access to justice in environmental matters" [\[AARHUS\]](#). This turned out to be a unheard-of challenge to the information policies and technologies of all the governmental authorities concerned.

In Germany, the Federal Environmental Agency ("Umweltbundesamt", UBA) decided to develop an "information broker" as a single-point access to any public information of the authorities on the federal and state ("Laender") level. They called this the German Environmental Information Network. GEIN was presented as the Agency's contribution to the EXPO 2000 and has become one of the favourite models for governmental information portals in Europe.

2.1. XML-Based Internet Information Brokerage

In the 90s of the bygone century, most of the environmental authorities started to use the internet for public information. Now we can say: most of the commonly requested information is accessible via the Web (currently GEIN deals with 140,000 static Web pages published by 70 information providers).

But still the information is distributed among many places, and each information provider follows his own rules and preferences. There is no harmonisation but being "in the internet", and there is (almost) no localisation support but common search engines. Thus the end user, though amidst of a wide variety of environmental information, just happens to find some of it without being able to compare it to what *could* be found.

While developing the GEIN broker, we experienced that a growing amount of information providers connected their databases to the internet dynamically. This information can not be detected using search engine crawler methods. So we established a distributed query mechanism using XML and HTTP requests. This might be called an early Web Service, responding with result sets of hyperlinks as an answer to a search condition. Currently we have 9 database servers participating in Germany, containing more than 500,000 accessible information objects.

We implemented all this as an XML based Web solution. From a technical point of view, GEIN has been buildt on "XML anywhere". XML has been used as a universal data format for storage [\[Tamino\]](#), retrieval [\[XPath\]](#), processing [\[DOM\]](#) and communications (XML-via-HTTP). We were one of the international wide spread projects who invented the "wheel" of communicating XML in the body of HTTP requests and responses, which short time later has been standardized as [\[SOAP\]](#) by W3C. In all, GEIN demonstrates

".. state-of-the-art web techniques such as XML, Java Servlets, and HTTP communication, ... providing a fast and dynamic, thesaurus-based knowledge management." (from the Agency's marketing flyer).

2.2. Automated Knowledge Management

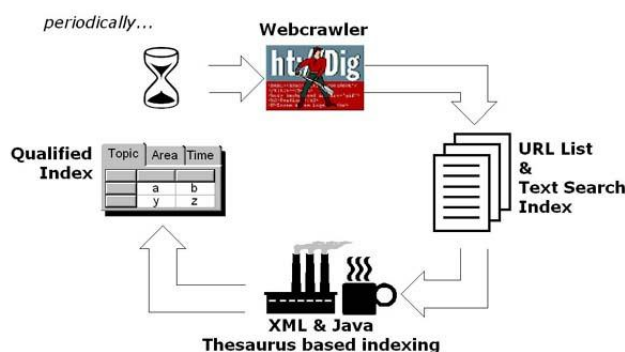
In the field of Knowledge Management, [\[GEIN\]](#) uses three semantic structures:

- a thesaurus of currently 39,143 environmental terms,
- a gazetteer including the intersections between 48,213 geographical objects of all kinds,
- a synopsis of historical and contemporary events that have affected the environment, currently 544 events.

Still any of these collections is separated from the others, and each has an individual data model.

Besides that, we integrated a conventional full text search engine. GEIN has chosen an Open Source engine [\[ht://Dig\]](#). These tools have been integrated into an highly automated, permanent indexing cycle.

Figure 2. The Automated Indexing Cycle



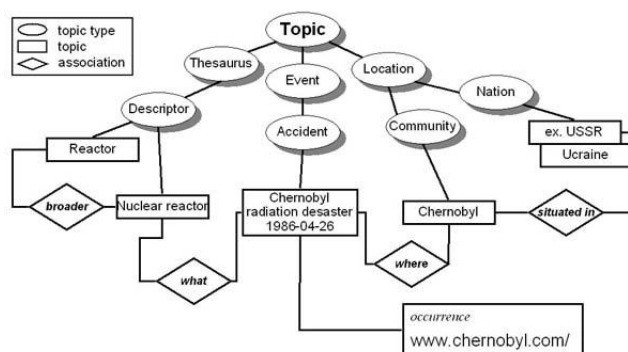
The ht://Dig crawler starts the indexing cycle, controlled by a list of domains that the crawler is not allowed to leave. This crawler generates a simple text index used by GEIN's full text retrieval facility. As a second "give-away", it generates a list of all the URLs it has scanned.

The URL-list is input to GEIN's thesaurus-based indexing machine, which analyzes the content of each page to find terms, or synonyms, geographical names, and time notations. In an automatic mode, the indexing machine registers the most significant keywords covering topic, and area, and one time, or timespan notation.

3. A Topic Map for GEIN

[SNS] converts the thesaurus, the gazetteer, and the synopsis into one environmental Topic Map. Additional associations are used to semantically interconnect these three collections to provide some understanding of questions like: "What happened since Chernobyl?"

