

GEIN 2000 and beyond: Environmental Information in the “Semantic Web”

1. Workshop "Environmental Markup Language (EML)"
Thematic Group 4.6.1 "Informatics in Environmental Protection"
German Informatics Society (GI.eV)
Berlin, Humboldt University, 1999

<http://schlemmer.wiwi.hu-berlin.de/xml-eml.nsf/Workshop1EMLIndex>

Thomas Bandholtz, Sema Group GmbH

Abstract

GEIN (German Environmental Information Network) is an XML-based Web directory that is being developed for the Federal Environmental Agency in Germany. This report explains, in which way GEIN sees itself as a contribution to a global *Semantic Web*. It further discusses the basic concepts and some critical details of the usage of RDF vs. “plain” XML in Web meta-data management. The information model and the namespace of the GEIN-“profile” may serve as a first foundation for the Environmental Markup Language to be developed.

Content

1 Introduction.....	74
2 What is the “Semantic Web”?.....	74
3 XML in the German Environmental Information Network.....	78
3.1 Information Brokerage.....	79
3.2 XML Document Types in GEIN.....	81
3.3 The GEIN Index and RDF.....	88
4 “AnyML” and the Complexity of Environmental Information.....	92
4.1 Special Fields of Environmental Information.....	93
4.2 Integrating MLs and the Use of Namespaces.....	94
4.3 Modeling of Measurement Networks.....	95
5 Conclusions.....	97
6 Literature.....	98

1 Introduction

As we were in the middle of designing a comprehensive Web index for environmental information in XML, we got surprised by the first announcement of „EML'99“ while working on the project German Environmental Information Network (GEIN) in the late summer of 1999. We felt competent enough to turn towards the EML initiative in order to enhance this initiative with our advanced meta-data approach.

From the beginning, EML and GEIN completed each other in an ideal way, whereas EML came primarily from the approach of corporate environmental information systems and developed especially environmental data by itself. But GEIN has nationwide meta-data about publicly available environmental information from offices and public institutions.

Reading the following: “The goal is the discussion and development of an initial draft of an XML-based language for the description of environmental data”, one can say: GEIN had already provided part of that goal at this time.

An explicit analysis of the current status of environmental data to be used and a strategic work-up with the global concept of the semantic Web were necessary as well as the discussion of the role of XML.

2 What is the “Semantic Web”?

Probably every reader knows these sentences from other articles and books containing various examples of “readable” XML notations. We do not want to add anything to this, but we would like to concentrate on the “readability” in a more detailed way.

The various publications quote the World Wide Web Consortium (W3C) about XML very frequently, if there are considerations about the *definition* of the Extensible Markup Language. Whereas their strategic motivation is mentioned only very little, this it can be found in W3C under “Design Issues – Architectural and philosophical points” (W3C Design Issues 1999).

Tim Berners-Lee, director of W3C and one of the “inventors of the WWW” (initially at CERN, today Principal Research Scientist at MIT Laboratory for Computer Science), explains “the thinking behind the specifications” there. These reports are directed towards a “technical community, to explain reasons, provide a framework to gain consistency for future developments, and avoid repeated discussion once resolved”.

The visions which need to be realized by using XML are explained in a series of articles under the title “A roadmap to the Semantic Web”. These articles were published beginning in 1998 until now. According to these articles, the Web as an in-

formation space was not only developed for the human-to-human-communication.

Nevertheless, a standard is spread with HTML which supports primarily information readable by people. The approach of the Semantic Web develops languages, however in order to express information in a format readable by machines, as well.

Using XML, the Web becomes potentially a global database: “The Semantic Web is a web of data, in some ways like a global database.” This database needs semantic descriptions based on a general model on a high level of abstraction. This general model is the Resource Description Framework (RDF). This global database needs furthermore a query language, which Berners-Lee still connects to RDF very often: “It is clearly important that the query language is defined in terms of RDF logic.”

The Semantic Web can hereby work as a distributed net of diversely structured data. However, these structures are explicit with XML. The Resource Description Framework (RDF) allows the presentation of all the data in a distributed and global catalogue. The query language (XML Query Language, XQL) allows queries against this catalogue as well as queries against the databases described in the catalogues.

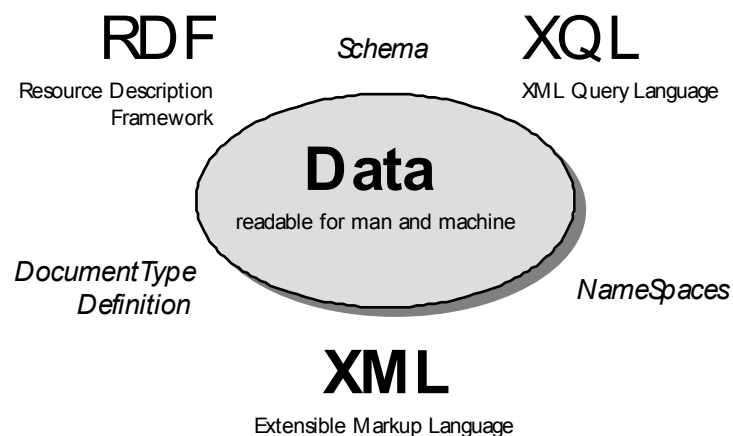


Figure 1: Components of the Semantic Web

XML, RDF and XQL form the language basis for the Semantic Web (supplemented by numerous special standards like XSL, XSLT, XPointer and Xlink). Whereas the necessity of a query language seems to make sense, the purpose of RDF appears not

in a very clear way, since catalogue information can be designed in XML from the beginning on.

