Information, Organization, and Management

Unit 7: The Semantic Web: A Web of Data

Prof. Dr. Martin Hepp

http://www.heppnetz.de
mhepp@computer.org

http://www.heppnetz.de/teaching/img/

IMG - Unit 7
Contents

• The Semantic Web Vision
• Core Components
  – A Uniform Naming Schema
  – A Data Model and Exchange Format
  – Languages for Machine-suitable Vocabularies
• Core Tools
  – Editors: Protégé
  – Repositories: OWLIM
  – Reasoners: e.g. Pellet, Racer
  – Applications: e.g. Tabulator
• Standards and Related Work
  – Query Languages: e.g. SPARQL and RQL
  – GRDDL and Microformats
Resources

• Textbook
  Antoniou/van Harmelen: 
  *A Semantic Web Primer*
  MIT Press 2004

• W3C Web Page
  – [http://www.w3.org/sw/](http://www.w3.org/sw/)
Limitations of the Current Web: Search by Words

• Limited recall and precision due to
  – Synonyms („Car“ and „automobile“)
  – Homonyms („Jaguar“ [animal] vs. „Jaguar“ [car brand])
  – Spelling variants („organize“ [AE] vs. „organise“ [BE])
  – Spelling mistakes
  – Multiple languages

• No means to specify the relation between a resource and a term
  – Sell / buy / collect stamps
Limitations of the Current Web: Combining Web Content and Reuse

• Is it cheaper to buy a certain outdoor jacket in Austria, in Switzerland, or in the US?
  – Austria: http://www.outdoorshop.co.at/
  – Switzerland: http://www.outdoorshop.ch/

• Problem
  – Identify type of good and/or make and model
  – Currency conversion
  – Freight charges
The WWW: An Unprecedented Wealth of Data

• Between 15 – 30 billion Web pages as of 2007
• Many pages are „deep pages“, i.e. entry pages to back-end databases
Semantic Web: The Vision

"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

Core Tasks in Information Processing

- **Search** for information
- **Extraction** of information
- **Processing and transactions**
- Consistent **updating** of information
- **Discovery of implicit facts**
- **Distribution and Access Control**

*Example:*

```
<PRODUCT>
  <NAME>Model 1</NAME>
  <PRICE cur="EUR" vat="yes">4</PRICE>
</PRODUCT>
```

*Order:*

5 bolts M3x15

*Questions:*

- Which documents must be updated if a specific fact becomes invalid?
- Each supplier of bolts is also a member of the category „supplier of hardware“.
- Who is allowed to access and modify which parts of the data?
What is the Semantic Web?
The Gopher – HTTP/HTML Analogy

1990: Gopher
Textual information

mid 1990s: HTML
Multimedia added

2006 and beyond: RDF, OWL
Meaning added

Gopher screenshot courtesy of the University of Minnesota.
Search by Logical Expressions

„I am looking for an accommodation in Tyrol with a shuttle service to a ski resort for Dec 8-10, 2008“

→ Web pages that contain offers for accommodations (huts, B&B, hotels,...) in those villages in Tyrol (Telfes, Neustift, Innsbruck,...), for which there is a shuttle service to a ski resort; and from this set only those with available rooms for Dec 8-10, 2005.
The Semantic Web on One Slide

- Unique **Identifiers** for Resources.
- A **Data Model** for statements about resources, which is compatible with the design principles of the Web (especially with its distributed nature)
- **Ontologies**, i.e. consensual domain models with some formal semantics.
  - Ontology **Languages**
  - Ontologies
The Semantic Web: Embedding Machine-accessible Meaning

mid 1990s
HTML – Multimedia added
Two Paths to Automation

• Empower machines to process natural language (HLT approach)

• Encode data in a machine-suitable way.

