Semantic Web Overview

Diane Vizine-Goetz
OCLC Research

Outline

- Semantic Web vision
- Core technologies
- OCLC Web services

The Vision

"The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation." [1]

More on the Vision

". . .information on the web needs to be in a form that machines can 'understand' rather than simply display. The concept of machine-understandable documents does not imply some magical artificial intelligence allowing machines to comprehend human mumblings. It relies solely on a machine's ability to solve well-defined problems by performing well-defined operations on well-defined data." [2]

Core technologies

- eXtensible Markup Language (XML)
- Resource Description Framework (RDF)
- Ontologies
- Software agents

XML (eXtensible Markup Language)

- Standard designed to transmit structured data to Web applications
- Describes structure & content
- Provides syntactic interoperability
- XML namespaces qualify element names uniquely on the Web in order to avoid conflicts between elements with the same name

Metadata in HTML

<body>

Title: Automatic Classification and Content Navigation Support for Web Services

Creator: Traugott Koch

Creator: Diane Vizine-Goetz

Subject: Automatic classification

Subject: Knowledge organization

Publisher: OCLC

Date: 1999

Type: Text

Identifier: http://www.oclc.org/research/publications/arr/1998/koch_vizine-goetz/automatic.htm

Language: en

</body>

Metadata in XML

```
<?xml version="1.0" ?>
<metadata xmlns:dc="http://purl.org/dc/elements/1.1/">
  <dc:title>Automatic Classification and Content Navigation Support for Web Services</dc:title>
  <dc:creator>Traugott Koch</dc:creator>
  <dc:creator>Diane Vizine-Goetz</dc:creator>
  <dc:subject>Automatic classification</dc:subject>
  <dc:subject>Knowledge organization</dc:subject>
  <dc:publisher>OCLC</dc:publisher>
  <dc:date>1999</dc:date>
  <dc:type>Text</dc:type>
  <dc:identifier>http://www.oclc.org/research/publications/arr/1998/koch_vizine-goetz/automatic.htm</dc:identifier>
  <dc:language>en</dc:language>
</metadata>
```

RDF (Resource Description Framework)