In “Why is the RDF model not exactly the XML model” (1999) the author Berners-Lee explains the necessity of RDF with the necessity of a definite differentiation between – one could say: data and meta-data. The “example with jam” should be quoted at this point:

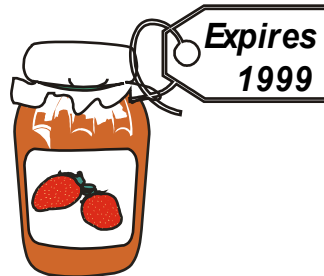


Figure 1: “What expires? The label, or the jam?”

At the same time the label is easily readable by humans. Everyone would understand that the jam is expired and not the label if one would see the jar (the “document”) on a shelf in the basement. This is actually due to a convention in the head of the observer what leads to the correct interpretation of the label (as a statement about the “document”). Those structures are not very easily readable by machines. One could say, RDF presents the machine-readable convention for labels of every kind.

Yet a “general RDF engine” needs to be provided understanding every of those labels spontaneously, since all are based on the same semantic fundamentals. Every statement on the label has got the structure of a triple in the following format:

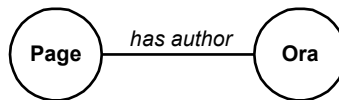


Figure 3: Triple “Ora is the author of page”

This example is among the most favorite ones at W3C. It stands for: “Ora is the author of page”, whereas page means Web page and Ora² stands for a person who is the author of this Web page. That is strongly suggestive of a relation in the Entity Relationship Model and indeed every RDF-statement can be presented in an ER diagram.

The RDF-machine can parse every label because it can be sure that every statement has the format of such triples in RDF. Therefore there is exactly one single valid expression, i.e. serialization of this triple in RDF

```
<rdf:description about="page">
  <author>Ora</author>
</rdf:description>
```

whereas the same statement can be serialized in various ways in XML. The following four (valid) examples are taken from the design issues:

```
<author>  <uri>page</uri>
  <name>Ora</name>
</author>
```

To receive more statements about „page“, for instance to mention its title, another tag must be introduced within the <rdf:description> block like

```
<title>Simple RDF-sample</title>.
```

The Dublin Core Metadata Initiative (<http://purl.org/DC/>) with their “Simple Content Description Model for Electronic Resources” was one of the first groups of users who applied RDF for their own purposes. This is a data record which goes back to bibliographic information and which slightly enlarges it (dc:type and dc:format) in order to be able to describe other media such as books, as well. Meanwhile, the English government had indexed its Web site (<http://www.open.gov.uk/>) completely with RDF with the utilization of Dublin Core.

```
<rdf:RDF xmlns:rdf="http://w3.org/TR/1999/PR-rdf-syntax-19990105#"
  xmlns:dc="http://purl.org/dc/elements/1.0/#">
  <rdf:Description rdf:about="http://www.open.gov.uk"
    dc:creator="Neil Pawley"
    dc:title="open.gov.uk home page"
    dc:subject="front page, introduction, welcome"
    dc:description="the open.gov.uk introduction page welcomes the web user
and provides a single entry point to uk public sector information on the
internet."
    dc:publisher="CCTA"
    dc:date="1999-09-01"
    dc:type="text"
    dc:format="text/xhtml"
    dc:language="en" />
</rdf:RDF>
```

According to the point of view of the author, this usage of RDF is actually valid XML (attributive style), but it is not in conformity with RDF. Therefore the “general RDF machine” is going to fail. Those problems are discussed in chapter 3.2.3.

3 XML in the German Environmental Information Network

The first contributions to the „roadmap to the Semantic Web“ were published just then as the project “GEIN 2000” began in the summer of 1998. Even the XML standard was only a few months old. Certainly, it was not possible to use both standards as a basis for the plan of the project at that time.



Figure 4: GEIN surrounded by the logos of the information providers (Federation and Federal States)

In any case, it was unquestionably dealt with the information brokerage and data exchange via internet.

At first, there was an order to audit the ANSI standard “Z39.50” which is established in all the libraries worldwide. Moreover it allows mechanisms similar to namespaces in XML using so-called profiles. Therefore a “GEIN-Profile” should be defined, if the suitability of Z39.50 (soon called only “Z” within the project slang) can be proven.

However, “Z” is without doubt a pre-Internet standard. This can also be seen in the structure of the communication-protocols and in the hypothesis, that not only hyperlinks, but also professional information has to be communicated by themselves.

The very expensive system analysis of current information offerings in the winter of 98/99 showed, that a major part of information was already available on the Internet. This information is actually visualizing itself and this in a better way than with a general possible broker function. The trend was in general directed towards the fact, that all edited and actively published environmental information should be present in the Web.

3.1 Information Brokerage

bro-ker ['broukər] *s* **1.** *Br.* Altwarenhändler *m*, Trödler *m*. **2.** *econ.* (*a.* Börsen-, Kurs-, Wechsel)Makler *m*, (*a.* Heirats)Vermittler *m*, Mittelsmann *m*; *s*'s note Schlußschein *m*.
bro-ker-age *s* **1.** Maklerberuf *m*, -geschäft *n*. **2.** Maklergebühr *f*, Courtagage *f*; buying *v* Einkaufsprovision *f*.

Figure 4: German translation of *broker*
 (source: Langenscheidt Grosswörterbuch)

A concentration on the brokerage as a core business resulted for GEIN out of that. This means a broker's call between the interests and inquiries of the users (“interested laymen” and “professionals”) on one hand and the over 50.000 Web pages (source 1999) of the information offered on the other hand.

