Towards semantic web: adding meaning and trust to the web by XML

Airi Salminen
University of Jyväskylä
http://www.cs.jyu.fi/~airi/

TUCS 28.11.2002

Outline

1. Milestones of the web
2. What is XML?
3. Why XML evolved
4. What is semantic web?
5. Metadata on the web
6. XML as metadata
7. The RDF model
8. Semantic web architecture
9. XML-based languages for semantic web
10. Related research at the University of Jyväskylä

1. Milestones of the web

1960-1980 ... Infrastructure for the Internet

- RFC = Request for Comments
- TCP/IP

1986 ... SGML (Standard Generalized Markup Language)

1991 ... WWW, HTML, Internet Society

1. Milestones of the web

1992 ... computers connected to the Internet > 1000.000

1994 ... W3C = World Wide Web Consortium

1996 ... PICS = Platform for Content Selection

1998 ... XML, Dublin Core

1999 ... RDF = Resource Description Framework

2000 ... computers connected to the Internet > 100.1000.000

2. What is XML?

XML = Extensible Markup Language

A set of rules for defining and representing information as structured documents for applications on the Internet; a restricted form of SGML (Standard Generalized Markup Language)

T. Bray, J. Paoli, C. M. Sperberg-McQueen, and E. Maler (Eds.), Extensible Markup Language (XML) 1.0 (Second Edition), W3C Recommendation 6 October 2000, http://www.w3.org/TR/2000/REC-xml-20001006

2. What is XML?

- Rule 1: Information is represented in units called *XML documents.*
- Rule 2: An XML document contains one or more *elements.*
- Rule 3: An element has a name, it is denoted in the document by explicit markup, it can contain other elements, and it can be associated with *attributes.*

and lots of other rules ...

2. What is XML?

Example of an XML document

```
<?xml version = "1.0"?>
<poem author = ”Murasaki Shikibu” author_born = ”974”>
<info_link  xmlns:xlink=”http://www.w3.org/1999/xlink”
    xlink:type="simple”
    xlink:href=
    ”http://digital.library.upenn.edu/women/omori/court/murasaki.html”>
       About the author
</info_link>
<stanza>
<line>This life of ours would not cause you sorrow</line>
<line>if you thought of it as like </line>
<line>the mountain cherry blossoms</line>
<line>which bloom and fade in a day. </line>
</stanza>
</poem>
```

Note: The text of the line elements is taken from http://www.slip.net/~knabb/rexroth/translations/japanese.htm, containing Kenneth Rexroth’s translations of Japanese poetry

2. What is XML?

XML is a metalanguage, not a specific language

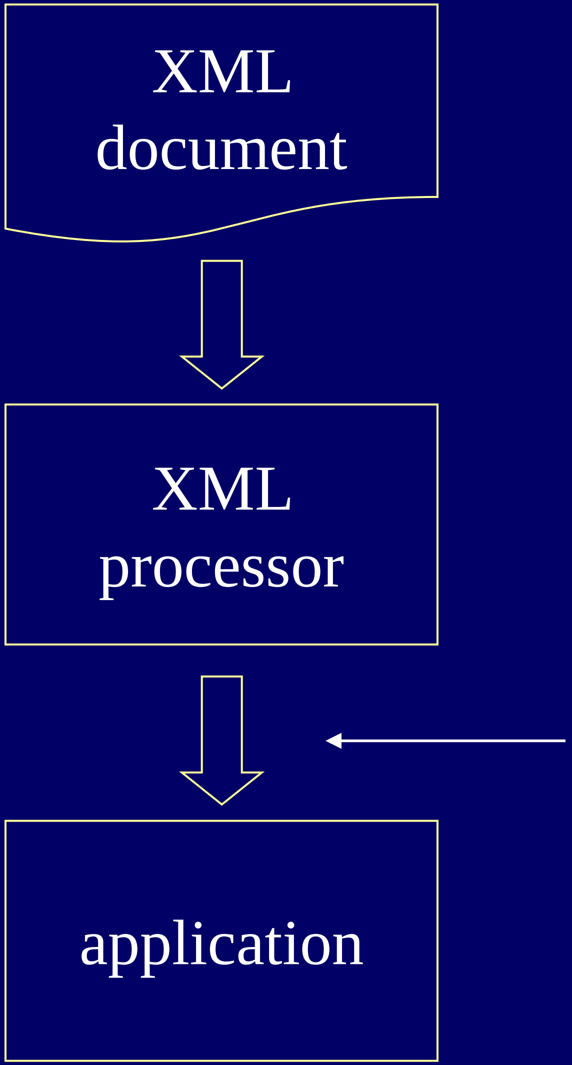
- Defines the rules how to mark up a document — does not define the names used in markup.
- Includes capability to prescribe a document type by a collection of declarations to constrain the markup permitted in a class of documents.
- Intended for *all* natural languages, regardless of character set, orientation of script, etc.