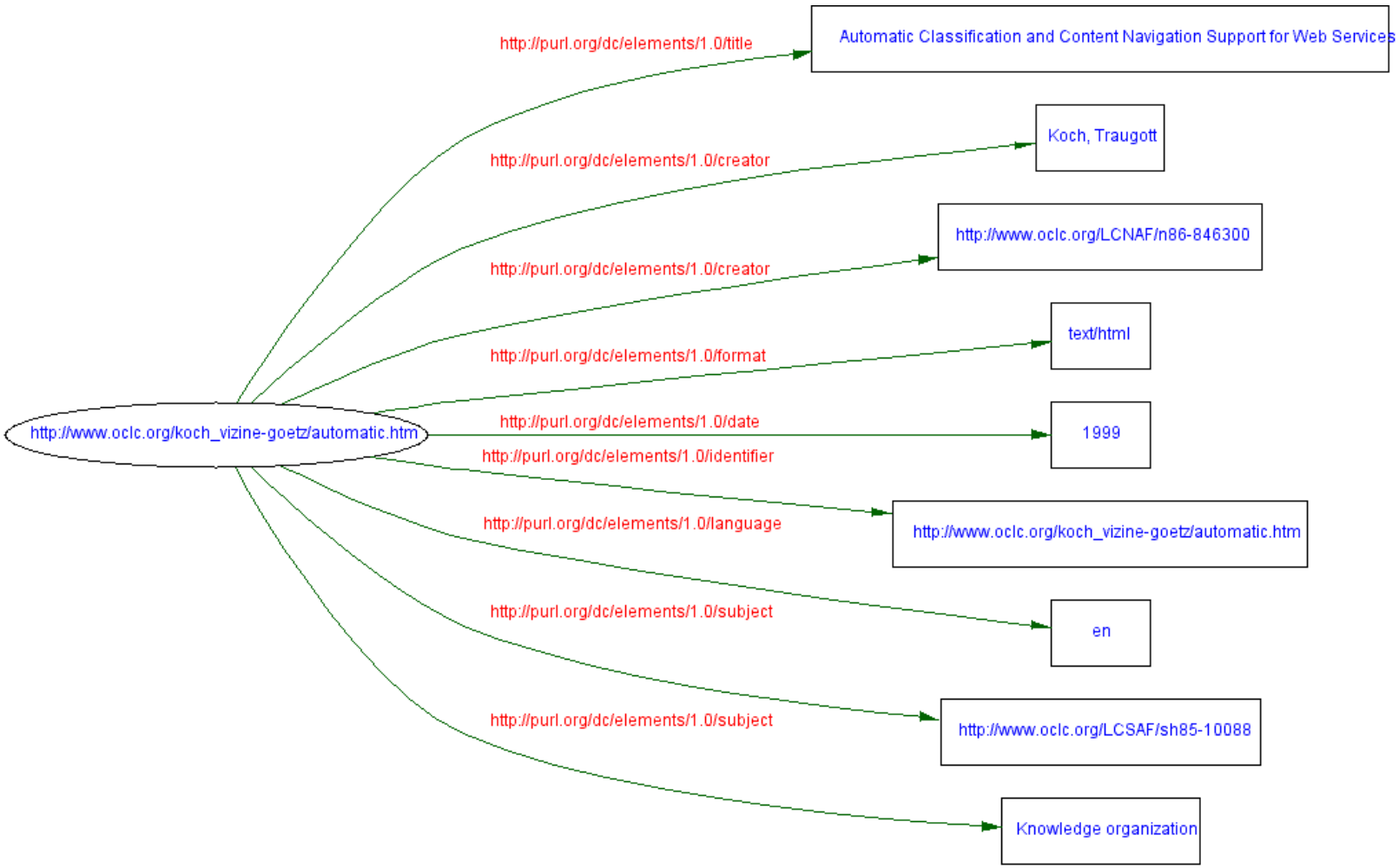
- Provides a mechanism for encoding meaning
- Simple way to state facts (e.g., properties, characteristics) about web resources
- Employs URIs to identify resources
- Data interoperability layer

URIs link concepts to unique definitions

- dc:creator
  - Traugott Koch
    - http://www.oclc.org/LCNAF/n93-57973
  - Diane Vizine-Goetz
    - http://www.oclc.org/LCNAF/n86-846300
- dc:subject
  - Automatic classification
    - http://www.oclc.org/LCSAF/sh85-10088
  - Knowledge organization
    - http://www.oclc.org/LCSAF/sh85-10088