Figure 3. Converting Thesaurus, Gazetteer, and Calendar into one Topic Map

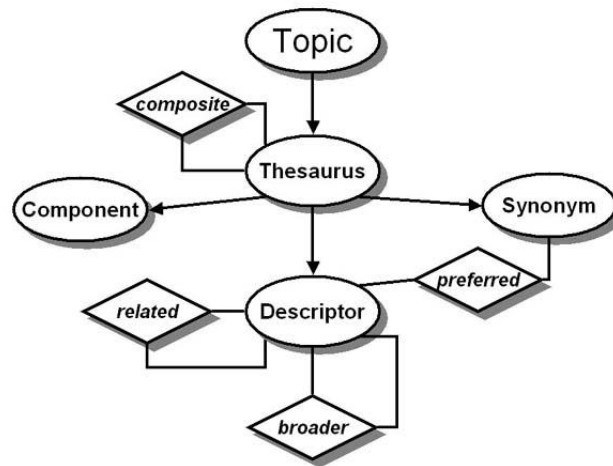


In Figure 3, the black ovals stand for *topic types*. Typically topic types build some hierarchy like in “/topic/location/nation”, very similar to a class hierarchy.

The rectangles stand for topic type instances, or to say it more simple, for topics. The rhombs indicate associations between topics.

This overview diagram contains only a snippet of the whole Topic Map typology. To give a closer impression, Figure 4 shows all the topic types and association templates used to represent a classical thesaurus.

Figure 4. Topic Map Typology of a Thesaurus



The figure reads: A topic may be of type “thesaurus”. A “thesaurus”-topic may be a “descriptor”, a “synonym”, or a “component”. Any “thesaurus”-topic (including the subtypes) may have a “composite” association to any other “thesaurus”-topic. Only “synonyms” may point to a “descriptor” indicating that this is the “preferred” term. Any “descriptor” may be “related” to, or be the “broader” term of another “descriptor”.

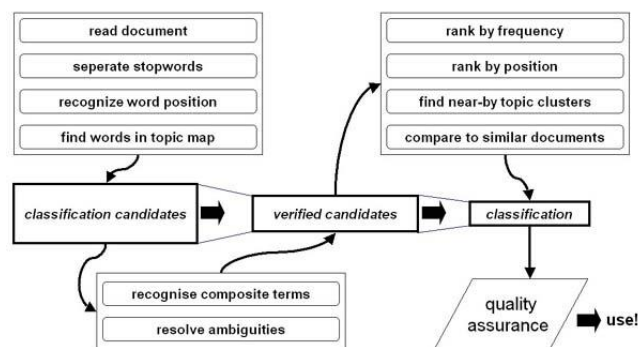
Now you may add more association templates between any type within the whole map and either the rather abstract “thesaurus” type (including all the subtypes), or more closely to “descriptor” only, as in the Chernobyl sample above.

4. Enhanced Auto-Classification

[GEIN] already has implemented a linguistic analysis of documents to suggest keywords that characterize each document. This analysis can also be used to process search criteria, so search criteria may be any kind of text, including the complete content of a given document. The new [SNS] will provide similar text analysis facilities which result in topic-based indexing of documents (passed by URL or as complete text).

But it is not enough to replace “keyword”-references by “topic”-references in the index. The whole process is shown in Figure 5. Generally we start by creating a large amount of possibly significant topics (classification candidates), and we try to reduce this set by “sorting signal from noise” using a rather linguistic approach. Most of the steps in this process do not depend on a Topic Map at all. But – far from this – the Topic Map has to support the requirements of a linguistic analysis by storing the word morphology of each name, and it has to know “stop-words” (known words without a significant meaning by themselves). This is an extension of the basic Topic Map concept as defined in [ISO 13250] (For those who are familiar with this standard: we used the facet concept to implement morphology – and even all the phonetics as well - in a Topic Map).

Figure 5. Auto-Classification Process

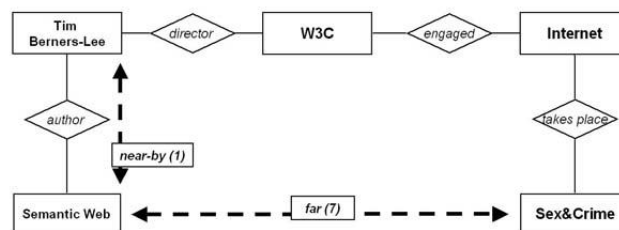


There is one important step in the process shown by [Figure 5](#) that utilizes the Topic Map paradigm in a way that would not have been possible without it: “find near-by topic clusters” The definition of Topic Maps in [\[ISO 13250\]](#) explicitly emphasizes the concept of a topic distance:

“A topic map defines a multidimensional topic space - a space in which the locations are topics, and in which the distances between topics are measurable in terms of the number of intervening topics which must be visited in order to get from one topic to another, and the kinds of relationships that define the path from one topic to another, if any, through the intervening topics, if any.”

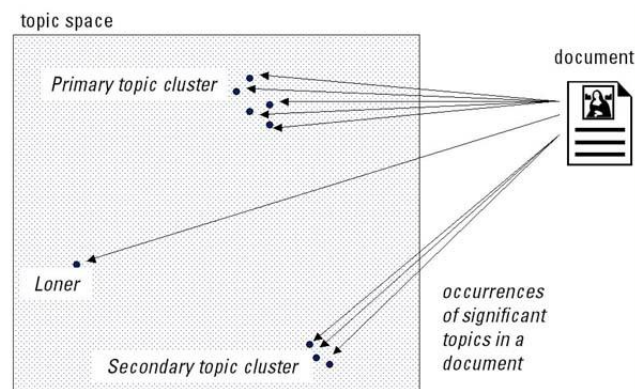
[Figure 6](#) shows a sample: in this Topic Map snippet the “near-by” distance is just “one association away”, while the “far” distance has to pass four associations and three intermediate topics.