The search engines such as Alta Vista or Yahoo count as state-of-the-art but are already seen as “classical” ones today. It is preferred to speak of “portals” today. The latter very much reflects a concentration on a certain clientele of information providers. This fact is actually helpful for the concept of GEIN. Nevertheless, the emphasis is mostly set on marketing aspects. Marketing aspects mainly focus on advertisement and especially on “personalization” of the visitor who therefore voluntarily provide information about their way of shopping.

This “personalization” should not take place in GEIN: Users performing retrieval will remain anonymous, because nobody wants an unknown person to enjoy one's own attempts.

Therefore we did concentrate on overcoming of borders of classical search techniques in three main areas:

1. Focus on the sites of the affiliated information providers
2. Inclusion of “dynamic” information offers
3. Semantic differentiation between search conditions and search techniques.

As done in various existing search engines (like Harvest, Ultraseek, ht://Dig), the first problem can be solved by the introduction of a limited list of server to be searched. In contrast, the other two points mentioned above require a bigger effort and innovative start.

The problem of dynamically generated pages that cannot be indexed by search engines does not have anything to do with dynamics if one looks closer. Even the (HTML-) output of CGI and Perl scripts or Java Servlets is readable and can be used by crawlers if they can be references by a known URL: The actual problems consist of integrated search functions and menu systems of the Web sites because they have to be hidden in order to restrict the direct readability of the URLs. Therefore it happens very often that press releases can be found as ordinary, static HTML-documents on the server without any directly connected hyperlinks. The universal access is implemented by using a search form instead. The crawler of a search engine is capable of detecting the format written in the HTML code but it is not able to understand (or to guess) the specific logic with their encoded searching conditions and selection criteria - this is always individually and not explicitly. Therefore these unlinked sites are not recognized by the search engine.

GEIN is tackling that problem in a way, that the search condition itself is handed over to the corresponding server and the server is answering with the own procedure. An arranged protocol is certainly necessary and the realization of an appropriate interface. GEIN uses HTTP Post requests for that reason, questions and answers are exchanged in XML in that case. The cooperation is necessary for the offers of information without a doubt. However, practice showed that this is very motivating and can be solved with a rapid success.

The third aspect – the semantic differentiation of the search conditions and the search procedure – is undoubted one of the most complex. It is important that the search string is not simply used as a non understandable chain of characters. GEIN differentiates between the categories *topic*, *space* and *time* instead. The Federal Office for Environment provides a very reliable professional thesaurus that contains a defined vocabulary for the category *topic*. Together with GISU (the GIS project of the Federal Office for Environment) we developed a geographic thesaurus for the category *space* which contains names - and especially the relations of location in between – of more than 50.000 geographic objects together with an efficient index. Even a so called environmental calendar is going to be developed for the category

time which is responsible to store exposed events along with their date and duration. Probably this calendar has got the chance of becoming a thesaurus by itself. The category *time* can certainly be used as (any exact) point of time or as a time period. Time periods can be kept open in one direction (sooner or rather later as X).

3.2 XML Document Types in GEIN

The uses of HTML to represent meta-data are too primitively for those purposes. The head-attribute "<title>", as well as the meta-tags "keyword" and "description" do not allow neither the illustration of the semantic differentiation nor the communication of conditions to search. Additional meta-tags can actually be defined freely in general. However, HTML does not provide any methods in order to represent one's own semantics of applications in a reliable way.

On that point the discovery of XML and especially RDF was made for the project. An XML namespace was created instead of the initially planned Z39.50-profile, which has got its stable second version as "G2k-profile" today. The abbreviation "g2k" was simply created by bringing together the project name of GEIN 2000 and the well-known abbreviation "Y2K" for the year 2000 problem.

This profile regulates the indexing using the topic-space-time semantics according to the given thesauri as well as the communication between search conditions and lists of results.

As these points were set a trial, the next attempt was performed – at first experimental - to model further data in XML instead of modeling in a relational way. In the meantime, the vocabulary (the thesauri as well) can be completely found in XML. Even the necessary data for administration (for instance the configuration of affiliated information providers) are modeled and stored in XML.

By the way at that point we use the XML-database Tamino by Software AG (<http://www.softwareag.de/tamino>) which manages native XML, that means a relational or object oriented mapping is not necessary. The data can be accessed using the designated Document Object Model provided by W3C. The query language is XQL.

3.2.1 The Vocabulary

The thesauri represent a controlled vocabulary (in the sense of *controlled term*) for the semantic differentiation. The vocabulary is used in particular for the indexing of the environmental information to be provided.

3.2.1.1 UBA Thesaurus

From the content point of view, the UBA thesaurus is an already often used professional vocabulary provided by the Federal Environmental Agency (Umweltbundesamt 1999). At this point the linguistic structure cannot be discussed. The following example provides an insight in XML:

```

<THESEIN ino:id="1204">
<RISN>18741</RISN>
<ADATUM>950215</ADATUM>
<ART>9</ART>
<DATERF>900906</DATERF>
<ENDC>CA</ENDC>
<ENDUNG>.</ENDUNG>
<ENDUNG>E</ENDUNG>
<ENDUNG>EN</ENDUNG>
<ENDUNG>ES</ENDUNG>
<ENDUNG>S</ENDUNG>
<GESDRU>J</GESDRU>
<SWAF>ORGANOCHLORPESTIZID</SWAF>
<SWSRT>ORGANOCHLORPESTIZID</SWSRT>
<SWVF>Organochlorpestizid</SWVF>
<SWVFU>Organochlorpestizid</SWVFU>
<THSISN>00018741</THSISN>
<TYP>ND</TYP>
<USEISN>00006072</USEISN>
<USEISN>00028719</USEISN>
</THESEIN>

```

The structure and nomenclature that is used here was taken in a one by one from the original database export of UBA (Adabas) in order to assure a simple quality control of the data transformation.