2. What is XML?

Document type declaration for a poem

```
<!DOCTYPE poem [
<!ELEMENT poem   (info_link? title?, stanza+)>
<!ATTLIST poem
       author CDATA  #REQUIRED
       author_born  CDATA    #IMPLIED>
<!ELEMENT title            (#PCDATA) >
<!ELEMENT info_link    (#PCDATA) >
<!ATTLIST info_link
     xmlns:xlink  CDATA  #FIXED "http://www.w3.org/1999/xlink"
     xlink:type  CDATA    #FIXED "simple"
     xlink:href  CDATA    #REQUIRED >
<!ELEMENT stanza        (line+)   >
<!ELEMENT line             (#PCDATA) >]
```

2. What is XML?



XML document

XML processor

application

may or may not be "validating"

"XML Information Set"

3. Why XML evolved

After the breakthrough of WWW and HTML there was an urgent need for a new, common data format for the Internet

- Needs:
  - Simple, common rules that are easy to understand by people with different backgrounds (like HTML)
  - Capability to describe Internet resources and their relationships (like HTML)
  - Capability to define information structures for different kinds of business sectors (*unlike* HTML, like SGML)

3. Why XML evolved

- Needs (cont’d):
  - Format formal enough for computers and clear enough to be human-legible (like SGML)
  - Rules simple enough to allow easy building of software (*unlike* SGML)
  - Strong support for diverse natural languages (*unlike* SGML)

4. What is semantic web?

The abstract representation of data on the World Wide Web, based on the RDF standards and other standards to be defined. It is being developed by the W3C, in collaboration with a large number of researchers and industrial partners

W3C Semantic Web Activity,
http://www.w3.org/TR/2001/sw/

4. What is semantic web?

An extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation

Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, Scientific American, May 2001.
http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html

4. What is semantic web?

- Web resources consist of primary resources and metadata resources.
- Metadata resources related to the meaning, use, and trustworthness of the (primary) resources.
- Metadata resources first class web resources.
- Metadata in standardized formats readable both by software and people.

4. What is semantic web?

- Formats based on XML and RDF.
- Major portion of the primary resources written in various natural languages used in various communities.
- Homogeneous metadata about heterogeneous content.
- Enabling merging of resoursers.

4. What is semantic web?

- Automated reasoning about meaning and trustworthness.
- Enabling extensive cooperation of software.
- Enabling and requiring cooperation of people in communities having shared understanding of the meaning of the content and shared values.
- Development coordinated by W3C.

5. Metadata on the web

metadata = data about web resources

about

- documents
- databases
- applications
- services

5. Metadata on the web

Examples of metadata

About a document

- title
- creator
- subject
- format
- identifier
- description
- publisher
- rights

Can be given, for example, by Dublin Core elements

5. Metadata on the web

Examples of metadata (cont'd)

About a document repository

- structure (DTD, XML Schema)
- words in the content (indexes)
- concepts and their meanings (ontologies)

5. Metadata on the web

Examples of metadata (cont'd)

About metadata in a repository

- vocabularies of the markup (namespace, DTD, XML Schema)
- vocabularies in the metadata descriptions (RDF Schema)
- data types in the schemas (XML Schema type definitions)

5. Metadata on the web

Examples of metadata (cont’d)

- users of an application
- access rights related to the resources of a community
- annotations for a document (Annotea)
- business process where documents are created

5. Metadata on the web

metadata classifications

embedded	external
centralized	distributed
created by people	created by software

6. XML as metadata

- The markup used in a document serves as metadata in relationship to the character data
- The declarations associated with a class of documents serve as metadata in relationship to the documents.

