URN:NBN:fi:jyu-2007954 Issued 2001-05-21 © Jyväskylän yliopiston kirjasto

Jyväskylän yliopiston kirjaston tiedotuslehti
Verkkomakasiini

FROM STANDARDS TO THE INVISIBLE WEB

XML, SFX,... and other abbrieviations: A Metalink to the future
EUSIDIC Spring meeting 22.-23.3.2001 Lille, France

Pekka Olsbo
Jyväskylä University Library
Publishing Unit

EUSIDIC (European Association of Information Services) is an independent expert organization established in 1970, whose main function is to create forums and communication channels for people working in the information business. More than 100 different organizations from 16 European countries are involved in its activities. More about EUSIDIC at http://www.eusidic.org/

The EUSIDIC seminar in Lille was attended by more than 90 persons from 15 countries, all interested in XML and other issues connected with structured publishing. The issues dealt with in the seminar can be divided into three complementary sections: 1. The development of XML language and XML-based database solutions, 2. taxonomies and knowledge management and 3. the invisible web - search engines and web resources invisible to them. The seminar materials are available on EUSIDIC web pages.

All solutions and models presented at the seminar were based on XML (Extensible Markup Language) language and its applications. XML 1.0 version specification was published in February 1998. During the following three years XML has developed into a fairly well-known format used in many applications. XML is regarded as a format which will help solve the problems connected with electronic publishing. It is standardized and it will distinguish between the content and the appearance of a document. This guarantees good search features and it meets the requirements of long-term storage. There has long been a clear demand for these features, especially within industry and other fields producing technical reports and manuals. Also the university world and especially libraries have been investing in XML research and the utilization of its features. In recent years, the standardization of XML and its extensions has advanced significantly and its use has spread practically to all fields of electronic publishing. XML is regarded as a format which with time will replace HTML as the markup language of the web.

Extensible Markup Language and its development

At the EUSIDIC seminar, Josef Mattes, IT director of FIZ Karlsruhe presented a paper on the development and significance of XML-language. The use of XML has spread to all activities on the web. XML applications have been constructed for the needs of documents, databases, web publishing, e-commerce, and operating systems. Originally XML was developed for the needs of web publishing. There was a need for a standardized format which would be entirely independent of application and operating system and which would in practice enable any change needed.

When we talk about XML, we talk about an entire XML language family. XML has been developed from a more extensive SGML-standard, from which HTML, for example, has also been developed. The development of the XML standards is the responsibility of the World Wide Web Consortium (W3C), which is further developing tens of XML-based standards.

Standards connected with XML can be roughly divided into three groups. There seems to be differing views on what these groups should be called. Josef Mattes divides the XML-family into three group in the following way (Links connect to the W3C definition page):

1. XML Grammar contains information about XML syntax and it incorporates a host of standards, such as, XML Name Spaces, (a mechanism which ensures

the unambiguity of the logical structure of XML documents), XSL style sheets, XML Schema, Xpath, Xlink etc. These definitions are used to present limitations and regulations about how XML is constructed.

2. XML Protocols, which are responsible for the movement of data between systems and applications. These include, for example, RosettaNet, OASIS-ebCML and SOAP.

3. XML Vocabularies define the use of XML in various applications and they include discipline-based special definitions contained in the schema or DTD (Document Type Definition)(see for example W3C http://www.w3.org/XML/1998/06/xmlspec-report.htm and DocBook http://docbook.org/). These kinds of field-specific applications include, among others, CML (chemistry), MathML (mathematics), SMIL (multimedia) and RoboML (Human-Robot interface Using Agents Communicating in an XML-based Markup Language).

Airi Salminen, again, uses the following names for the above XML-families:

(See for more detail Salminen, Airi: Summary of the XML Family of W3C Languages http://www.cs.jyu.fi/~airi/xmlfamily.html)

1. XML Accessories,
2. XML Transducers
3. XML Applications.

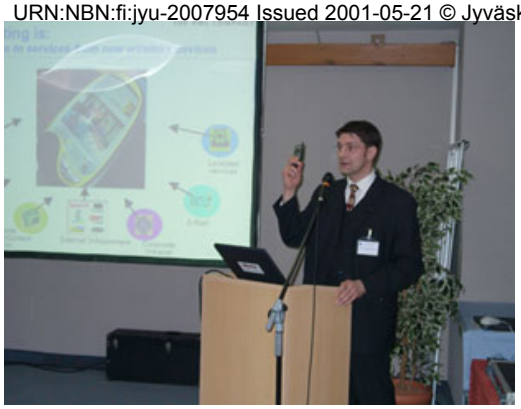
The W3C development group initially defined ten objectives for the development work:

1. The language must be well-suited for the internet environment.
2. It must support an extensive number of different applications from information description in structured databases to multimedia publishing.
3. The language must be SGML compatible.
4. The writing of XML document processing programs must be easy.
5. There should be a minimum number of optional features for each function.
6. XML documents should be clear and comprehensible.
7. The language must be developed quickly.
8. The language must be made formal and compact.
9. The creating of XML documents should be easy.
10. Scarceness of markings is not an end in itself.

