

The German Environmental Information Network (GEIN)

Thomas Bandholtz¹, Richard Bös² and Maria Rüter³

Abstract

Das Umweltinformationsnetz Deutschland (GEIN) vereint ein weites Spektrum an Umweltinformation von öffentlichen Einrichtungen und Behörden im Internet, die zuvor allein über zahlreiche individuelle Websites erreichbar war. Das "Portal der Deutschen Umweltinformation" bietet *jedem* Internetbenutzer einen unkomplizierten Zugang über eine einzige Adresse, sowie vielfältige Unterstützung bei der Wahl geeigneter Suchbedingungen durch Fach-Vokabular, geografische Namen und einen interaktiven Umweltkalender. Dahinter steht ein eigener XML-"Namensraum" (*namespace*), der die Kommunikation zwischen den einzelnen Informationsanbietern und ihrem gemeinsamen "Broker" auf eine stabile semantische Grundlage stellt. GEIN verfügt über ein thesaurus-basiertes Indizierungsverfahren, sowie über einen schnellen Verteilungsmechanismus für Anfragen, die nicht aus dem Index beantwortet werden können.

Seit der offiziellen Eröffnung am 9. Juni 2000 fand GEIN starke Akzeptanz in der allgemeinen Öffentlichkeit wie in der Fachwelt.

1. GEIN - the Network

The German Environmental Information Network (GEIN) consolidates a wide range of information currently distributed across many different Web sites run by public institutions in Germany, such as environmental authorities, agencies and ministries at the federal and Land (state) levels. It acts as an information broker for environmental information in Germany, or, as GEIN claims for itself, as "*the* portal to German environmental information."

¹ Sema Group, Cologne (D), thomas.bandholtz@sema.de

² Federal Environmental Agency, Berlin (D), richard.boes@uba.de

³ Federal Environmental Agency, Berlin (D), maria.ruether@uba.de

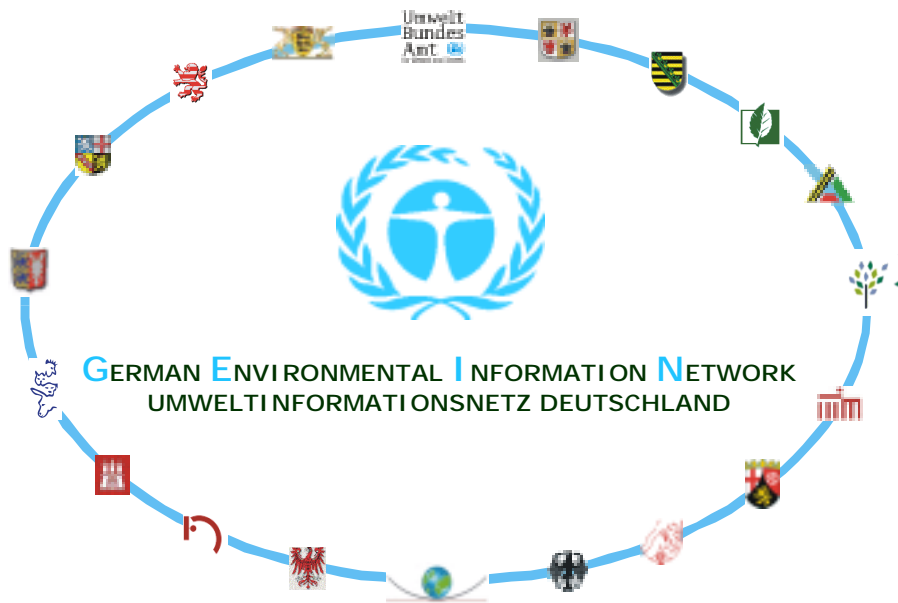


Figure 1
Logos of the Initial Information Providers

GEIN brings together 50⁴ suppliers of environmental information. This adds up to:

- more than 80,000 individual Web pages,
- a growing number of database interfaces (dynamic Web databases).

The Federal Environmental Agency (Umweltbundesamt) has established this countrywide information network as a resource for sharing its experience and knowledge on a national and international level and supporting its active involvement in the promotion of environmental protection.