Metadata in RDF

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.0/">
<rdf:Description about="http://www.oclc.org/research/publications/arr/1998/koch_vizine-goetz/automatic.htm">
<dc:title>Automatic Classification and Content Navigation Support for Web Services</dc:title>
<dc:creator>Koch, Traugott</dc:creator>
<dc:creator>http://www.oclc.org/LCNAF/n86-846300</dc:creator>
<dc:format>text/html</dc:format>
<dc:publisher>OCLC</dc:publisher>
<dc:date>1999</dc:date>
<dc:identifier>http://www.oclc.org/research/publications/arr/1998/koch_vizine-goetz/automatic.htm</dc:identifier>
<dc:language>en</dc:language>
<dc:subject>http://www.oclc.org/LCSAF/sh85-10088</dc:subject>
<dc:subject>Knowledge organization</dc:subject>
</rdf:Description>
```


http://www.oclc.org/koch_vizine-goetz/automatic.htm
http://purl.org/dc/elements/1.0/title
Automatic Classification and Content Navigation Support for Web Services
http://purl.org/dc/elements/1.0/creator
Koch, Traugott
http://purl.org/dc/elements/1.0/creator
http://www.oclc.org/LCNAF/n86-846300
http://purl.org/dc/elements/1.0/format
text/html
http://purl.org/dc/elements/1.0/date
1999
http://purl.org/dc/elements/1.0/identifier
http://www.oclc.org/koch_vizine-goetz/automatic.htm
http://purl.org/dc/elements/1.0/language
en
http://purl.org/dc/elements/1.0/subject
http://www.oclc.org/LCSAF/sh85-10088
http://purl.org/dc/elements/1.0/subject
Knowledge organization

Ontologies

- An ontology formally defines a common set of terms that are used to describe and represent a domain (e.g., librarianship, medicine, etc.)
- Ontologies include computer-usable definitions of basic concepts in the domain and the relationships among them
- Ontologies are usually expressed in a logic-based language

Ontologies

- A web ontology language, the logic layer, will provide a language for describing the set of inferences that can be made for a collection of data