3.2.1.2 Geo-Thesaurus Umwelt (GTU)

The Geo-thesaurus is a new development of GEIN in cooperation with the Geographic Information System Environment (**GISU**) of the Federal Office for Environment. The following example shows (shortened) two Geo-objects of a different type (community and mountains) which show an overlapping of locations (<CELL>36032</CELL>).

```

<GEOBJECT>
<GEOTYP>GEMEINDE</GEOTYP>
<GID>0577400012</GID>
<UID>GEMEINDE0577400012</UID>
<NAME>Borchen</NAME>
<CELL>35511</CELL>
<CELL>35512</CELL>
....
<CELL>36032</CELL>
....
<CELL>36293</CELL>
<CELL>36294</CELL>
</GEOBJECT>
<GEOBJECT >
  <GEOTYP>GEBIRGE</GEOTYP>
  <GID>10018</GID>
  <UID>GEBIRGE10018</UID>
  <NAME>Paderborner Hochfläche</NAME>
  <HOEHE>248</HOEHE>
  <CELL>36032</CELL>
</GEOBJECT>

```

This figure illustrates the major intention of the GTU: The illustration of relations of locations between geographic objects of various kinds which describe the related reference of environmental information. GTU is restricted to the scale 1:200.000 today, that means all smaller objects are not contained in the GTU, either because of the cartographic displacement and abstraction. But there are over 50.000 remaining objects which should be a sufficient number concerning the vocabulary and its consideration nationwide.

The <CELLS> point at an abstract grid with a fragmentation of 3X3 kilometers. A high dissolving fragmentation was calculated for every object. After this happened once, the started cells only were indexed with an ID, so that at the moment of the access the data it is not necessary to perform any calculations at all.

The further implementation showed, that even the access via the common cells is not workable anymore because there are often arising more than 1000 disjunctions. That leads to another derivative – the explicit storage of the intersections for every single object.

```

<intersection uid="GEMEINDE1306000056"
  typ="GEMEINDE" term="Parchim" cov="2"/>
<intersection uid="GEMEINDE1306018075"
  typ="GEMEINDE" term="Stralendorf" cov="2"/>
<intersection uid="GEMEINDE1306021028"
  typ="GEMEINDE" term="Granzin" cov="2" />
<intersection uid="GEMEINDE1306021050"
  typ="GEMEINDE" term="Lutheran" cov="6" />
<intersection uid="GEMTEIL63302"
  typ="GEMTEIL" term="Lancken" cov="3"/>
<intersection uid="GEMTEIL63303"
  typ="GEMTEIL" term="Beckendorf" cov="3" />
<intersection uid="LANDSCHAFT15846"
  typ="LANDSCHAFT" term="Mecklenburgische Seenplatte" cov="1" />
<intersection uid="NATURRAUM22"
  typ="NATURRAUM" term="Mecklenburgische Seenplatte" cov="1" />
<intersection uid="WASSEREINZUGSGEBIET592" typ="WASSEREINZUGSGE-
BIET" term="Elde-Müritz-Wasserstraße" cov="1" />
<intersection uid="KREIS13060" typ="KREIS" term="Parchim" cov="1" />

```

By the way, these are the intersections of the community Rom in Brandenburg. In this case, there are geo-objects shown that share at least one single common cell. “Cover” is a metric that specifies the extent of overlapping.

Therefore the computation of all overlappings of more than 50.000 objects is now reduced to one single database access.

A more detailed discussion about the purpose and the capability of this method does not belong here at this point. This example only shows a way of using XML to store those objects.

3.2.1.3 Environmental Calendar

GEIN used the standard ISO 8601 for the naming of points in time and time periods, that means for instance “1999-12-31-11T23:59” or “2000-01-01T00:00”. But also “1869” or “2000-03” would be valid examples.

The work with the thesauri led to the idea to use even exposed points in time in the sense of *controlled terms* in order to allow the user to ask questions like what happened “since Tschernobyl” without the necessity to remember the exact date of that event.

```

<EVENT language="de" type="Katastrophe" id="bv2400">
<NAME>Bodenversalzung</NAME>
<SINGLETIME year="-2400">2400 v.Chr.</SINGLETIME>
<DESCRIPTION>Etwa um diese Zeit werden aus dem Zweistromland
Mesopotamien große Schäden als Folge unzureichender Bewässer-
ungssysteme bzw. fehlender Entwässerungssysteme gemeldet.</DESCRIP-
TION>
</EVENT>
<EVENT language="de" type="Konferenz" id="5k1999">
<NAME>5. Klimaschutzkonferenz</NAME>
<STARTTIME year="1999">1999-10-25</STARTTIME>
<ENDTIME year="1999">1999-11-05</ENDTIME>
<DESCRIPTION>
Ziel der Konferenz in Bonn ist es, die 1998 getroffenen Festlegung in
<LINKTO target="ba1998">Buenos Aires</LINKTO>
weiterzuführen. Es sollen Regelungen und Instrumente für die Verminder-
ung der Treibhausgase erarbeitet werden.
</DESCRIPTION>
</EVENT>

```

Certainly, the environmental calendar has not the meaning of a thesaurus on a large scale. Though it defines the use of language synonyms for time data and therefore makes a contribution to the semantics of the vocabulary. This approach is certainly capable of being developed in the future.