The German Environmental Information Network was conceived under the title "GEIN 2000" in a close cooperative effort between the Federal Government and the Federal Länder in the framework of an Environmental Research Plan (UFOPLAN) project managed by the Federal Environmental Agency.

The development of the environmental information network was entrusted to the Cologne-based company Sema Group.

⁴ in July 2000

List of the initial information providers (2000-05-01)

Bundesbehörden: Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit
Umweltbundesamt
Umweltpräsentationssystem umwelt deutschland
Bundesanstalt für Geowissenschaften und Rohstoffe
Bundesamt für Naturschutz
Bundesamt für Strahlenschutz
Bundesamt für Seeschifffahrt und Hydrographie
Rat von Sachverständigen für Umweltfragen (SRU)
RAL/Umweltbundesamt Umweltzeichen "Blauer Engel"
Baden-Württemberg: Ministerium für Umwelt und Verkehr Baden-Württemberg
Landesanstalt für Umweltschutz Baden-Württemberg
Bayern: Bayerisches Staatsministerium für Landesentwicklung und Umweltfragen
Berlin: Senatsverwaltung für Stadtentwicklung, Umweltschutz und Technologie Berlin
BLUME - Berliner Luftgüte Messnetz
Stiftung Naturschutz Berlin
Brandenburg: Ministerium für Landwirtschaft, Umweltschutz und Raumordnung
Bremen: Senator für Bau und Umwelt
Hamburg: Freie und Hansestadt Hamburg – Umweltbehörde
HaLm Online (Online-Service des Hamburger Luftmessnetzes)
Arbeitsgemeinschaft für die Reinhaltung der Elbe
Hessen: Hessisches Ministerium für Umwelt, Landwirtschaft und Forsten
Hessisches Landesamt für Umwelt und Geologie
Multimediales Auskunfts- und Recherche-System MARS
Mecklenburg-Vorpommern: Umweltministerium Mecklenburg-Vorpommern
Landesamt für Umwelt, Naturschutz und Geologie Mecklenburg-Vorpommern (LUNG)
Niedersachsen: Umweltministerium Niedersachsen
Niedersächsisches Landesamt für Ökologie (NLÖ)
Alfred Töpfer Akademie für Naturschutz
Nationalparkverwaltung Harz
Niedersächsisches Landesamt für Bodenforschung
Institut für Geowissenschaftliche Gemeinschaftsaufgaben
Bezirksregierung Braunschweig
Bezirksregierung Hannover
Nordrhein-Westfalen: Ministerium für Umwelt, Raumordnung und Landwirtschaft
Landesumweltamt NRW
TEMES - Aktuelle Ozon-Messdaten
Rheinland-Pfalz: Ministerium für Umwelt und Forsten
Saarland: Ministerium für Umwelt
Sachsen: Sächsisches Staatsministerium für Umwelt und Landwirtschaft
Sächsisches Landesamt für Umwelt und Geologie
Sachsen-Anhalt: Ministerium für Raumordnung, Landwirtschaft und Umwelt
Landesamt für Umweltschutz Sachsen-Anhalt
Schleswig-Holstein: Ministerium für Umwelt, Natur und Forsten
Landesamt für Natur und Umwelt
Nationalpark Schleswig-Holsteinisches Wattenmeer
Thüringen: Thüringer Ministerium für Landwirtschaft, Naturschutz und Umwelt
Thüringer Landesanstalt für Umwelt
Thüringer Landesanstalt für Geologie

Figure 2
Initial Information Providers in GEIN

2. GEIN - the Application

GEIN runs on a public web server and is accessible by everybody via one of the common browsers without any special requirements.



Figure 3
The Page Header at <http://www.gein.de>

GEIN offers several ways of finding information.

First, there are three "static" collections that can be browsed by users with a more general interest:

- direct links to the homepages of the network members,
- recommended "portal" pages assembled by environmental topics,
- relevant events selectable from an environmental calendar.

While the links and portal pages have the form of hyperlinks, the calendar contains descriptions of each event itself. Each event can be made the starting point of a qualified search.