6. XML as metadata

```
<?xml version = "1.0"?>
<poem author = ”Murasaki Shikibu” author_born = ”974”>
<info_link  xmlns:xlink=”http://www.w3.org/1999/xlink”
    xlink:type="simple”
    xlink:href=
    ”http://digital.library.upenn.edu/women/omori/court/murasaki.html”>
       About the author
</info_link>
<stanza>
<line>This life of ours would not cause you sorrow</line>
<line>if you thought of it as like </line>
<line>the mountain cherry blossoms</line>
<line>which bloom and fade in a day. </line>
</stanza>
</poem>
```

6. XML as metadata

**This life of ours would not cause you sorrow
if you thought of it as like
the mountain cherry blossoms
which bloom and fade in a day.**

Lisätietoa runoilijasta

6. XML as metadata

Metadata expressed in the markup :

- The document is called a poem and it consists of elements called info_link and stanza, and the stanza consists of elements called line.

- The author of the poem is Murasaki Shikibu, born in 974.

- The element info_link with the text content ”About the author” is a simple link referring to the Web resource at http://digital.library.upenn.edu/women/omori/court/murasaki.html

- …

6. XML as metadata

Also DTD provides metadata

```
<!DOCTYPE poem [
<!ELEMENT poem   (info_link? title?, stanza+)>
<!ATTLIST poem  author CDATA  #REQUIRED
                author_born  CDATA   #OMITTED>
<!ELEMENT title              (#PCDATA) >
<!ELEMENT info_link          (#PCDATA) >
<!ATTLIST info_link
    xmlns:xlink  CDATA  #FIXED "http://www.w3.org/1999/xlink"
    xlink:type  CDATA   #FIXED "simple"
    xlink:href  CDATA   #REQUIRED >
<!ELEMENT stanza             (line+)   >
<!ELEMENT line (#PCDATA) >]
```

6. XML as metadata

The metadata provided by the DTD

Vocabulary: poem, stnza, line, author, ...

Structure:

- The documents are called poems.
- A poem may contain an element called title and it always contains one or more elements called stanza.
- A poem may be linked to a resource by a simple link.
- For each poem there is information about the author and possibly about the year of birth of the author.