Figure 6. Topic distances



We might criticize the semantics of this sample – it does not look very intelligent, but here it just serves to illustrate the idea of topic distances. [Figure 7](#) shows a utilization of the distance paradigm. After having found a crowd of topics that apparently occur in a document, we analyse the distances between them to figure out clusters of relatively adjacent topics. The one with the most topics in it is regarded to be the primary cluster, most probably indicating the most significant area of interest covered by the document. If we detect some isolated “loners” in this topic space, it is not easy to decide whether this is an error of the processing algorithm, or a well-detected, significant aspect of the document classification.

Figure 7. Near-by Topic Clusters



As you may imagine, this works in a quite experimental state, and it is highly dependant of the semantic intelligence represented by the Topic Map used. I.e. add a direct association between “semantic web” and “internet” in [Figure 6](#), and everything looks completely different. BTW, this stresses the significant importance of *qualified taxonomists* to be editors of any Topic Map. A well developed algorithm, running on a “stupid” Topic Map, must result in nonsense. Or – like anywhere: garbage in – garbage out (the same will be the case for “stupid” documents ...).

5. Semantic Web Services

Today, [\[Web Services\]](#) has become a buzzword. [\[GEIN\]](#) been using a (pre-SOAP) XML-via-HTTP communication since 1999, and so we have been very close to SOAP and Web Services all the time.

There was a particular reason to search for something like Web Services – resulting from the distributed architecture shown in [Figure 1](#). We always have strived for some semantic “harmonization” of the portal and all the contributors – but this is a completely heterogeneous landscape. Many of the contributing sites have asked for a copy of the GEIN thesaurus engine, but – despite the general Java portability of the code – there was a considerable diversity of operating systems and Java versions, not to talk about database implementations used. To avoid a huge trouble with versioning and configurations (including the distribution issues), we always answered these requests with the recommendation to use Web Services, or – more early – a customized XML-via-HTTP communication before Web Services had been defined. The European Environmental Agency must have experienced something similar when they demanded:

“... a set of emerging standards and practices from the e-business world will have to be adopted by the environment sector. These include but are not limited to XML/edi, SOAP, ebXML, UDDI, ISO/IEC 11179, and more.” [\[CDS 01\]](#)

This was an encouraging statement for the German authorities, being confronted with several obstacles by routing problems when passing secure firewalls of some contributors – but these problems still exist. Finally, I prefer finding fast but secure routing solutions to distributing a database application in a heterogeneous field.

Anyway, [\[SNS\]](#) will provide access to all the functionality as a Web Service, starting with a simple Topic Map search, up to auto classification services:

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 an enhanced, topic-map-based linguistic analysis to extract the most characteristic topics to describe a given text fragment (“document”).

We do not support Topic Map *editing* via Web Services here– as this is done by a special team of responsible taxonomists. They use a separate Web user interface that cannot be accessed by anybody outside the team.

6. Knowledge Taxi and the XML Topic Map Engine

In parallel to the project progress described above, the XML Network and the Competence Center Content Management of SchlumbergerSema started to experiment with a XML Topic Map Engine [\[XTME\]](#) in late 2000. XTME is included in the project scenario as shown in [Figure 1](#).

XTME supports the following basic functionality:

6.1. Topic Map Design and Editing

The first step in creating a Topic Map is to specify the topic types and association templates to be used. This is up to the Topic Map owner. Although the owner may modify types later, this will not be very comfortable for the users. Types should be rather persistent – at least some basic core types. XTME offers some ready-to-use types like “Event”, “Thesaurus”, or “Location” that you may use as the first level of the type hierarchy.

An association template specifies which topic types may be used by the associations that apply this template. I.e. an association as “IsSituatingIn” should only interconnect a “Location” with another “Location”, and not with an “Event”. “Event”-topics may be interconnected with “Location”-types using “HappenedIn” associations, etc.. Some examples are shown in [Figure 3](#).

Once you have set up the typology you will proceed creating the topics and associations. Most use cases will want to “import” some legacy taxonomies, such as customized thesauri, gazetteers, classification systems, or simple keyword lists already in use. This will be relatively easy if you manage to convert the existing taxonomy to any XML format first, and use XSL Transformations [XSLT] to convert it to the target schema, and then import this into the XTME database.

The topic map owner, optionally a dedicated team of taxonomists, may create and edit topics and associations at any time, but this should be done carefully, or it will confuse the users and possibly affect the usability of the whole map. XTME (like any other Topic Map engine) can validate the consistency between your decisions and the existing map, and it can move topics from one type to another, but it cannot discuss the semantic quality of your decisions.

XTME will include some statistics about the usage of types and associations, and a comfortable notification mechanism, so any user can register proposals about missing (or apparently erroneous) topics or types. But the owner has to decide, as he will be the only one who knows all the existing topics and associations, and the reason why everything has been created this way, or has not been created for some reasons.

XTME tries to support this process and prohibits that any user can edit the map on-the-fly without having taken into account the over all consequences. If you think this is not necessary, you simply may register “all users” to be taxonomists – but we would not recommend that. (It would be something like allowing every stranger to modify a city map).