The latter is one of the three dynamic search facilities:

- conventional text search throughout the network
- qualified search using special terms (thesaurus based), by topic, area and time
- special search fields providing more in-depth access to selected topics.

Each query is resolved against an index of all (some 80,000) web pages, which are accessible by crawler mechanisms. Further on, the queries are broadcasted to a growing number (7 initially) of independent database servers that maintain their own corresponding search facilities. All the results are assembled in one master result list.

An unusual but very helpful feature is the assistance provided by two large thesauri, one of them containing environmental vocabulary and the other listing some 50,000 geographical names of different classes, with a complete knowledge about their intersections. Starting with the user's colloquial question, GEIN proposes the corresponding technical and geographical terms, which will be more exact and can be found as keywords in GEIN's qualified index (Figure 4). Time conditions may be selected from the environmental calendar. GEIN uses the same text analysis for the user's query as it does when indexing the web pages. Thus the query input

may consist of complete lines of text copied via the clipboard, taken from previous search results or different documents.

The screenshot displays the GEIN (German Environmental Information Network) Thesaurus search interface. At the top, the GEIN logo and name are visible, along with navigation links like 'HOME', 'About GEIN', 'About', 'Links', 'Help', and 'deutsch'. Below the logo, there are search options: 'Textsearch', 'Thesaurus', 'Specials', and 'eScout'. The main content area shows the user's input: 'How was water quality in Berlin from 1990 to 2000'. The analyzed input is shown as '[water quality, Berlin, 1990, 2000]'. The interface is divided into three sections for keyword suggestions:

- Topic - check the chosen terms!**: Suggests 'water quality' and 'Landesbehörde (Berlin)'. It includes a text input field for 'Enter more terms:' and a dropdown menu for 'What shall be found?' set to 'Many of the chosen terms ("or")'.
- Area - check the chosen names!**: Suggests 'Berlin District', 'Berlin Communities', 'Neuenhagen bei Berlin Community', 'Schöneiche a. Berlin Community', and 'Berlin State ("Land")'. It includes a text input field for 'Enter more names:' and a dropdown menu for 'What shall be found?' set to 'Many of the chosen terms ("or")'.
- Time - check the chosen dates!**: Suggests '1990' and '2000'. It includes a text input field for 'Enter more dates:' and a dropdown menu for 'What to look for?' set to 'single date or time span'.

At the bottom, there is a summary bar: 'You may now search with the chosen items ...' followed by a 'search now' button.

Figure 4
GEIN Proposes Keywords

The user may navigate in the vocabulary to select related but more suitable terms in two further windows shown below in figures 7 and 8.

For the inexperienced user, there are two levels of support:

1. Guidance on how to operate GEIN can be found under Help.
2. A *Scout* helps the user to find an appropriate approach to the information sought and to select the correct technical terms for the search.

The server log shows that most of the users do not need any guidance on the self explanatory navigation system, which has a very compact and "classic" feeling – no gimmicks, no banners, no plug-ins, no slow loading applets, no frames.

The lack of a true geographical *input* medium is quite tolerable given GEIN's wide knowledge of 50,000 geographical objects (and their intersections) by name. There are plenty of geographic views in the information presented in the collections and result lists.

3. GEIN - the XML Namespace

Early information brokers, like GELOS, GILS, CDS, JRC, EIONET should be well known to the UI community. All of them are based on pre-Internet standards like Z39.50 or SGML. More recently, the German Federal Environmental Agency has presented GEIN as an environmental information broker based on new-generation Internet XML technology, implemented by Sema Group. The data is stored on Software AG's Tamino XML server.

Basically, XML is: tags (<title>) and attributes (<title xml:lang="en">) that are used to markup content, the *text* between the tags, as in

```
<title xml:lang="en">GEIN - the XML Namespace</title>.
```

The names and the meaning of the elements and attributes are not predefined by XML, but agreed by a specific community concerned with a particular application. That's why XML is called "extensible". In fact, almost every usage is an extension.

XML *namespaces* are the solution to a very simple but ugly problem (following Murphy): if one community defines its own elements and attributes, there will always be another community using the same names with a different meaning.