7. The RDF model

RDF = Resource Description Framework

a model for describing web resources

RDF Specification: http://www.w3.org/TR/REC-rdf-syntax/

resource anything that can be identified on the Internet; identification by URI

examples: file, service, site, part of a file, book, person, company

7. The RDF model

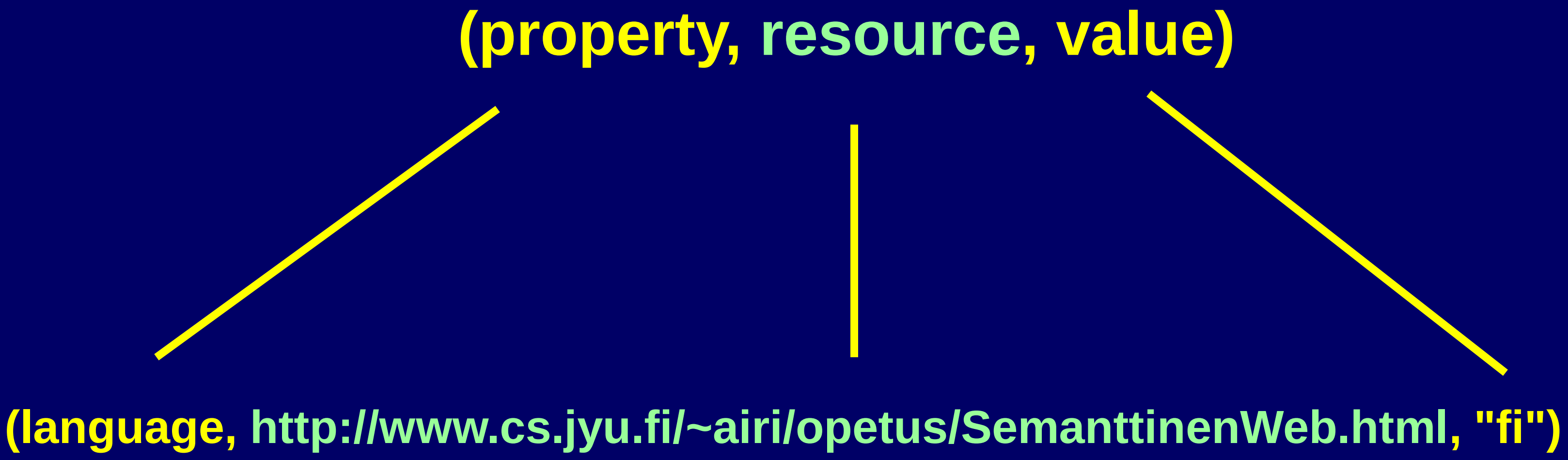
Examples of resources

resource	URI
home page of a course	http://www.cs.jyu.fi/~airi/opetus/SemanttinenWeb.html
Department of CS & IS at the University of Jyväskylä	http://cs.jyu.fi