According to Josef Mattes, XML development has proceeded in accordance with objectives. There have been no major setbacks in the development work.

Christian Schöter from *Software AG* presented an interesting outlook on future information society, where personified interfaces to personified information resources are brought to the use of all mankind. This aroused lively discussion about inequality and the shortage of available resources. What would be nicer than to give everyone a future hand-held computer for the handling of all everyday matters. Unfortunately reality is different. The fact is that 90 per cent of the world lives beyond the reach of these future images.

In industrialized countries, mobile applications with user recognition will, however, become more common with time. This is made possible particularly by XML, because at the moment it is the only format considered to be so universal and standardized that all applications can deal with it. This ideal picture naturally calls for major commitment from all hardware and software providers. However, Schöter considers this fully feasible and the only possibility which can guarantee inter-communication between different applications.



(Photo Dagmar Marek)

XML and mobile equipment development, for more details see: http://www.softwareag.com/xml/

SFX and the linking of information contents

The information seeker's problem is often that s/he does not get the information s/he wants from available information resources. Organizations distributing information have their own databases, which only provide information contained in them. There are also organizations that have been founded to link up and gather information coming from different sources. As a rule, these information gathering and distributing organizations are called libraries.

Information obtained via libraries has its restrictions, too. A search in a database usually focuses on only one database or on the databases offered by one service. Electronic journals, for example, consist of the databases of several service providers. ABI Inform and EBSCO HOST, among others, are such service providers.

When we submit a search to electronic journals offered by EBSCO HOST, for example, the search only focuses on the journals of the service in question. If we want to do the search also in other databases, we have to enter other databases. But wouldn't it be nice if the EBSCO search at the same time also produced information from other available databases? The same search function would focus not only on the databases provided by EBSCO HOST, but also on databases provided by, for example, *Cambridge Scientific Abstracts* and *SilverPlatter*. If the person seeking information so desired s/he could also choose to only search for meta data.

SFX, Content Sensitive Reference Linking has been designed to solve these problems. By means of SFX, an institution's electronic collections, both licenced and freely distributed, can be defined and distributed to end users in the desired way. Thereby all users are identified according to the organization or environment they come from. User recognition and the definition of the metadata objects of available information resources is done using OpenURL. OpenURL is a standard protocol, which verifies the compatibility between the information source and the local service provider. For more, see http://www.sfxit.com/openurl

The SFX system is in use at Gent University where electronic journals of over 60 publishers and more than 30 other databases are in joint use. SFX can be tried out on Ex Libris SFX pages at http://www.sfxit.com

Within the XML-standard family, Xlink, XML Linking Language, has been developed to serve similar needs as SFX. Xlink defines the structures that make it possible to add to XML documents elements needed in the creation and description of links. An ambitious (utopian) aim is to unite the information resources of the whole world. As Bohdana Stoklasova from the National Library of Check Republic noted, there are many problems along this road. How to make everybody use the existing standards? Not to mention the lack of economic and also intellectual resources, which affects most of the world. It is often a lot of work to overturn old structures and attitudes.

Taxonomy

However effective search systems there are available, the information must still be retrieved. Information or information about information must always be organised so that it is possible to find it. Knowledge/Information Management and the

classification or taxonomy of information have taken on a new meaning in the organization of the ever expanding electronic material. In the EUSIDIC seminar, some order to this chaos was sought by Susan Batley, *University of North London* and Alan Gilchrist, *TFPL Ltd*.

The purpose of a taxonomy and a classification is to create order to chaos. It is about the simplification of information, by means of which we try to find descriptions that combine subjects. By means of this descriptive information we build "subject trees", which create organized order. Without organized order it is impossible to find the right information.

When we start to structure information, for that is after all what taxonomy is about, we have to first find out what needs and tasks the taxonomy seeks to answer. We also have to analyse the resources, information and skills as well as the contents, value and use of the material (Subject analysis). We are talking about the same structuring of information when we define DTD, for example, in order to change over to XML-based publishing. DTD or Document type definition contains a list of elements, attributes, notations, entities and their interrelations in the document. DTDs define a host of rules regarding document structure (Harold 2000, 214). In the same way, taxonomies can be thought to describe rules regarding information structure.

According to Susan Batley, the most essential thing in the definition of taxonomies is unambiguity. Each subject should be given only one category and everybody should have a common understanding of the terms used, if at all possible.