The GEIN-namespace (*g2k*⁵, see <http://www.gein.de/2000/profile-11.htm>) is a working example. *g2k* defines tags like <*g2k:topic*>, which most probably would be equivocal in the XML-world without the prefix *g2k*. Figure 5 (next page) shows an example of *g2k* metadata describing a document which is identified by an URI (Unified Resource Identifier).

The same "profile" contains an example that makes use of two namespaces, here "*g2k*" and "*dc*" (Dublin Core, Figure 6, next page).

In the same way, this profile could integrate existing namespaces like those of GELOS, GILS, CDS, JRC, or EIONET if each of these defined their de facto namespace by means of XML. This is not real *harmonization*, but it is already progress if things can be *integrated* where they cannot be *harmonized*. In any case, do they really have to be harmonized? There are different domain-specific interests behind each of these namespaces, and too much harmonization would also lead to a certain loss of focus of these individual efforts. As Figure 6 shows in a very simple case, XML can be used with integrated namespaces, because each of them remains explicit.

⁵ In 1999, GEIN was named "GEIN 2000", and everybody was talking about "Y2K" instead of "Year 2000", so we invented G2K as GEIN 2000's nick name.

```

<?xml version='1.0' encoding="iso-8859-1"?>
<g2k:G2K xmlns:g2k="http://www.gein.de/2000/profile-11#" >
  <g2k:description uri="http://www.kingfisher.de/doc.htm"
    xml:lang="en" date="2000-02-02">
    <g2k:portal>NL</g2k:portal>
    <g2k:class>NL10</g2k:class>
    <g2k:title>The King Fisher</g2k:title>
    <g2k:abstract> About his habits and reservates.</g2k:abstract>
    <g2k:topic uid="28753" term="kingfisher" rank="10"/>
    <g2k:topic uid="16963" term="bird species" rank="5"/>
    <g2k:area uid="110000000" term="Berlin" type="town" rank="8"/>
    <g2k:area uid="NR-12" term="Havelland" type="reservat" rank="7"/>
    <g2k:time event="cal/2000" from="1999-12-31" to="2000-01-01"/>
  </g2k:description>
</g2k:G2K>

```

Figure 5
Example of a resource description using the g2k namespace.

```

<?xml version="1.0" encoding="iso-8859-1"?>
<g2k:G2K xmlns:g2k="http://www.gein.de/2000/profile-11#"
  xmlns:dc="http://purl.org/metadata/dublin_core#" >
  <g2k:specialQuery scope="Literature"
    xml:lang="en" mid="4711" match="and">
    <dc:Creator>Dr. B. King-Fisher</dc:Creator>
    <dc:Subject>Bird species, Kingfisher</dc:Subject>
    <g2k:topic uid="7827" term="Kingfisher" />
  </g2k:specialQuery>
</g2k:G2K>

```

Figure 6
Mixing two namespaces (g2k and dc) in one description

In the first version of its profile, GEIN tried to comply with the very restrictive rules of the Resource Description Framework (RDF), the W3C's recommended metadata layer based on and written in XML. RDF only allows a "triplet" form of semantic assignments, with the intention that any general RDF engine in the world can immediately handle any Resource Description conforming to RDF. In the end, it was decided that GEIN should use unrestricted XML for the internal network semantics and communication, but it still supports an external RDF interface.

4. GEIN - the Vocabulary

What seems to be more important in this context is the use of a well-defined *vocabulary*. Having defined the tags, GEIN proceeded to use thesauri and classifications in the "tagged" content. GEIN uses a complex thesaurus comprising more than 20,000 environmental terms and another 50,000 geographic names. Moreover, there is an environmental calendar that associates dates of important events with common "nicknames" (i.e. "since Chernobyl"). In Figure 5, the attributes "uid" (<g2k:topic> and <g2k:area>) and "event" (<g2k:time>) represent references to identifiers of these thesauri/calendar. This is the only way to handle homonyms, synonyms and dialects in metadata.

The **environmental thesaurus of the Federal Environmental Agency** contains some 8,700 "descriptors" (preferred terms) and 17,800 "non-descriptors" (as synonyms, broader and narrower terms).