"Modell 1 kostet 4 EUR inkl. MwSt"

<PRODUCT>
  <NAME>Modell 1</NAME>
  <PRICE cur="EUR" vat="yes">4</PRICE>
</PRODUCT>

<PRICE cur="EUR" vat="included">4</PRICE>
Overview: Semantic Web Stack

• Identifiers: URIs
• Data Model: RDF
• Vocabulary: RDF-S and OWL
• Data
• Tools
• Applications
Universal Resource Identifiers (URIs)

• The Web is an information space in the sense that things in it have an “address”.

• These “addresses” (= names = identifiers) are called Universal Resource Identifiers (URI).

• An information object is “on the web” if it has a URI (those are sometimes called First Class Objects (FCO)).

• Universal: The Web is declared to be able to contain, in principle, every bit of information accessible by networks.

  http://www.ibm.com

  ftp://heppnetz.de/license.txt

  urn:ISSN:1560-1560

URI Syntax: `scheme://abcd/efg/h`

- URIs are divided into **schemes**
  - http:
  - ftp:
  - urn:
- URIs can be of **unlimited length**, which allows for importing any other universal space.
- In some schemes, **slashes** are used to indicate a **hierarchically structured name** or address.
  - `//` = top node
  - allows for relative addressing
  - relative references are important for scalability
  - can be extended (`///`, `/////` etc.)
  - not supported by all schemes (cf. RFC2396, p. 4)

URI Schemes

• Schemes **partition** the URI space into subspaces.

• Schemes are used to
  – **group similar types of information objects** (e.g. services, connection end points), and to
  – represent their different roles in the protocols.

• Schemes **can add or clarify properties of a set of URIs**, e.g. wrt the concept of **ownership** or **persistence**.

URI Scheme Prefixes

• Registered URI scheme prefixes
  – List maintained by IANA
  – http://www.iana.org/assignments/uri-schemes

• Unregistered URI schemes
  – Public Unregistered Schemes
  – Private Schemes

• Although many schemes are named after protocols, this does not imply that the only way to access the respective resource is via the named protocol. (cf. RFC 2396)