6.2. Storage and Retrieval

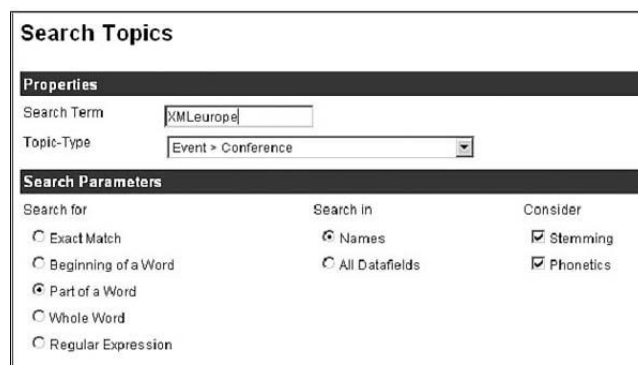
XTME supports storage and retrieval of Topic Maps, including the necessary administrative data. The internal data access layer is designed as an abstract interface, to allow a free choice of the database behind. Currently we use the [Tamino]XML Server to benefit from native XML storage and [XPath] as a supported query language. But it is possible to use any database engine. Generally, there is a mapping of the XML Schema to a relational model including SQL as the query language. So XTME can be integrated with most of the existing environments.

In principle, the storage could be implemented using XML documents in the file system and any open-source XPath or [XML Query] processor. But we do not recommend this for a production system currently.

6.3. Methods and User Interface

Besides the design and editing features described above, XTME supports several query methods with many options, as shown in the sample screen shot in Figure 8.

Figure 8. Search options (screen shot snippet)



The screenshot shows a web interface titled "Search Topics". It is divided into two main sections: "Properties" and "Search Parameters".

Properties:

- Search Term:** A text input field containing "XMLEurope".
- Topic-Type:** A dropdown menu currently showing "Event > Conference".

Search Parameters:

Search for	Search in	Consider
<input type="radio"/> Exact Match	<input checked="" type="radio"/> Names	<input checked="" type="checkbox"/> Stemming
<input type="radio"/> Beginning of a Word	<input type="radio"/> All Datafields	<input checked="" type="checkbox"/> Phonetics
<input checked="" type="radio"/> Part of a Word		
<input type="radio"/> Whole Word		
<input type="radio"/> Regular Expression		

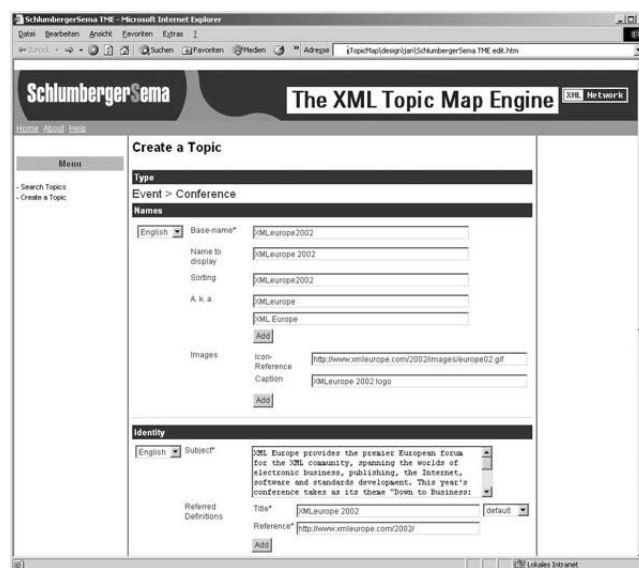
Once you have found some topics, you can navigate along the associations, or follow links to the definitions and attached information.

There is a basic, bilingual “out of the box” user interface ready for use. This is implemented based on Java Server Pages and XSL Transformation [XSLT]. An application can edit several configuration files to adapt the look&feel (wording, symbols, colors, etc.) to any corporate or community style guide. Modules of the XTME user interface can be integrated into any more complex layout used by the application community.

On the other hand, the default layout of the customisable user interface (Figure 9) leaves space to integrate additional features to be used in any local environment. This might be applied for the taxonomists, while other users will find XTME dialog modules integrated in their common working situation.

All this is Web based only. So it perfectly fits into public, or Intranet knowledge management systems that use a Web browser as the the user interface platform.

Figure 9. Default layout of the customisable XTME user interface



6.4. Interfaces

XTME is designed to be a "plug-in" that can be integrated with any Document/Content/Knowledge Management solution by means of a Web Service, i.e. using SOAP calls, or even HTTP GET requests.

I.e. if you are using a customisable Content Management System that lacks an elaborated handling of taxonomies (like Tridion, Imperia, and many others) you can use XTME as a plug in for these systems.

If you want to integrate any system with XTME more closely (like we did it for SNS), you may prefer to use the XTME Java API (Figure 1).

6.5. Basic Classification Support

XTME itself does not contain a full, general auto-classification method ready to use, as parts of this method are dependant of the specific topic types and association templates. You may use the basic classification support stand alone, but this will not get the most of the semantic intelligence of the Topic Map. I.e. when you use the thesaurus model (Figure 4), you will try to inspect the broader terms of the topics found in your document, and you certainly will differentiate between synonyms and descriptors in a well defined way. This will produce much better results than only considering some type of topic and association without understanding the semantic meaning of the specific association template and the associated topic types.

XTME implements a general document model that assigns topics to words, and words to sentences that are located in chapters. There are basic statistical ranking methods. But many steps of the processing model shown in Figure 5 can only be implemented with a knowledge of the specific semantic model in use.

In our use case, the application specific [SNS] module uses the XTME document model to implement the possibilities of auto classification with an unrestricted awareness of the specific semantic model. Remember that SNS is developed in a public R&D project. We are experimenting, and we would not be SchlumbergerSema if we hadn't in mind to find some more generalized methods that may be plugged into the XTME document model for general use.