The English version of this thesaurus is compatible with the English thread of the "General Multilingual Environmental Thesaurus" (GEMET) maintained by the European Environment Agency (EEA) in Copenhagen. GEMET is based on the integration of different national thesauri in Europe. In the current version 2.0 it contains "EnVoc" of UNEP - Infoterra and will probably be integrated in the USA by EPA.

GEMET contains 5,300 descriptors, 1,260 synonyms, and a complete glossary in 12 languages (English and most Western European languages). Thanks to this compatibility GEIN could be opened to all of these languages spoken in Europe (and world wide).



Figure 7
Thesaurus Navigation in the GEIN User Interface

The **Geographical Thesaurus of Environment** (GTE) was developed from scratch by GEIN with "a little" help by GISU, the Geographical Information System on Environment, maintained by the Federal Environmental Agency. In fact, GISU has calculated every intersection between one of the 50,000 geographical objects and a nation-wide 3km-grid, which provided solid ground for the further speed-optimized indexation of objects and intersections used by GEIN. The GTE covers different topics like cities and communities, administrative regions, different types of protected areas, mountains, rivers, lakes, sea und landscapes. Using a suitable navigation interface, one can easily stroll around, starting in a town, following the river, strolling in a biosphere reserve, climbing a mountain and so on to find the location that is relevant for the query in mind.



Figure 8
Geographical Names and Intersections in the GEIN User Interface (abbr.)

5. GEIN - the Indexing Machine

To make this possible, the vocabulary must be mapped to the information presented in GEIN. GEIN does this by maintaining two dynamic indices:

1. a conventional search engine's text index, and
2. a qualified index, which is thesaurus-based and structured according to the categories of topic, area and time.

The first of these does not differ from any typical search engine's index, taking into account that a thesaurus can never be perfect. This index is built by a common search engine crawler. The introduction of a limited list of servers to be indexed is easy to manage in various existing search engines (like Harvest, Ultraseek, ht://Dig, Altavista). After we tested the first three of them, we decided to use ht://Dig. This engine is open source (like Harvest), but it is well supported and supports high quality (like Ultraseek). Ultraseek and Altavista work fine (like ht://Dig) but are quite expensive. There is no reason to spend this money for non-commercial services.

nb

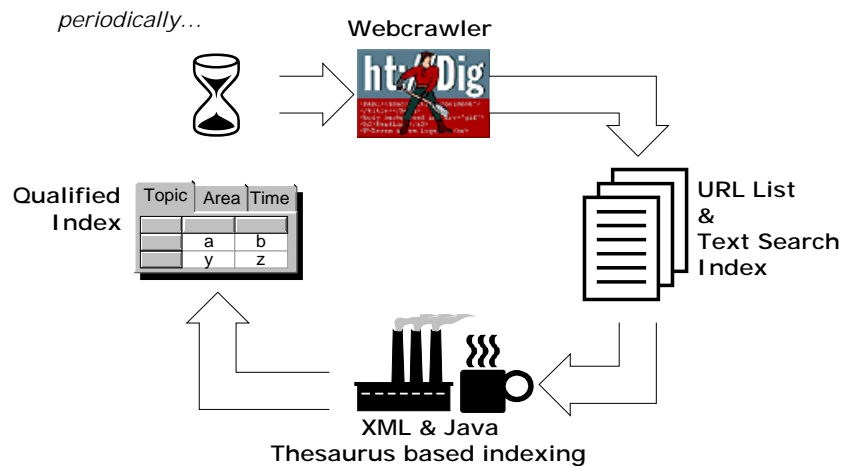


Figure 9
The Indexing Cycle

ht://Dig's 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 Text Search facility. As a second "give-away", it generates a list of all the URLs it has scanned.

GEIN has a second crawler written by the developers team themselves, which specializes in website consistency checking. It is thorough rather than fast, which is why we use it for site-checking on demand only.

The URL list is input to GEIN's thesaurus-based indexing machine, which analyses the content of each original page to find terms or synonyms, geographical names and time notations. In an automatic mode, the indexing machine registers a maximum of ten most significant keywords in topic and area, and a time or timespan if more than one time notation was found.