3.2.2 The GEIN-Index

This index suggests a <description> for every offered information within GEIN. This information consists of:

1. The typical search engine result information (URL, title and abstract)
2. Topic-space-time references on the vocabulary described above
3. Additional references to the environmental classification of the Federal Environmental Agency (<g2k:class> and <g2k:portal>)

```

<?xml version='1.0' encoding="iso-8859-1"?>
<g2k:G2K xmlns:g2k="http://www.gein.de/2000/profile-11#" >
  <g2k:description uri="http://www.any-mu.de/doc.htm"
    xml:lang="de" date="2000-02-02">
    <g2k:portal>AB10</g2k:portal>
    <g2k:class>AB10</g2k:class>
    <g2k:title> Titel des Dokuments </g2k:title>
    <g2k:abstract> unformatierte Kurzbeschreibung des Dokuments
    </g2k:abstract>
    <g2k:topic uid="28753" term="Luftverunreinigung" rank="10"/>
    <g2k:topic uid="16963" term="Meßstellennetz" rank="5"/>
    <g2k:area uid="GEMEINDE1100000000" term="Berlin"
      type="GEMEINDE" rank="8"/>
    <g2k:area uid="NATURRAUM12" term="Havelland"
      type="NATURRAUM" rank="7"/>
    <g2k:time event="cal/4711" from="1999-12-31" to="2000-01-01"/>
    <g2k:time at="1990" />
  </g2k:description>
</g2k:G2K>

```

Detailed explanations can be found in “g2k-Profil” (see 3.2.3 “The GEIN-Namespace”). Here you can get some hints to essential details.

The elements <topic>, <area> and <time> obviously refer to the explained vocabulary in the described differentiation topic-area-time. The identifier and the ID (for <time> optional) are provided for each item there.

The contained redundancy can be seen as a concession directed towards a practicable application of this index.

One has to understand in that case that the identifier mentioned in the term-attribute is not necessarily uniquely defined! That counts especially for the geographic names in <area>.

The used *uid* is unique within the whole GEIN-vocabulary but it is not understandable without further accesses. Certainly, there is an implemented HTTP-Service by GEIN which is able to answer with the appropriate vocabulary-definitions after passing the uid. This actually seemed not appropriate for the netwide communication. The only phrase of UID would have provoked a critical net load under the current conditions on one hand and a spontaneous access to the index would not have been possible on the other hand since the communication with the HTTP-service would have been necessary to implement.

The solution provided allows an access using a two steps mechanism:

1. the use of term by itself whereas certain ambiguity similar to a search of a free text has to be taken into account
2. the use of uid which excludes ambiguity and allows access to further attributes (lower-and non lower terms, synonyms, terms for locations and so on).

3.2.3 The GEIN Namespace

XML-tags like <GEOOBJECT>, <EVENT> and so on can certainly be defined with different semantics in various contexts. If the context can be seen as internal data which are exclusively used within GEIN the context can be defined as an implicit one. Within GEIN, <EVENT> is uniquely preserved for an event described in the environmental calendar.

This uniqueness will be lost as soon as data is exchanged or generally published. This applies especially to the GEIN index explained in the following, which must be considered a public Web index in the sense of the Semantic Web.

Even if neither an implemented “general RDF machine” nor an equivalent XML broker are known at this time, we considered it advisable, to secure all precious data against any kind of ambivalence.

This will be guaranteed by the „g2k“-namespace which is documented in a “profile”(GEIN 2000 Profile) readable by humans.

This namespace will be referenced in the beginning of each document, for example by the line:

```
<g2k:G2K xmlns:g2k="http://www.gein.de/2000/profile-11#" >
```

The applied URL shall primarily function as a globally unique key. The W3C specification does not require a simultaneous definition for the namespace at this point, neither as a formal XML scheme nor as content explanation. However, it is considered practical to insert such a document at this place in order to provide immediate access to the underlying semantics to anyone dealing with a “g2k” document in any context.

The uniqueness of the single tags is achieved by starting with the prefix “g2k” followed by a colon.

The globally ambivalent <topic> turns into a globally unique tag by using <g2k:topic>. It must be said that this is only true in the condition that no other application in the world is using the abbreviation g2k as well. In connection with the URL which is assigned to the attribute g2k:xmlns uniqueness can still be achieved without leaving the context of the document. This uniqueness is based on the internationally controlled authorization of domain names in the internet: www.gein.de can only exist once.

3.2.4 Further Documents Types

Besides the vocabulary, also administration data are modeled and stored in XML as for example the intern index of information providers and the description of interface configurations and the appropriate administrators.

Without doubts, this kind of things could as well be modeled and stored relationally. This is also valid for the vocabulary and the indexing of the provided information. But the problem with the relational model is that it does not represent an exchange format at the same time and therefore it stays ambivalent as long as the data model and the physical storage type are not known. These problems could be avoided; however, we preferred for GEIN for reasons of uniformity to administrate all data in XML. Precise examples shall not be given here since they are not comprehensible without detailed explanation: they are designed only for intern use.

3.3 The GEIN Index and RDF

The attentive reader will have noticed that the examples of chapter 3.2. do not contain any RDF notations. Here we had to undertake a tactic turn – in the last minute.

In fall 1999 we presented our model for examination to the “rdf-interest” forum of the W3C (<http://www-rdf-interest@w3.org>). It turned out that the complexity of our approach could not uniquely be approved with the already stable fundamentals of the still very young standard. Although the “official” approving RDF-parser (SirPac) accepted our application of the standard as flawless the members of the forum believed to see a wrong approach in our semantic which led to a discussion about a possible mistake by the approving parser. The project was under time pressure but the discussion in the W3C forum was extended to more and more basic topics and deviated more and more from a concrete and binding recommendation for our problem.