6.6. Index of Classified Documents

The Topic Map paradigm contains an concept of topic occurrence which means:

“groupings of addressable information objects around topics (occurrences)” [ISO 13250]

Of course XTME supports this in its topic model. But a document index is something else, it rather means “groupings of topics around addressable information objects”. There are *several* topics that are significant for a document. Occurrence means that there are several documents significant for a topic (i.e. as a definition, or application sample).

While XTME includes the occurrences anyway, it not necessarily contains a document index. Some applications already have a document index themselves (i.e. when you plug XTME into a Content Management System). Some applications may not need any document index at all. XTME includes a Schema for document indexes, and it supports some user assistance on the way from his initial query to a set of topics that will serve as the final search condition. In our use case, SNS needs a document index , and so this part of the XTME Schema is implemented by SNS.

There is another reason to distinguish occurrences from a document index that is relevant for portals: A portal like [GEIN] is designed to cover information from some specified contributors only. References to these documents will be kept in the document index only. But if you want to link to the relevant definitions of the topics themselves you probably not want to be restricted to the selected contributors of your portal. I.e. the sample in Figure 3 contains an occurrence that links to <http://www.chernobyl.com/> which is definitely *not* a domain owned by a German governmental authority, and so it could not be referenced from within the document index.

7. XML Topic Map Schema

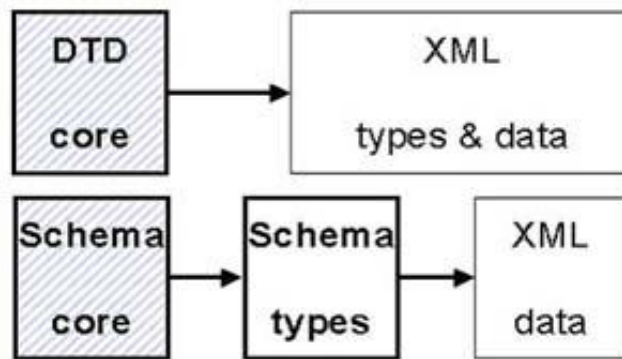
When we started to think about the XML document structure suitable for Topic Maps, we found a SGML DTD contained in ISO 13250, and we found *TopicMaps.org* (<http://www.topicmaps.org/>), at that time an independant group of people who were developing a XML DTD named [XTM].

7.1. DTD vs. Schema

But we also found the XML schema activity within the Architecture domain of W3C [XML Schema]. Watching this specification process we found it highly presumably that this would result in a W3C recommendation soon. We made only one tiny mistake: it resulted in *three* recommendations :-).

While the XTM DTD was developed in a very abstract way, i.e. using topics as topic types, we intended to develop a truly extensible TM schema that would contain type definitions and association templates, and completely separate this from the Topic Map data itself. With XTM, the typology resides in the Topic Map data (Figure 10).

Figure 10. Separating typology from data and core definitions



7.2. Using Inheritance to Define Topic Types

Similar to a programming language like Java, XML Schema supports complex data types and inheritance.

You may define `<topic>` as a complex data type and you may define `<event>` as an extension of `<topic>` which means that `<event>` inherits all the definitions given in `<topic>`. Then you may define `<conference>` as an extension of `<event>`, and so on.

This would result in a Topic Map serialization where some topics are tagged `<topic>`, while others are tagged `<event>` or `<conference>`. As `<topic>` is defined rather abstract (means a topic without any type information) in most Topic Maps then there will be no `<topic>` tag at all. This is a consequence that can be found in Martin Bryan's approach at [\[BRYAN 01\]](#). Martin uses "substitution groups", which is a valid approach, but may be not eligible here.

We regarded this to be not reader-friendly – though a parser or any processing engine will have no problem after having interpreted the schema. But what about [\[XPath\]](#)? How to address a topic by search criteria when you do not know the type in advance?

XML Schema allows to write something like:

```
<topic xsi:type="eventType">
```

in a conforming XML document as an equivalence for:

```
<event>
```

if `<event>` is a descendant of `<topic>`. This solved our problem, and you may access the type as a simple attribute – any DOM parser understands this, and any XPath expression can handle this efficiently.

7.3. Defining Association Templates

Association templates are a little more complex. I.e. we wanted to specify that an association template "what-happened-where" may have a member "what" and a member "where".

This looks easy. But as a template, this should contain some restrictions about the members. "what"-members must be of the type `<event>` (including the descendants of `<event>`), and "where"-members must be of the type `<location>` (or descendant).

This led us to the limits of XML Schema possibilities. We did not find a way that the parser could control these restrictions while validating a document – a very interesting challenge, and it would need several pages to summarize what we have tried.

Finally we contented ourselves with a path attribute containing a string-serialization of the path-of-inheritance, as in “/event/conference”, with <topic> as the root (symbolized by the leading ‘/’). Now we could use simple regular expressions like “/event/*” to express that the type may be any of the descendants of event. This is easy to be used, also in XPath expressions.

For association templates, we defined an abstract complex type <member> containing a general topic reference and a path attribute, and we restricted this type by fixing the path attribute to a certain value for each specific member type. This schema snippet is shown in [Figure 11](#).