The quality seems to be much better than we suspected, but the information providers can override each set of keywords with their own selection, based on the same text analysis. Up to June 2000, this has happened only 32 times, but this frequency may grow rapidly as usage of the system increases. Pages indexed by information providers are continuously checked to ensure they are up to date, but the automatic indexer does not modify them. In case of inconsistencies the information provider is notified.

The whole cycle takes one or two days with some 80,000 pages, depending on how much new or updated indexing has to be done. The crawler completes the task in one night, but the indexing machine handles only about 12,000 pages in 24 hours. Thus one complete indexing takes 7 days, but if only 12,000 pages have changed in the second cycle it will take only one day after the crawler has run.

6. GEIN - the Query Broker

An outstanding feature of GEIN is that it makes dynamic Web databases accessible to users, which is otherwise hidden from conventional search machines. This function is of special importance as such databases mainly contain data on specific subjects and it makes these data easy to retrieve together with other information.

The problem of the "dynamic" sites not being indexed by the search engines does not have much to do with dynamics if one takes a close look. Even the (HTML) output of CGI- and Perl-Scripts or Java Servlets is readable and can be used by crawlers if they can receive a known address (URL). The problem lies in the wide spread use of *forms* in HTML. The crawler of a search engine is able to detect the form and its controls in the HTML code – but it is not able to understand (or to guess) their specific meaning. Therefore thousands of unlinked pages, accessible only via the original form, are not recognized by any search engine.

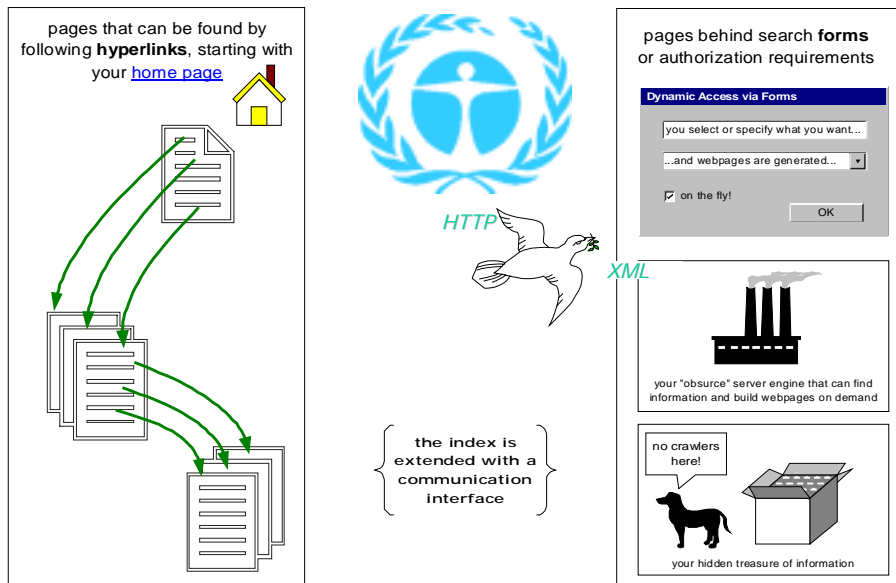


Figure 10
Broadcasting Queries

GEIN solves this problem by broadcasting the search condition to the corresponding servers of information providers. Of course there has to be an interface on the server that handles this request – just like any other post request sent by a browser. A communication protocol has been defined within the XML-namespace already discussed in Chapter 3.

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<g2k:G2K xmlns:g2k="http://www.gein.de/2000/profile-11#">
  <g2k:simpleSearch xml:lang="de" mid="4711" match="and">
    <g2k:class>LU10</g2k:class>
    <g2k:word>air pollution</g2k:word>
    <g2k:word>Berlin</g2k:word>
  </g2k:simpleSearch>
</g2k:G2K>
```

Figure 11
A "simpleSearch" Request Sent by GEIN