(Photo Dagmar Marek)

Alan Gilchrist, again, presented facts that have given new meaning to taxonomies and the classification of information in a world overwhelmed by excessive information. The web which is open to everybody contains at the moment some 2,5 billion documents. Every day the amount increases by about 7.3 million. When we include the "deep web" or the "invisible web", i.e. various databases, intranets and other documents out of reach of search engines, the total amount of documents on the web is estimated at more than 500 billion. Microsoft's intranet alone includes 3 million documents. So we are talking about incomprehensible figures. The Internet's most comprehensive search engine Google has indexed some 1.35 billion web-pages. This gives us an idea of how much of the material on the net is out of reach of search engines. And yet, 95 per cent of those 500 billion documents are fully open to everybody.

It is not only a question of the material on the net being poorly described or in such a format that present search engines cannot find it. It is also a question of people not knowing how to use search engines correctly. Gilricht also presented some statistics on search behaviour. A sample of one billion searches made with search engines showed that over 72 per cent of search statements contained no more than two search terms. Over 20 per cent of search statements did not contain one single search term. In addition, 80 per cent of information seekers did not combine their terms with any operator. On the other hand, we have to bear in mind that, for example, Google automatically adds an AND operator between the terms, unless otherwise expressed. For more on taxonomies see TFPL Ltd: http://www.tfpl.com

Invisible Web

What are then the majority of web resources that modern search engines and agents cannot find, however intelligent searches we make? At the EUSIDIC seminar this question was elucidated by Chris Sherman from *Search Wise.Net*.

Sherman estimated that the "real" invisible web is probably about 2-50 times bigger than the "visible" web. According to him, the reason why these estimates about the size of the invisible web vary so much, lies in the way it is calculated. Calculations which have yielded 500 billion as the total of documents or information units in the invisible web have also taken into account the information sent to earth by satellites, which is probably rather impossible to calculate by unit.

(Photo Dagmar Marek)

In other words, the invisible web consists of all the material that search engines either don't find or they have not been programmed to find. Often the information content of this undiscovered material is considerably higher than on the visible web, since this material includes among other things databases which require registration. The reason why search engines don't find these databases is simply that search engines cannot fill in registration forms. It is estimated that the web has over 250,000 such unreachable databases.

Other undiscovered materials include certain file formats (PDF, Flash-applications, Office files etc.), most of real-time data (stock exchange, weather and flight schedules) and dynamically generated pages (cgi, javascript, asp and pages whose URL-address has "?").

Since the invisible web represents such a significant part of all the material on the web, services have been developed which specialize in distributing information about these invisible resources. Such services are offered by, for example, *Intelliseek* (http://www.invisibleweb.com and http://beta.profusion.com), *Complete Planet* (http://www.completeplanet.com/) and *Librarian´s Index to the Internet* (http://www.lii.org). On the pages of Complete Planet you will also find general information about the invisible web.

Chris Sherman also listed some significant "invisible" resources in various disciplines. Below is a short list of them

Computer Science

MacAfee World Virus Map: http://www.mcafee.com

- the name tells you what it is about

ResearchIndex: http://www.researchindex.com

- a digital library of scientific literature indexing, among other things, articles saved in Postscript and PDF format

Dictionaries & Languages

EuroDicAutom: http://eurodic.ip.lu

- Machine translation into 12 European languages. Works, but rather slowly. (the hook is krog in Danish)

Verbix: http://www.verbix.com/index.html

- the verbs of more than 50 languages. Translations from jiwarlin language, for example

Art & Artists

ADAM (Art, Design, Architecture & Media Information Gateway): http://www.adam.ac.uk/

- works like a virtual library, i.e. seeks links to suitable resources on the web.

Artcyclopedia: http://www.artcyclopedia.com/

- a search engine for artists, works of art and art museums

The attachment contains a complete list of Sherman's links.

Future

At present, most of the materials invisible to search engines are in some other form than the web's basic format HTML. For example, a very popular document distribution format on the web, PDF, is unreachable by most search engines. Google is the first search engine that has started to index also PDF documents on the web. One could imagine that other services soon follow suit. There are also many other formats that search engines cannot index at the moment.

So, organizations providing search services are facing heavy investments. Sherman predicts that during this year at least two big search service providers will have to close down because of lack of resources. At the same time, service providers of a completely new type will be entering the market. Intelligent agents, *"killer applications", Inktomi, Index Connect, Ultraseek, WhizBang* and *wrappers* are terms and applications that we are sure to be bumping into.

Although search systems are developing all the time, the problem of the invisible web will never disappear. At the same speed as new search engines develop, develop also new formats and more and more complicated applications. The only way to reduce the proportion of invisible material is to increase web publishing based on standards. This requires commitment also from software houses and application developers. And always when commitment is required there are problems to be expected.