Figure 11. Member type definitions in an association template

```
<xs:complexType name="KTmemberType" abstract="true">
  <xs:attribute ref="xlink:href"/>
  <xs:attribute name="mpath" type="xs:string" />
</xs:complexType>
[...]
<xs:complexType name="eventMemberType">
  <xs:complexContent>
    <xs:restriction base="KTmemberType">
      <xs:attribute ref="xlink:href" use="required"/>
      <xs:attribute name="mPath"
        type="xs:string" fixed="/event/*"/>
    </xs:restriction>
  </xs:complexContent>
</xs:complexType>
```

We added a similar path attribute with the same fixed value to the definition of the topic type <event>.

A fixed value attribute, once defined in the Schema, needs not do be repeated in the conforming data. The parser will complement this attribute automatically when it is missing in the data, or it will raise an error if it has been set to any value different from the fixed value.

I never have been real enthusiastic about this approach, as the application of the template definition still is not really controlled by the parser. But at least the application will find a clear statement that can be processed very easy and efficient. I.e. when you want to list topics that may be referred to in a <member> instance, it simply compares the path attribute of <topic> to the mpath attribute of the <member> type.

7.4. Extensions of the ISO 13250 Standard ?

When I was talking about extending <topic> above, I just talked about building a type hierarchy – but nothing has been really extended yet. We also use XML schema inheritance to extend specific topic characteristics of certain subtypes. I.e. any <event> should have some temporal extent assigned to it. As far as I understand [\[ISO 13250\]](#) recommends to use facet types in this case:

“In addition, information objects can have properties, as well as values for those properties, assigned to them externally. These properties are called ‘facet types’.”

This is not made really clear, and it has been written without considering the XML schema inheritance (which was just being developed at that time). The [\[XTM\]](#) DTD does not use facet types at all.

If you want to express: “XMLEurope2002” (a <topic>) started at “2002-05-20” (the starting point of the temporal extent), XTM DTD would require that you make “2002-05-20” a topic itself, and interconnect both by an association of type “started-at” (or similar). Certainly this could express the fact. The problem is: I do not want “2002-05-20” to be a topic, I do not want to be forced to make any topic characteristic a topic itself. This might become a rather philosophic discussion as long as you do not consider processing. If I want to find all events that took place in May, 2002, I can use a query like:

```
"select all topics having path='/event/*' and date='2002-05*'"
```

(abstract notation) which is easy to express in XPath, or SQL as well. But – I need a date attribute within the topic object. Try this with the association style!

So, as we have been looking for an efficient production format, we used inheritance to add characteristics (<xsd:element> or <xsd:attribute>) to topic types whenever needed. Another example: we add a [GML] -bounding-box to any location. We use all this very restrainedly – a Topic Map is not a universal database containing any information about the topics, it is only a general map. But we use such extensions whenever we feel that we would need an attribute to navigate and search the map.

8. Interoperability

8.1. Interchange Formats

In December 2001 the [XTM]1.0 DTD has become an annex of [ISO 13250], and since that you find some companions using “ISO 13250 XTM” when they talk about XTM. This refers to the annex, it does not say that XTM is ISO 13250. ISO 13250:2000 already contained a SGML format („HyTime metaDTD“), accompanied by a general note that any XML format would be “legal” as well:

“As the Extensible Markup Language (XML), a World Wide Web Consortium recommendation, is a subset of SGML, (...) XML can be also used as a base notation for Topic Maps.”

XTM has been accepted by ISO/IEC JTC1/SC34 as a second valid interchange format that does not add any normative definitions. Jim Mason, Chairman of ISO/IEC JTC1/SC34 stresses this very clearly in his posting from 2001-12-14 <http://lists.oasis-open.org/archives/topicmaps-comment/200112/msg00012.html>):

“BOTH the HyTime metaDTD originally published in the standard and the XTM 1.0 DTD are now equally parts of the standard. Each one provides an approved ISO/IEC mechanism for interchange of Topic Maps. (...) We need to keep clear that the transfer serializations are not the definition of Topic Maps: The standard is the definition. SC34 intends that the supplementary standards will clarify the meaning of Topic Maps without changing their essential nature. (We also recognize that other transfer serializations are possible, outside the standard.)”

Mason also refers to a short document by Biezunski, Bryan, Newcomb „Differences between XTM 1.0 and the HyTime-based meta-dtd“ [BBN 01]. This document shows that there is a considerable tolerance in comparing the standard definitions to those of any interchange format. Among these tolerated differences is one aspect important to our approach:

“Facets are not mentioned in the XTM DTD.”

(Remember XTM uses topics and associations instead). Besides this, the document clearly states that XTM is not designed to be a processing format:

“Neither HyTime-based nor XTM-based topic map documents are “ready-to-use” by application-specific logic.”

But that exactly was what we have been looking for. The XML schema in use by [XTME] is not designed to be a general interchange format, but just an implemented example of a processing format, and we will not submit it to any standardization committee with the intention to make it another annex of any standard. Remember that everybody is free to use any kind of data format in his own processing environment.

In the case of interchange of a Topic Map or of Topic Map fragments, Jim Mason’s posting quoted above clearly expresses a general tolerance (“other transfer serializations are possible, outside the standard“). In my opinion, the choice of interchange formats depends on the community that implements this interchange.

As we have been discussing Web Services in this article, we should be aware that the Web Service Description Language [WSDL] includes a <wsdl:types> section, where you should include the schema used by your service. This provides a solution for everyone to offer his Topic Map interchange services in any chosen format, including

an understandable description of the format. Of course, this does not solve the problem to understand any specific semantic of a service format. That is why a well defined, but maybe less efficient, interchange format that is understood by everyone still helps.