```

<?xml version="1.0" encoding="iso-8859-1" ?>
<g2k:G2K xmlns:g2k="http://www.gein.de/2000/profile-11#">
  <g2k:simpleResultSet mid="4711" xml:lang="de">
    <g2k:description uri="http://www.any-mu.de/doc.htm">
      <g2k:class>LU10</g2k:class>
      <g2k:title>Air Pollution in Berlin</g2k:title>
      <g2k:abstract>Introduction ... </g2k:abstract>
    </g2k:description>
  </g2k:simpleResultSet>
</g2k:G2K>

```

Figure 12
A "simpleResultSet" as Responded by an Information Provider

Figures 11 and 12 give a simple (fictional) example of the exchanged content. Of course, *simpleResultSet* usually contains more than one *description*. If they become too many, the response may summarize its results in one *metaResult* which points to a separate result set maintained exclusively by the responder himself. To handle the thesaurus-based detailed search conditions (topic-area-time), there is a *detailedQuery* and a *detailedResultSet* that correspond with the GEIN index structure as shown in Figure 5.

GEIN uses HTTP-Post-Requests for this communication, which are sent by Java Servlets. The corresponding servers use CGI, Perl, or Servlets to respond. Some cooperation was required on the part of the information providers in the beginning, but practice showed that this challenge was motivating and could be negotiated quickly and successfully.

7. GEIN - the Future of Environmental Information

Since it was opened by the German Environmental Minister Jürgen Trittin on 9 June this year, GEIN has been heavily accessed with an average of 5,000 hits a day. The first 18 days resulted in 131,115 hits by 14,130 visitors from 8,842 IP addresses. This (and the numerous e-mails) shows, that we have come to a point where users find their web navigation dreams satisfied, at least in the first instance.

One thread of further development will extend the presented facilities both in function and comfort. As the mass of information grows, GEIN has to keep up with capacity. There are many new providers who want to get in, and the mail traffic, newsletter, and forum need some form of lean but continuous moderation.

Another thread is the idea of a more generalized Environmental Markup Language (EML), which has been discussed in Berlin last year in a national context, and was presented to an international community at ISESS 2000 this June.

EML means a set of recommendations for the national and international usage of XML (eXtensible Markup Language) in the communication of environmental information.

EML consists of two parts:

- EML MetaData
- EML Data eXchange

EML MetaData is a special namespace and vocabulary used by information brokers that are part of the *Semantic Web* and have a special focus on environmental concerns. Any kind of information on the WWW that considers itself a contribution to global knowledge about the environment should index itself with EML MetaData.

EML Data eXchange consists of a core set of attribute and document type definitions to be used when exchanging any kind of environmental data other than metadata. Both attribute and document types may be extended or overwritten for the needs of special subdomains.

In the future, we expect a set of tools to be recommended and maintained by an agency that takes responsibility for EML.

Following these ideas, GEIN might be the start of an international semantic web of environmental information. If you are interested to see what it involves, visit <http://www.gein.de> for a first-hand impression.

8. References

- Arndt, Hans Knud, Bandholtz, Thomas, Günther, Oliver, Rüter, Maria, Schütz, Thomas (2000): EML - the Environmental Markup Language. In: Proceedings of the Workshop Symposium on Integration in Environmental Information Systems (ISESS 2000).
- Bandholtz, Thomas (2000): GEIN 2000 und darüber hinaus: Umweltinformation im "Semantic Web". In: Tagungsband 1. Workshop "Environmental Markup Language (EML)" der Fachgruppe 4.6.1 "Informatik im Umweltschutz" der Gesellschaft für Informatik (GI).
- Bilo, Michael, Streuff, Hartmut (2000): Das Umweltinformationsnetz Deutschland GEIN2000 – Fachliche Anforderungen an ein Forschungs- und Entwicklungsvorhaben. In: Tagungsband 3. Workshop des Arbeitskreises "Hypermedia im Umweltschutz" der Gesellschaft für Informatik (GI) am 23./24. März 2000.
- Treffler, Peter (2000): Das Geographische Informationssystem Umwelt. In: (GISU), In: Sicad Geomatics (Hrsg.): Kartographie und GIS im Umweltbereich; Grundlagen, Anwendungen, Beispiele und Trends, Hüthig, Oktober 2000