In this difficult situation (not unusual for research projects) we decided to leave the obviously unstable RDF approach. The G2K profile commented:

“During debates about the standardization context it turned out that RDF does not yet contain the required stability of definitions. Some structures are still at issue. Often, suggestions for application containing unexpected use of elementary regulations are made which in a first place must be judged according to its consequences. Besides this, a discussion for a ‘simplified syntax’ was started.

It would be contra productive to burden our intern communication with this (as such necessary and delightful) development continuously. That's why from now on we exclusively use XML with the (known) g2k namespace for GEIN. XML alone (without RDF methods) allows the user a big definition autonomy for his own procedures so that the stability could only be limited by the practical value of the regulations.

Anyhow, the future role of RDF is continuously considered as strategic. That is why the GEIN broker (at a later point) will provide a HTTP service which formats the XML documents into RDF. This filter can keep up with the further development of the RDF and can therefore continuously enable compatibility to the standard without interfering into the intern communication.” (GEIN 2000 Profile)

How did that happen? Lets first take a look at the abstract structure of the G2K index as an entity-relationship-model, which is beyond any kind of serialization in XML.

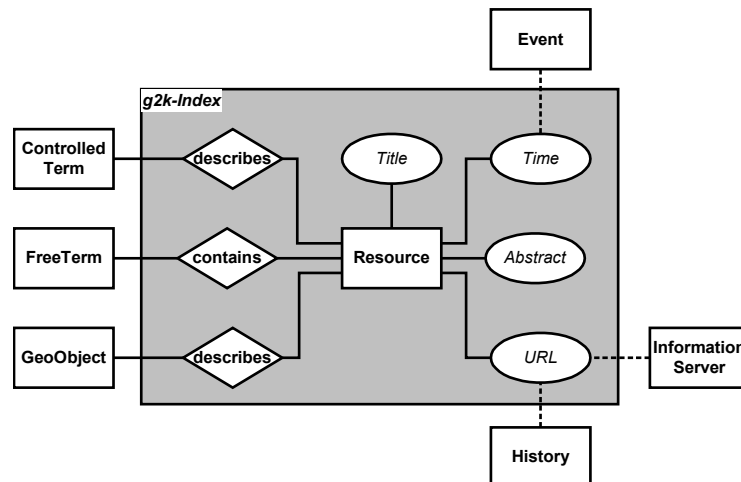


Figure 6: The GEIN index as an entity-relationship-model

At a first sight, this could be broken up into a series of RDF triples, as for example:

GeoObject – is spatial relation to – resource

and so on. We have already remarked that the terms in our vocabulary are not always unique. This is especially true for geographical names. For this reason we decided to indicate the (unique) *uid* aside of the term and, for quicker orientation, in the case of GeoObjects also the type (as to say rural commune, river, nature reserve and so on). For relational procedures this corresponds to an application appropriate de-normalization.

Additionally we had decided to assign a ranking to each catchword, which means a number between 1 and 10, which indicates the importance of the catchword for the document.

Lets take the example “Hessen”. It is not only the name of the Federal State but also of a rural commune. A statement (in prose) to the space relation would therefore contain the following information.

The Bundesland Hessen (see the uid b14711) is a very significant (10) space relation of the document page.

Lets try to express this as a triple according to the model of

Ora Lassila is the author of page

we will get a very complicated structure in the following form:

[(“Hessen“,“b14711“,“Land“)10] is a g2k:area of „page“

This would approximately correspond to a situation with various Ora Lassilas, whereby it could also be the name of a machine that generated the document.

We suggested at this point to put this down into RDF:

```
<rdf:Description about = "page">
  <g2k:area
    term = "Hessen"
    geotype = "Land"
    uid ="b14711"
    rank =10
  />
```

```
</rdf:Description>
```

We found this readable and suitable to any application interest. Furthermore it still is valid XML. But, it violates the triple structure of RDF. This structure would only allow:

```
<rdf:Description about = "page">
  <g2k:area = "Hessen" />
</rdf:Description>
```

□ which is not a unique information, or:

```
<rdf:Description about = "page">
  <g2k:area = "b14711" />
</rdf:Description>
```

□ which alone is not comprehensible and does not even contain ranking information.

RDF allows however the reference from one triple into other triples with the help of `rdf:resource`. Our complete statement could then be broken up into a series of single statements.

1.

```
<rdf:Description about = "page">
  <g2k:area rdf:resource="#007"/>
</rdf:Description>
```

That means: A space relation (`<g2k:area>`) of "page" is situated in `rdf:resource „#007“` (artificial key).

2.

```
<rdf:Description rdf:ID="#007">
  <g2k:term rdf:resource="b14711"/>
  <g2k:rank> 10 </g2k:rank>
</rdf:Description>
```

This spatial relation is described more precisely in „b14711“, we will assign it a significance (rank) of 10. This is information about the occurrence of "b14711" in the document "page" (connected with the previous information through the `rdf:ID`).

We still do not know what „b14711“ actually means. Therefore the following information is needed:

3.

```

<rdf:Description rdf:ID=" b14711">
  <g2k:term> Hessen </g2k:term>
  <g2k:type> Land </g2k:type>
</rdf:Description>
</rdf:RDF>

```

Here, at last, it is explained that we deal with the Federal State Hessen. This third information is no longer talking about space relation in the document but it describes the object with uid "b14711" for itself, so that information about other documents can as well indicate to this description.