But I have my doubts that XTM is of the kind to become widely accepted as the de-facto interchange format used by a growing Topic Map community in the future. I believe that I conform with the standard when I say: This still is an open issue. But I also want to state clearly that I see no problem to convert XTME schema serialization to XTM and back – it will work without loss.

8.2. Published Subject Identifiers

Any topic refers to a subject. If I have a topic named “XMLEurope2002” this probably will have this conference as its subject.

ISO 13250 contains a definition of “subject” that I personally consider to be a very exceptional masterpiece of standardization definitions (not a joke!):

“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] distinguishes addressable subjects from non-addressable subjects. I.e. if I design my personal Topic Map, and I create a topic named “cloud”, and I only describe the subject as:

“condensed water in the air sometimes looking like pink elephants in the evening sun.”

within the identity attribute of my topic, this is a non-addressable subject. It is a valid subject, remember “whatever” can be a subject. But if I create a topic that is talking about “XMLEurope2002”, I might not want to give my own definition but just refer to <http://www.xml europe.com/2002/>. This is an *addressable* subject, and I would consider this to be a public subject descriptor following the ISO 13250 definition:

“A subject descriptor (...) which is used (or, especially, which is designed to be used) as a common referent of the identity attributes of many topic links in many topic maps.”

Now XTM has introduced a slightly different concept of a published subject:

“A published subject is any subject for which a subject indicator has been made available for public use and is accessible online via a URI. A published subject indicator is therefore any resource that has been published in order to provide a positive, unambiguous indication of the identity of a subject for the purpose of facilitating topic map interchange and mergeability.”

This stresses somehow what already has been said in the definition of public subject descriptor („... which is designed to be used ...“). It is not quite clear that <http://www.xml europe.com/2002/> has been created with this intention, including all the criteria of a published subject.

After TopicMaps.org has migrated to become an OASIS member section in late 2001, an OASIS Technical Committee has been founded to give this a more specific meaning, see

- Topic Maps Published Subjects (PubSubj) Technical Committee (<http://www.oasis-open.org/committees/tm-pubsubj/>).

In his „Recommendations for Documentation of Published Subjects” [VAT 02] Bernard Vatant describes the main objective:

“The general purpose is that topic maps interoperability needs non-ambiguous definition of subjects (represented by topics), that should be provided by stable resources, made available online through trustable publication process.”

There are many stable sets of identifiers in common and reliable use, i.e. ISBN codes for publications, or United Nations Location Codes for ports and other locations [LOCODE]. If you arrived here by plane, your ticket will contain the code „BCN“ for the target airport Barcelona. This is a fragment of the United Nations Location Code „ES BCN“ for Barcelona, Spain. (There is also a code “VE BLA” for Barcelona, Venezuela. I guess your airline did not mistake this). So what, while you are in the air, United Nations would have changed the codes, or just discontinued to support any code at all? I am sure your plane would have touched down at the Barcelona air port (Spain) anyway, as you did not use a topic for this flight.

You may consider this to be a foolish example, but to me it makes clear what is needed. Talking about topics instead of planes again, the current state of [LOCODE] works for travel applications, but not as a set of published subjects. There is a URL pointing to the recommendation (<http://www.unece.org/cefact/rec/rec16en.htm> and another URL pointing to the whole set of codes <http://www.unece.org/cefact/locode/service/main.htm>), but there is no URI pointing to „ES BCN“, and if there was, it would not point to a topic in the current situation.

Especially in this application field there has been founded a second TC, the

- OASIS Topic Maps Published Subjects for Geography and Languages (GeoLang) TC (<http://www.oasis-open.org/committees/geolang/>).

While this article is been written the work of both committees has just started. In my opinion, there has to be more work on technical details, especially on

1. How to address a published subject using a persistent Uniform Resource Identifier [URI], independent of the “media type” used to publish the subjects; and
2. Which format should be used to publish the subject, and what should be returned exactly?

I would exceed the scope of this paper if I would get deeper into this discussion. If you have a closer interest, feel welcome to read the archives of both committees that may be accessed using the URLs given with both TCs.

To come back to the [SNS] use case, I can state two objectives here:

1. SNS will publish more than 90,000 subjects in the scope of environmental protection, and SNS will try to reference subjects that have been published by other authorities.
2. SNS will use a Web Service to publish subjects. It will be bound to a HTTP Get request to support a single URI for each subject, and the Web Service Description [WSDL] will contain a schema of a “Topic Map Fragment” to be returned by the service. This may be the [XTM]-DTD rewritten in XML Schema language, but it does not have to be.

Currently there is another TC, having its kick-off in March, that will cover related Web Services issues, but not especially dedicated to Topic Maps:

- OASIS Web Services for Remote Portals (WSRP) TC (<http://www.oasis-open.org/committees/wsrp/>)

The discussion is ongoing, but there will be more clearance till the date of the XML Europe conference in May, hopefully.

9. Conclusion

A project like [SNS] is living on the cutting edge of standardization. Every now and then you come to a point where you have to rack your own brains. You try to conform with the gist of what has been written - you cannot follow literally, as your special kind of usage has not been defined yet. In the world of XML, this is a typical every day situation, as XML really is *extensible*.