- For example, a search program using an ontology might look only for resources described by precise concepts, from a given set of KO resources, instead of simple keywords (see RDF example)

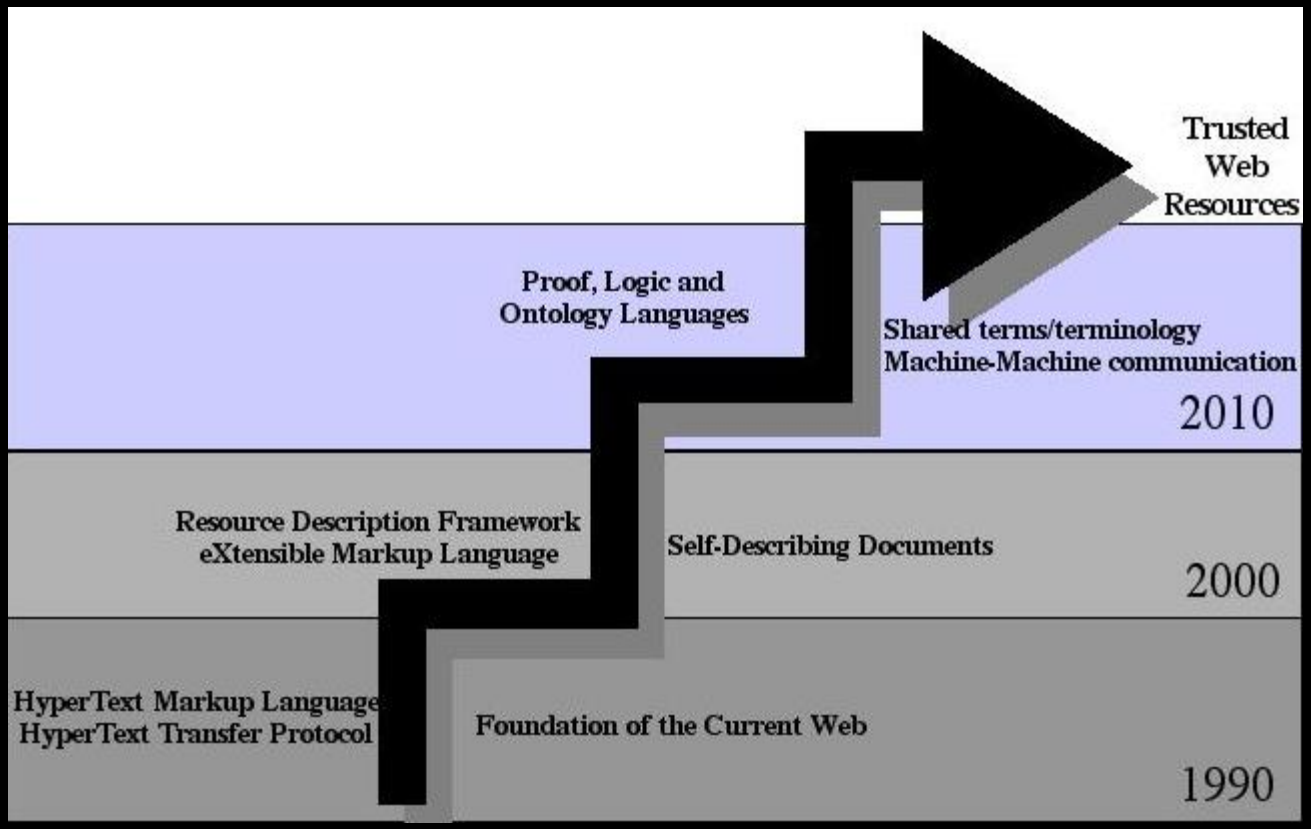
Ontologies, taxonomies, vocabularies, etc.

- Ontology - used to describe knowledge organization resources with varying degrees of structure
  - Linguistic and lexical ontologies (WordNet)
  - Vocabularies (Dublin Core)
  - Taxonomies (Yahoo, Open Directory)
  - Thesauri (AAT, INSPEC Thesaurus, MeSH)
  - Classification schemes (DDC, UDC)
- Web ontologies might use one or more of the above KO resources

Software agents

- ". . .programs that collect Web content from diverse sources, process the information and exchange the results with other programs" [1]
- Software agents will become effective as more well-defined content & other agents become available

Layers of the Semantic Web [2]



Recap vision and goal

“ …aim of the SW [Semantic Web] vision is to make Web information practically processible by a computer. Underlying this is the goal of *making the Web more effective for its users*…by the automation or enabling of things that are currently difficult to do: *locating content, collating and cross-relating content, drawing conclusions from information found in two or more separate sources*.” [5]

Caveat

"… the new technology, like the old, involves asking people to make some extra effort, in repayment for which they will get substantial new functionality -- just as the extra effort of producing HTML markup (HyperText Markup Language) is outweighed by the benefit of having content searchable on the web." [2]

OCLC Web Services

- Unbundle metadata services from CORC system
  - Extract metadata from resource
  - Automatically assign subject terms
  - Control names and subjects

OCLC Web Services

- Offer a range of terminology services that supports multiple
  - Terminology resources
  - Methods and Services
  - Protocols
  - Specifications for knowledge organization resources

Unrestricted Terminology Resources