In comparison to the first version which concludes all information in a single description-block we do get a very confusing working model. This model can be parsed without problems using a general RDF engine. But we doubt the content of this information to be understandable in an efficient way. It has to be considered that every document has not only a few space relations, but also moreover relations to the used topic and the time.

The index contributes primarily the communication in the environmental information network. It seems realistic to design a specialized semantic working model within this domain that is understandable for everyone and able to be worked with in an efficient manner. This attitude is accepted within the RDF-interest-forum:

"Many, many people have decided that using XML with XML Schema is a better solution than using RDF for their information. RDF does have benefits for certain form of information, like simple metadata (Title, Author, Copyright, etc.), and for describing web-sites (about each, etc.), but there is **always** a more efficient representation in plain XML. The disadvantage of XML, even with the XML Schema, is that everyone has to figure out how to parse your particular representation. With a general RDF engine, you don't." Perry A. Caro, Adobe.

The necessity to understand that working model results for the external users without any doubt. But we do doubt that the external users understand the content of the formal clear RDF-notation more easily after they have parsed this.

4 "AnyML" and the Complexity of Environmental Information

After this detailed report about the application of XML in GEIN a few comments concerning the concept of a general EML are mentioned now.

GEIN is basically limited to the modeling and communication of meta-information, where a complex vocabulary gets involved. The avoidance of ambivalences within the vocabulary is very demanding.

GEIN “acts” as a broker with URLs, which represents Web pages showing environmental information in words and pictures. But GEIN is not dealing with modeling and communication of environmental information by itself.

Concerning this a few basic and exemplary thoughts are following.

4.1 Special Fields of Environmental Information

The first question is: Is there actually a special field “environment”? One can doubt that⁴. The specific subject matter and problems are obvious taking other (undoubted) special fields into account like chemistry or mathematics: CML, the “Chemical Markup Language” was designed based on the necessity, a XML (originally SGML) notation for the description of molecular structures. The Mathematical Markup Language (MathML) (W3C MathML 2000) is working on the notation of differential, integral and so on.

But what is actually the field of environmental information?

chemistry regional planning
 agriculture geology
 physics medicine metrology
 process engineering politics
 biology transportation geography
 library science

Figure 6: Examples of professional fields dealing with environmental protection

We obviously deal with a cross-sectional discipline. If we would like to describe the reaction of a chemical substance as an environmental influence for example we have to consider the fact that arranged professional procedures for the nomenclature and description of this chemical substance already exist. We come across with a known problem of standardization and not to invent existing and approved procedures again if possible. In “Attribute Set Development Guide” of the Z39.50 agency there you can find a quotation of a contribution to a discussion using a fictitious “Galaxy”-Profile to describe the desired procedure:

“Whoever defines the Galaxy attribute set will be directed for each potential access point to look at existing attribute sets to see if there is already an appropriate attrib-

ute for that access point, and define an attribute for that access point only if there is not. This means (in term of the new architecture) if the developer of the Galaxy set needs 'title' and if the semantics of the bib-2 title match the desired semantics of the Galaxy title, then no title Abstract attribute is needed for the Galaxy attribute set (the Galaxy *Profile* will describe how title is to be searched, i.e. using bib-2; and the Galaxy attribute set may include a comment explaining why it doesn't contain a title attribute)." (Percivall 1998)

The same problem is also known in the world of EDI. If every group of users invents their own set of attributes once again „from the scratch“ in a new way, as even no one was thinking about related procedures, soon there will be countless application-“standards” not compatible with each other. Therefore they would not meet the goal of a real standardization.

4.2 Integrating MLs and the Use of Namespaces

Namespaces in XML represent an appropriate method to integrate attributes (or XML-elements) from different worlds with each other. Chapter 3.2.3. already referred to the “g2k”-namespace used for GEIN. Several namespaces can be used in the same document with the same procedures without creating a chaos even if the attributes should sound like the same. This can be seen with the following example of W3C:

```
<?xml version="1.0"?>
<!-- both namespace prefixes are available throughout -->
<bk:book xmlns:bk='urn:loc.gov:books'
         xmlns:isbn='urn:ISBN:0-395-36341-6'>
  <bk:title>Cheaper by the Dozen</bk:title>
  <isbn:number>1568491379</isbn:number>
</bk:book>
```

The namespaces “bk” and “isbn” are explicitly used side by side without having any conflicts to solve.

Even (fictitiously) <mathml:power>, <energie: power> and <political:power> could be used in one statement in the same manner.

4.3 Modeling of Measurement Networks

The modeling of reusable document types is going to be the major assignment for the development of an EML. Those have quite a resemblance with object-classes. Measuring sequences of measuring data in the field of environmental observation are one of the most frequent applications for the exchange of environmental data.

First we should have a look at two of today's used formats, the so called "CSV"-format of the administration-agreement data exchange Bund-Länder, and one EDI-format used in Sachsen-Anhalt.