Airi Salminen	http://cs.jyu.fi/henkilot/asalminen
Home page of Airi Salminen	http://www.cs.jyu.fi/~airi/

7. The RDF model

RDF description consists of statements

A statement is a triple expressing the value of a property of a resource:



(property, resource, value)

(language, http://www.cs.jyu.fi/~airi/opetus/SemanttinenWeb.html, "fi")

7. The RDF model

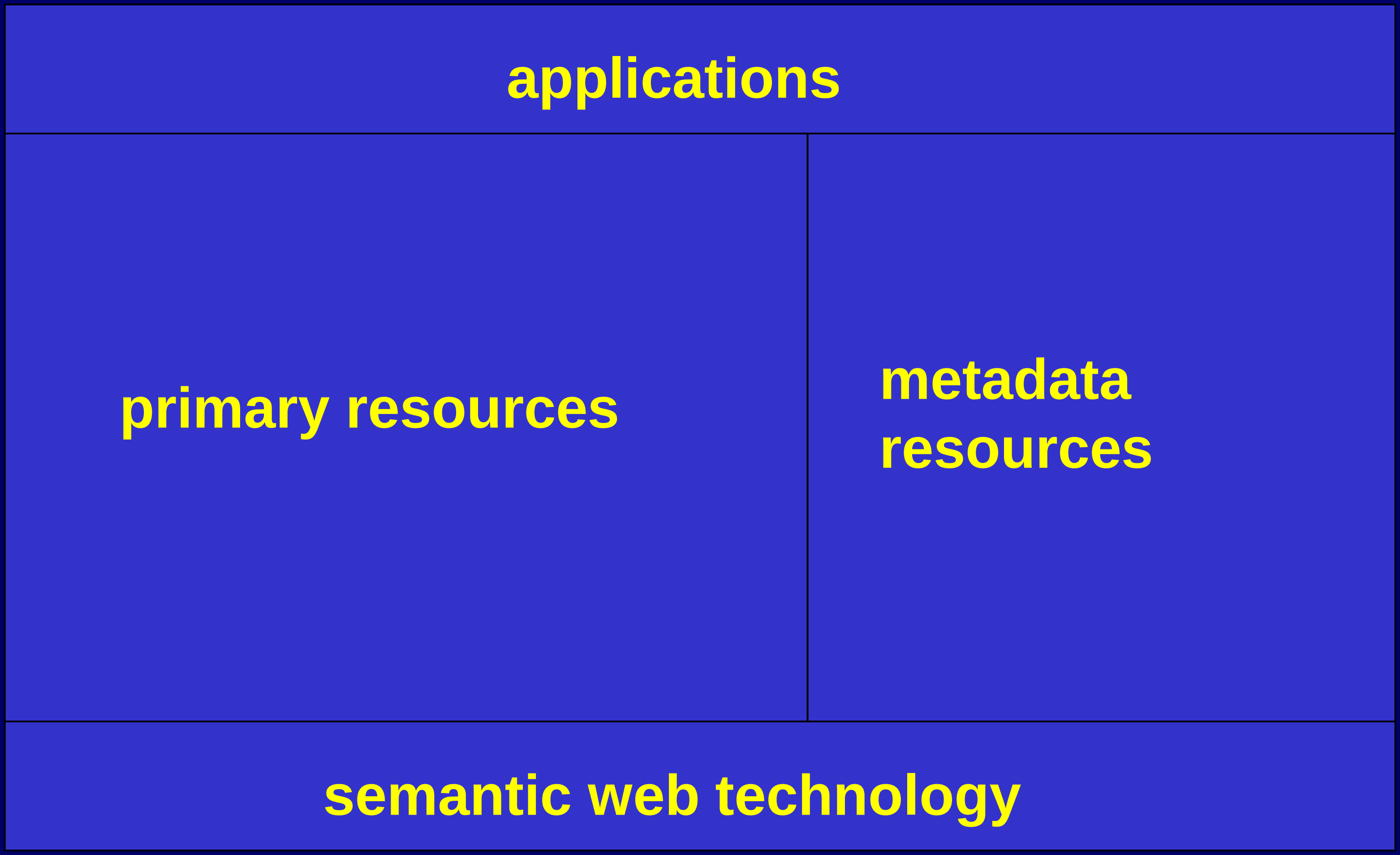
(dc:Creator,
http://www.cs.jyu.fi/~airi/opetus/SemanttinenWeb.html,
"Airi Salminen")