- Available now
  - LC Name & Subject Authority Files
  - LC Children's Headings (AC Program)
- In the queue
  - ERIC thesaurus & GEM subject headings
  - FAST (under development)
  - GSAFD file (form & genre categories for fiction)
  - LC Classification
  - MeSH

Restricted Terminology Resources

- Available now
  - Dewey Decimal Classification
  - PAIS Subject Headings
  - Sears Subject Headings
- Under discussion
  - Canadian Subject Headings (NLC)
  - RVM (Bibliothèque de l'Université Laval) & RAMEAU (Bibliothèque nationale de France)
  - SWD (Die Deutsche Bibliothek)
  - Te Pātakataka (Subject headings for New Zealand Primary Schools)

Multiple Protocols

- SOAP
- HTTP Get
- HTTP Post
- Z39.50

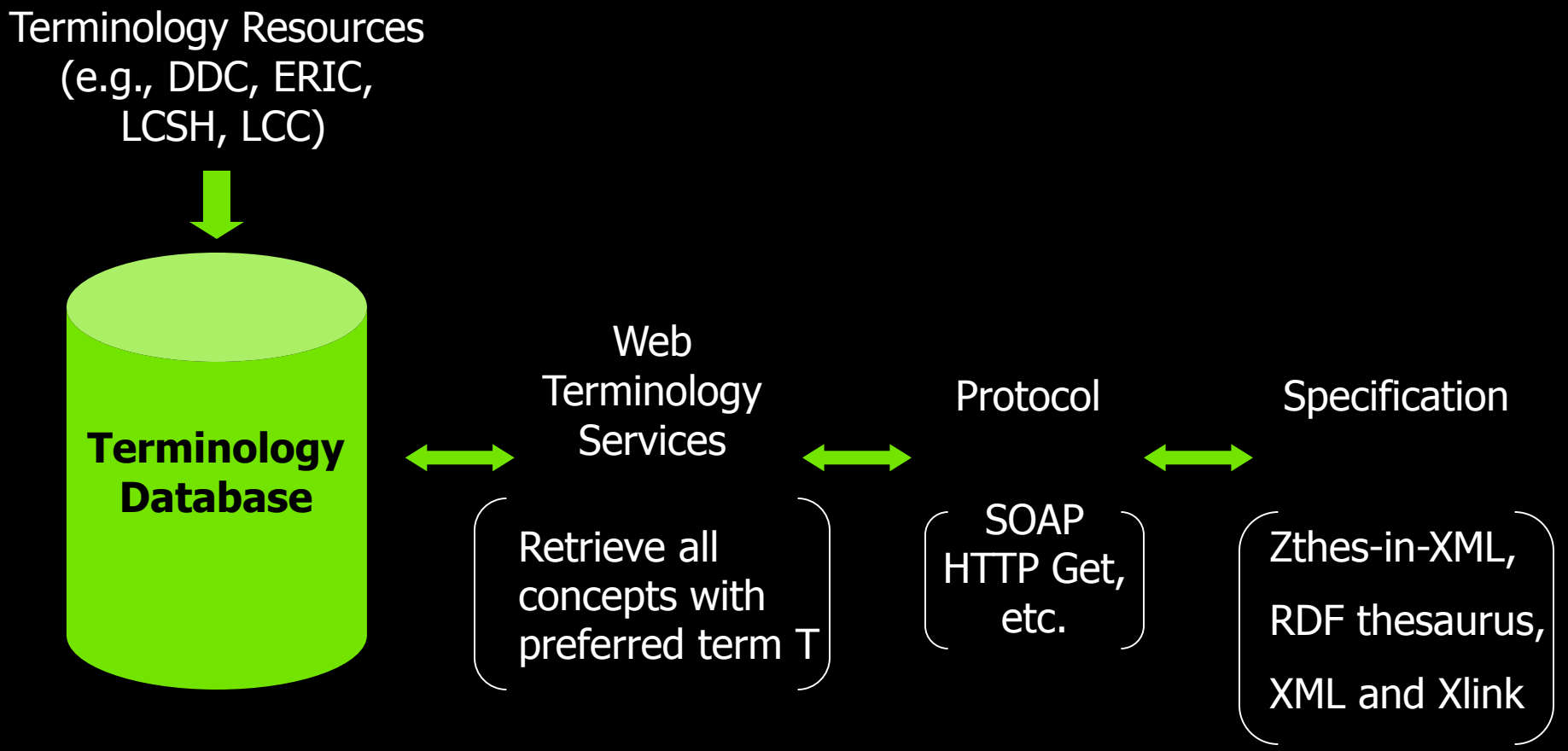
Multiple Specifications

- Zthes-in-XML
- MARC-in-XML
- RDF thesaurus specification
- XML and Xlink

Projects & Prototypes

- ePrint archive
  - Automated assignment of DDC categories and other controlled subject terms
- OCLC & Northwestern University
  - Provide a Web service to verify DDC numbers
- Prototype
  - LCCN Web Service Demo

Terminology Services



References & suggested resources

1. The Semantic Web by Tim Berners-Lee, James Hendler & Ora Lassila
   - http://www.sciam.com/2001/0501issue/0501berners-lee.html
2. Scientific publishing on the 'semantic web‘ by Tim Berners-Lee & James Hendler
   - http://www.nature.com/nature/debates/e-access/Articles/bernerslee.htm
3. Text markup and the cost of access by Jon Bosak
   - http://www.nature.com/nature/debates/e-access/Articles/bosak.html
4. XML and the Second-Generation Web by by Jon Bosak and Tim Bray
   - http://www.sciam.com/1999/0599issue/0599bosak.html
5. Building the Semantic Web by Edd Dumbill
   - http://www.xml.com/pub/a/2001/03/07/buildingsw.html

References & suggested resources

6. RDF Primer
   - http://www.w3.org/2001/09/rdfprimer/rdf-primer-20020315.html
7. Requirements for a Web Ontology Language
   - http://www.w3.org/TR/2002/WD-webont-req-20020307/