CSV is an abbreviation arising from comma separated variables and represents the most simple pattern of an exchange format. The following example shows a specific field of application.

```
station;parameter;date;
Value0030;Value0100;Value0130;Value0200;Value0230;Value0300;Value0330;Value0400;Value04
30;
Value0500;Value0530;Value0600;Value0630;Value0700;Value0730;Value0800;Value0830;Value09
00;
Value0930;Value1000;Value1030;Value1100;Value1130;Value1200;Value1230;Value1300;Value13
30;
Value1400;Value1430;Value1500;Value1530;Value1600;Value1630;Value1700;Value1730;Value18
00;
Value1830;Value1900;Value1930;Value2000;Value2030;Value2100;Value2130;Value2200;Value22
30;
Value2300;Value2330;Value2400
'DEBB050';'ozone';'19990901';
32;28;30;34;36;32;32;30;27;28;30;29;27;25;24;28;27;-999;34;45;49;58;52;48;30;37;30;44;57;50;71;
61;67;71;69;66;56;57;52;49;51;40;52;43;42;24;21;21
```

The paragraph above is the header of such a file. It gives a sequence first in which the real file can follow in the body of the file. The second paragraph contains the corresponding information. Every exchanged file contains various paragraphs. The semantics appears in general self-explaining. If you look closer it seems however that a high standard of implicit information needs to be guaranteed.

Firstly, all the variables are not really comma-separated but separated through a semicolon. Why? It was obviously counted on the transfer of decimals whereas the comma is used to separate decimals in Germany what means that it can not be taken for the separation of variables. What does "Wert0030" mean? If one looks closer it can be guessed that this stands for a value of a one half of an hour of the day

1.9.1999. But which value? A raw value? An average value? A maximum? The measuring station is obviously called 'DEBB050'. Where is it located? Which part of the measurement network? Which method is used to measure it? Which inquiry office can answer this question?

This exchange format is *working* astonishingly. But it can only work within a fine-meshed procedure, which does not even contain a reference.

The second example is a format which follows EDI-instructions. It was developed in the project Vision-Environment primarily for the exchange between the environmental ministry and regional administration in Sachsen-Anhalt in 1994 and is concerned about explicit information. Furthermore it allows a "fall over" of the way of observation between the columns of time or a single station and time period over all stations.

```

EDI
UNA:+.? '
UNB+UNOB:2+LAU-LUESA:HALLE+LSA-
MU:MAGDEBURG+981105:1526+0'
UNH+0+MESS1:1:931:MU:1'
MNZ+luesa'
MRE+o3+Stunde+Mittel+N+1998:11:05:15:-+1998:11:05:15:-'
MST+BROC'
MWT+1998:11:05:15:-+62'
MST+DECN'
MWT+1998:11:05:15:-+41'
MST+DEVC'
MWT+1998:11:05:15:-+12'
MST+DGCW'
....
UNT+42+0'
UNZ+1'

```

Not all possible questions can be answered explicitly even there. But the partners participating are known as well as the fact that they are mean values of hours. The time is mentioned close to the values. Therefore any single values can be handed in later or can be overwritten.

The general approved format like EML has to offer, should continue this paragraph containing the whole information in an explicit manner. This can be done via hyperlinks to the appropriate URL in the time period of the internet.

Finally the major entities are concluded for the synchronized measurement networks which have to be considered in the corresponding types of documents:

Operator	name and URL of the organization
Measurement Network	General name, link to the description
Station	name, link to position and description
Measuring components	measured indicator, method
Interval	time interval of the reported values
Aggregate	for example mean, min, max, percentile
Measuring unit	in which the values are reported
Time	validity of the value concerning the interval
Value	the "measured value", special cases optional

5 Conclusions

This is the presentation of nine short key phrases which from my point of view show the most important aspects of EML. It could not be dealt with all mentioned phrases in the same detailed way:

1. "EML" does not work wonders.
2. "EML" does not replace a professional agreement.
3. "EML" is not a new Markup-Language.
4. "EML" = XML + NameSpace(s) + types of documents.
5. "EML" is explicit and not a bit-scrooge.
6. "EML" can be used in a flexible way.
7. "EML"-documents communicate in the semantic web (RDF).
8. "EML" documents can be queries using XQL.
9. "EML" is international.

6 Literature

- Bandholtz, T. (1994a): Fachübergreifende Integration von Umweltdaten, in: Güttler, R./Geiger, W. (Hrsg.): Integration von Umweltdaten, 2. Workshop, Schloß Dagstuhl, Marburg, S. 49-65
- Bandholtz, T. (1994b): Interaktive Visualisierung mit Vision Umwelt, in: Denzer, R./Güttler, R./Deutsch, H. (Hrsg.): Visualisierung von Umweltdaten, 2. Workshop, Schloß Dagstuhl, Marburg, S. 21-37
- Berners-Lee, T. (1998): Semantic Web Road map,
<http://www.w3.org/DesignIssues/Sematic.html>
- GEIN 2000 Profile (2000): Profile,
<http://www.gein.de/2000/profile-11.htm>
- Open Module Foundation (OMF) (Eds.) (1999): Chemical Markup Language (CML),
<http://www.xml-cml.org>
- Percivall, G. (1998): Attribute Set Developers Guide, annotated outline, 18 Sep. 1998,
http://harp.gsfc.nasa.gov/~eric/attr_set_developers_guide.html
 oder/or
<http://lcweb.loc.gov/z3950/agency/attrarch/attrarch.html>
- Umweltbundesamt (1999): Umwettesaurus,
<http://www.umweltbundesamt.de/uba-datenbanken/thes.htm>
- World Wide Web Consortium (W3C) (Eds.) (1999): Design Issues: Architectural and philosophical points, <http://www.w3.org/DesignIssues/Overview.html>
- World Wide Web Consortium (W3C) (Eds.) (2000): Mathematical Markup Language (MathML), <http://www.w3.org/Math>
 oder/or
<http://xerxes.thphy.uni-duesseldorf.de/~vieth/subjects/www/w3c-specs/REC-MathML/Overview.html>
- World Wide Web Consortium (W3C) (Eds.) (1999): Namespaces in XML,
<http://www.w3.org/TR/1999/REC-xml-names-19990114/>