(dc:Language,
http://www.cs.jyu.fi/~airi/opetus/SemanttinenWeb.html,
"fi")

7. The RDF model

- RDF is intended to facilitate automated processing of Web resources
- RDF does not specify a mechanism for reasoning
- Intended to be used in a variety of application areas:
  - resource discovery
  - cataloging
  - by intelligent software agents
  - in content rating
  - to build a "web of trust" with digital signatures

8. Semantic web architecture



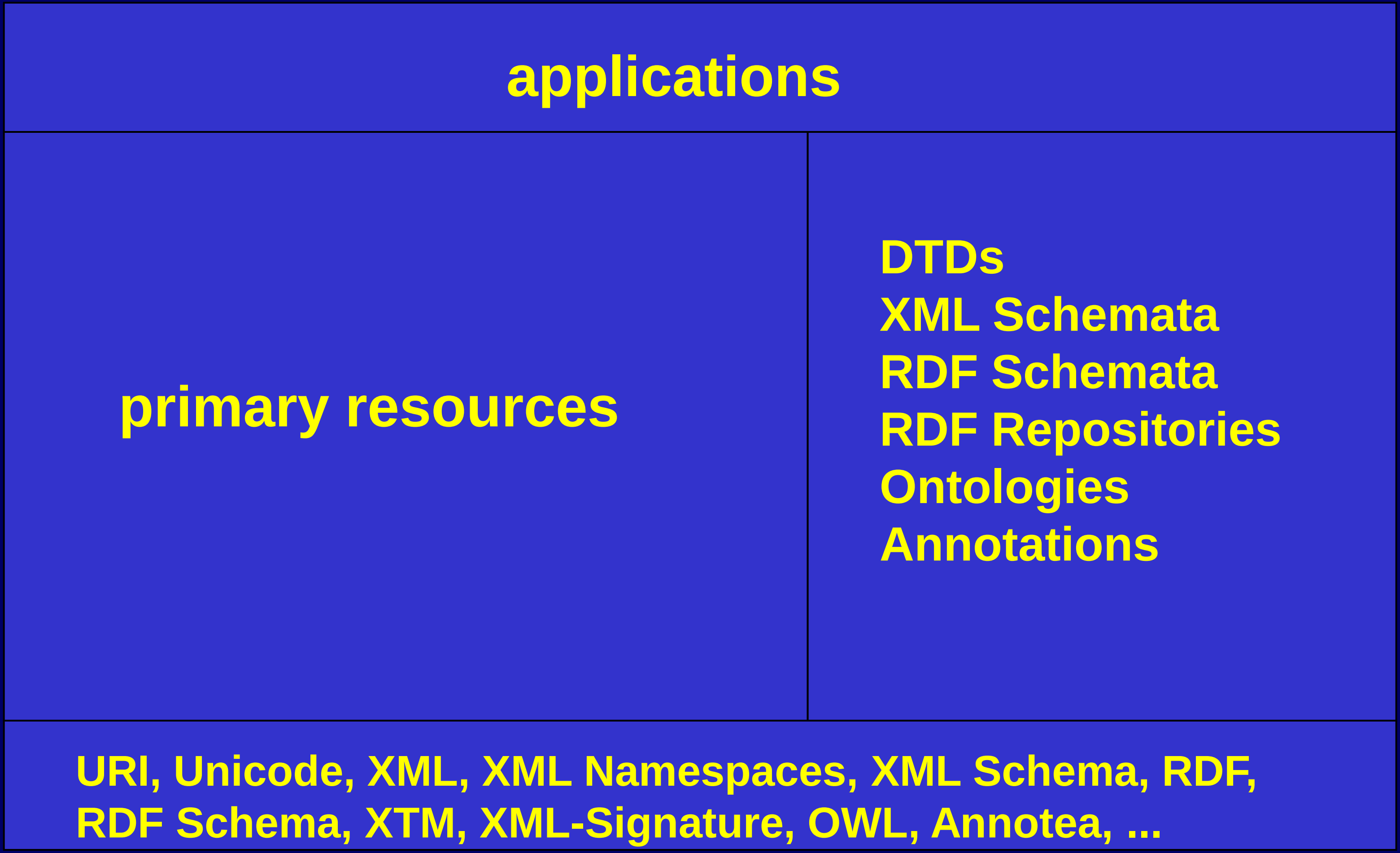
applications

primary resources

metadata resources

semantic web technology

8. Semantic web architecture



applications

primary resources

DTDs
XML Schemata
RDF Schemata
RDF Repositories
Ontologies
Annotations

URI, Unicode, XML, XML Namespaces, XML Schema, RDF, RDF Schema, XTM, XML-Signature, OWL, Annotea, ...

9. XML-based languages for semantic web

Languages for representing and defining structured documents

- XML
- XML Namespaces
- XML Schema

9. XML-based languages for semantic web

language	purpose
RDF	describing web resources
RDF Schema	defining RDF vocabularies
OWL	publishing and sharing ontologies on the web
XTM	Topic maps

9. XML-based languages for semantic web

language	purpose
XML-Signature	digital signatures
XKMS	public keys
P3P APPEL	privacy practices for web sites preferences regarding P3P policies
XML Encryption	encrypted data

10. Related work at the University of Jyväskylä

EULEGIS, European User Views to Legislative Information in Structured Form (Airi Salminen et al.) http://www.cs.jyu.fi/~airi/docman.html#eulegis

The purpose was to offer a consistent user interface to retrieve legal information created in different legal systems and at different levels - the European Union, a member state, a region, or a municipality. Utilized contextual metadata and ontologies in the user interface.

10. Related work at the University of Jyväskylä

DrElma: Digital Rights of Electronic Learning Materials (Pasi Tyrväinen et al.)
http://www.cs.jyu.fi/~airi/docman.html#DrElma

Steve Legrand (steveleg@hotmail.com), Using ontologies for text disambiguation

The main motivation behind this research is to improve the accuracy of linguistic parsers to benefit linguistic applications used in translation and language learning and other tasks, which use parsers for disambiguation.

10. Related work at the University of Jyväskylä

Airi Salminen, XML family of languages. Overview and classification of W3C specifications. Available at http://www.cs.jyu.fi/~airi/xmlfamily.html.

Airi Salminen, Semanttinen web. Home page of a course. Available at http://www.cs.jyu.fi/~airi/opetus/SemanttinenWeb.html.