XML is generally regarded as the saviour format because of its standardization and scaleability. It is, however, still difficult to say whether XML will share the fate of so many other standards: at first everybody is agreed on the excellence of the standard, but its transfer into practical measures is slow and remains incomplete. Finally everybody forgets that the standard ever existed. Except the developers of the standard. It has to be said for XML that in recent years its development has been very quick and there are already quite a lot of practical applications in use. So, XML has its possibilities and many think that XML is also the only possibility.

Sources

Batley, Susan. Access to Electronic Documents: Taxonomies in Information and Knowledge Management. Lecture. Lille 23.3.2001. Available: http://www.eusidic.org

Daubach, Marc. SFX: an open linking framework for libraries. Lecture. Lille 22.3.2001. Available: http://www.eusidic.org

DocBook.org: http://www.docbook.org

EUSIDIC, European Assosiaton of Information Services. http://www.eusidic.org

ExLibris: http://www.exlibris-usa.com/

FIZ Karlsruhe: http://www.fiz-karlsruhe.de/

Gilchrist, Alan. Corporate Taxononomies: Different Approaches to their Construction and Use. Lecture. lille 23.3.2001. Available: http://www.eusidic.org

Harold, Eliotte Rusty, 2000. XML-tehokäyttäjän opas. Satku – Kauppakamari. Jyväskylä 2000

Inktomi Corporation: http://www.inktomi.com/

Mattes, Josef. The swift success of XML. Lecture. Lille 22.3.2001. Available: . http://www.eusidic.org

Salminen, Airi, 2001. Summary of the XML Family of W3C Languages. Available: http://www.cs.jyu.fi/~airi/xmlfamily.html (viitattu 5.4.2001)

Schöter, Christian. The Return of Magic or how mobile-Commerce applications will revolutionize

life on the planet ... again. Lecture. Lille 22.3.2001. Available: http://www.eusidic.org

SearchWise.net: http:// www.SearchWise.net

Sherman, Chris. The Invisible Web. Lecture. Lille 23.3.2001. Available: http://www.eusidic.org

SoftWare AG: http://www.softwareag.com

TFPL Ltd: http://www.tfpl.com/index.html

WhizBang! Labs, Inc. http://www.WhizBang.com/

World Wide Web Consortium: http://www.w3.org

Appendix

Chris Shermans Invisible Web links

(Reproduced with kind permission by Mr Chris Sherman)

Gateways

- *Intelliseek*

- http://www.invisibleweb.com
- http://beta.profusion.com

Computer Science

- *MacAfee World Virus Map*

- http://www.mcafee.com

- *ResearchIndex*

- http://www.researchindex.com

Company Research

- *European High_Tech Industry Database*

- http://www.tornadoinsider.com/radar/

- *Kompass*

- http://www.kompass.com

Intellectual Property

- *Delphion Intellectual Property Network*

- http://www.delphion.com/

- *ESP@CENET (European Patent Office) Patent Database*

- http://ep.espacenet.com/

Dictionaries & Languages

- *EuroDicAutom*

- http://eurodic.ip.lu

- *Verbix*

- http://www.verbix.com/index.html

Arts & Artists

- *ADAM (Art, Design, Architecture & Media Information Gateway)*

- http://adam.ac.uk/

- *Artcyclopedia*

- http://www.artcyclopedia.com/

Real-Time Information

- *Flight Tracker*

- http://www.trip.com/ft/home/0,2096,1-1,00.shtml

- *J-Track 3-D Satelite Locator*

- http://liftoff.msfc.nasa.gov/realtime/Jtrack/Spacecraft.html

Maps and Driving Directions

- *MapBlast*

- http://www.mapblast.com

- *Streetmap.co.uk*

- http://www.streetmap.co.uk/

Government Info

- *Parline Database*

- http://www.ipu.org

- *United Nations Daily Press*

- http://www.un.org/news/

Health & Medicine

- *Economics of Tobacco Control Database*

- http://www1.worldbank.org/tobacco/database.asp

- *International Digest of Health Legislation*

- http://www.who.int

News & Current Events

- *Cold North Wind Newspaper Archiva Project*

- http://www.coldnorthwind.com

- *Financial Times Global Archive*

- http://www.globalarchive.ft.com

Science

- *Great Barrier Reef Online Image Catalogue*

-http://www.gbrmpa.gov.au/corp_site/info_services/library/index.html

- *Nuclear Explosions Database*

- http://ausseis.gov.au/databases

Transportation

- *Equasis (Merchant Ships)*

- http://www.equasis.org/

- *World Aircraft Accident Summary (WAAS) Fatal Airline Accident Subset*

- http://www.waasinfo.net/

Pekka Olsbo
Jyväskylä University Library
Publishing Unit