Since the XML Recommendation itself [XML] has been fixed in February, 1998, this standard provides a reliable and stable ground (the "second edition" only incorporated some known errata - not any meaningful modification). XML has not been modified, but it has been *used*, and may be "extended" by many applications. Some of them have become additional standards in their field, such as [XML Schema], [XPath], [ISO 13250], or [WSDL]. This

network of standards only defines the space that we are operating in, but it does not stipulate every this and that. I think that XML is expecting us to develop some creativity with each usage we are implementing.

Given an awareness of the different relevant facets of standardization, you may enrich this space by an independent usage example demonstrating that all the pieces really fit together.

Bibliography

- [AARHUS] UNECE, Environment and Human Settlements Division: Convention on Access to Information, Public Participation in Decision-making and Access to Justice in Environmental Matters (1998). <http://www.unece.org/env/pp/>
- [BA 2000] Bandholtz, T. / Börs, R. / Rüther, M. (2000): The German Environmental Information Network (GEIN). In: UI 2000
- [BA 2001] Bandholtz, T. (2001): 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>
- [BBN 01] Biezunski, Bryan, Newcomb: "Differences between XTM 1.0 and the HyTime-based meta-dtd". <http://www.y12.doe.gov/sgml/sc34/document/0277.htm>
- [BRYAN 01] Martin Bryan: XML Schema for ISO 13250 Topic Maps. Proposed Syntax 29th January 2001. <http://www.diffuse.org/TopicMaps/schema.html>
- [CDS 01] CDS and e-EIONET Work Conference 2001: Building environmental web services. <http://www.eionet.eu.int/events/cdsittag2001>
- [DOM] W3C Architecture Domain: Document Object Model (DOM). <http://www.w3.org/DOM/>
- [GEIN] Federal Environmental Agency of Germany ("Umweltbundesamt"): German Environmental Information Network (GEIN). <http://www.gein.de/>
- [GML] Open Gis Consortium: Geography Markup Language (GML) v1.0. <http://www.open-gis.org/info/techno/specs/00-029/GML.html>
- [ht://Dig] The ht://Dig Group: ht://Dig WWW Search Engine Software. <http://www.htdig.org/>
- [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>
- [LOCODE] UN/CEFACT: LOCODE - Code for Trade and Transport Locations. <http://www.unece.org/cefact/rec/rec16en.htm>
- [SNS] Federal Environmental Agency of Germany, SchlumbergerSema: Semantic Network Services. Research Project UFOPLAN-Ref. No. 20111612, promoted by BMU/UBA, Germany
- [SOAP] W3C Architecture Domain, Web Services Activity: SOAP Version 1.2 Part 0: Primer. <http://www.w3.org/TR/2001/WD-soap12-part0-20011217/>
- [Tamino] Software AG: Tamino XML Server. <http://www.softwareag.com/tamino/>
- [UI 2000] Computer Science for Environmental Protection '00. Environmental Information for Planning, Politics and the Public.. A.B. Cremers, Klaus Greve (eds.) ("Umweltinformatik aktuell", Band 26). Marburg 2000.
- [URI] IETF Network Working Group, T. Berners-Lee: Uniform Resource Identifiers (URI): Generic Syntax. <http://www.ietf.org/rfc/rfc2396.txt>

- [VAT 02] Bernard Vatant: OASIS Topic Maps Published Subjects TC Deliverables. 1. Documentation of Published Subjects - Requirements and Recommendations. <http://www.oasis-open.org/committees/tm-pub-subj/docs/recommendations/psdoc.htm>.
- [Web Services] W3C Architecture Domain: Web Services Activity. <http://www.w3.org/2002/ws/>
- [WSDL] W3C Architecture Domain, Web Services Activity: Web Services Description Language (WSDL) 1.1. W3C Note 15 March 2001. <http://www.w3.org/TR/wsdl>
- [XML] W3C: Extensible Markup Language (XML) 1.0 (Second Edition), W3C Recommendation 6 October 2000. <http://www.w3.org/TR/2000/REC-xml-20001006>
- [XML Query] W3C Architecture Domain: XML Query. <http://www.w3.org/XML/Query>
- [XML Schema] W3C Architecture Domain: XML Schema. <http://www.w3.org/XML/Schema/>
- [XTME] SchlumbergerSema: The XML Topic Map Engine (XTME). XML Network White Paper 2002-02
- [XPath] W3C: XML Path Language (XPath) Version 1.0, W3C Recommendation 16 November 1999. <http://www.w3.org/TR/xpath>
- [XSLT] W3C Style activity: XSL Transformations (XSLT) Version 1.0. W3C Recommendation 16 November 1999 (XSLT). <http://www.w3.org/TR/xslt>
- [XTM] Members of the TopicMaps.Org Authoring Group: XML Topic Maps (XTM) 1.0. TopicMaps.Org Specification. <http://www.topicmaps.org/xtm/index.html>

Biography

Wednesday, 22 May 11.00

Thomas **Bandholtz**

XML Network
SchlumbergerSema
Cologne
Germany
Email: TBandholtz@slb.com

Thomas works as a Senior Consultant and Project Manager for SchlumbergerSema, formerly Sema Group. He is also head of the SchlumbergerSema internal XML Network. He has been working on XML-based Thesauri and Gazetteers since years, using Topic Maps since 2000. Currently he is managing a R&D project developing a governmental Portal within the scope of environmental protection in Germany.

He is a member (prospective in March, 2002) of the OASIS Topic Maps Published Subjects Technical Committee, and a member of the OASIS Topic Maps Published Subjects for Geography and Languages (GeoLang) TC.