## URI Scheme Prefixes

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Name Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ftp</td>
<td>File Transfer Protocol</td>
<td>[RFC1738]</td>
</tr>
<tr>
<td>http</td>
<td>Hypertext Transfer Protocol</td>
<td>[RFC2616]</td>
</tr>
<tr>
<td>gopher</td>
<td>The Gopher Protocol</td>
<td>[RFC1738]</td>
</tr>
<tr>
<td>mailto</td>
<td>Electronic mail address</td>
<td>[RFC2858]</td>
</tr>
<tr>
<td>news</td>
<td>USENET news</td>
<td>[RFC1738]</td>
</tr>
<tr>
<td>nnrp</td>
<td>USENET news using NNTP access</td>
<td>[RFC1738]</td>
</tr>
<tr>
<td>telnet</td>
<td>Reference to interactive sessions</td>
<td>[RFC1738]</td>
</tr>
<tr>
<td>wais</td>
<td>Wide Area Information Servers</td>
<td>[RFC1738]</td>
</tr>
<tr>
<td>file</td>
<td>Host-specific file names</td>
<td>[RFC1738]</td>
</tr>
<tr>
<td>propro</td>
<td>Prospero Directory Service</td>
<td>[RFC1738]</td>
</tr>
<tr>
<td>z39.50s</td>
<td>Z39.50 Session</td>
<td>[RFC2065]</td>
</tr>
<tr>
<td>z39.50r</td>
<td>Z39.50 Retrieval</td>
<td>[RFC2065]</td>
</tr>
<tr>
<td>xid</td>
<td>content identifier</td>
<td>[RFC2282]</td>
</tr>
<tr>
<td>mid</td>
<td>message identifier</td>
<td>[RFC2392]</td>
</tr>
<tr>
<td>vennmi</td>
<td>versatile multimedia interface</td>
<td>[RFC2122]</td>
</tr>
<tr>
<td>service</td>
<td>service location</td>
<td>[RFC2609]</td>
</tr>
<tr>
<td>imap</td>
<td>Internet message access protocol</td>
<td>[RFC2192]</td>
</tr>
<tr>
<td>nfs</td>
<td>network file system protocol</td>
<td>[RFC2224]</td>
</tr>
<tr>
<td>acap</td>
<td>application configuration access protocol</td>
<td>[RFC2244]</td>
</tr>
<tr>
<td>rtsip</td>
<td>real time streaming protocol</td>
<td>[RFC2325]</td>
</tr>
<tr>
<td>tip</td>
<td>Transaction Internet Protocol</td>
<td>[RFC2371]</td>
</tr>
<tr>
<td>pop</td>
<td>Post Office Protocol v3</td>
<td>[RFC2284]</td>
</tr>
<tr>
<td>data</td>
<td>data</td>
<td>[RFC2397]</td>
</tr>
<tr>
<td>dav</td>
<td>dav</td>
<td>[RFC2518]</td>
</tr>
<tr>
<td>opaqueoctoken</td>
<td>opaqueoctoken</td>
<td>[RFC2518]</td>
</tr>
<tr>
<td>sip</td>
<td>session initiation protocol</td>
<td>[RFC3261]</td>
</tr>
<tr>
<td>sipt</td>
<td>secure session initiation protocol</td>
<td>[RFC3261]</td>
</tr>
<tr>
<td>tel</td>
<td>telephone</td>
<td>[RFC2806]</td>
</tr>
<tr>
<td>fax</td>
<td>fax</td>
<td>[RFC2806]</td>
</tr>
<tr>
<td>modern</td>
<td>modern</td>
<td>[RFC2806]</td>
</tr>
<tr>
<td>ldap</td>
<td>Lightweight Directory Access Protocol</td>
<td>[RFC2259]</td>
</tr>
<tr>
<td>https</td>
<td>Hypertext Transfer Protocol Secure</td>
<td>[RFC2818]</td>
</tr>
</tbody>
</table>

URI, URL, URN – Terminology

- **URI**: Universal (Uniform) Resource Identifier
- **URL**: Uniform Resource Locator
- **URN**: Uniform Resource Names
- Classical View: URI is either URL or URN
- Contemporary View:
  - URI space is partitioned by URI schemes into subspaces
  - **URN is one of the URI schemes** and defines a URI subspace; the respective URI scheme prefix is “urn:”
  - **URL is an informal concept**: A URL is a URI that identifies a resource via a representation of its primary access mechanism.

[W3C: URIs, URLs, and URNs: Clarifications and Recommendations 1.0, http://www.w3.org/TR/2001/NOTE-uri-clarification-20010921/]
Uniform Resource Names (URNs)

• A subspace of the URI space
• Partitioned by the URI Scheme prefix “urn:”
• Followed by a URN Namespace Identifier or ID (URN NID)
• urn:ISSN:1560-1560

[W3C: URIs, URLs, and URNs: Clarifications and Recommendations 1.0, http://www.w3.org/TR/2001/NOTE-uri-clarification-20010921/
URN Namespace Identifiers

• IANA Registry
  – [http://www.iana.org/assignments/urn-namespaces](http://www.iana.org/assignments/urn-namespaces)

• Examples
  – ISSN
  – ISBN
  – swift
URI Axioms (1)

- **Axiom 0 (Universality 1):** Any resource anywhere can be given a URI.
- **Axiom 0a:** Any resource of significance should be given a URI.
- **Axiom 1 (Global scope):** It doesn’t matter to whom or where you specify a URI; it will have the same meaning.

URI Axioms (2)

• **Axiom 2a (sameness):** A URI will repeatedly refer to the “same” thing (but the concept of identity might vary).

• **Axiom 2b: identity:**
  – The significance of identity for a given URI is determined by the person who owns the URI, who first determined what it points to.
  – The concept of a URI itself does not define the identity properties that exist between a URI and the resource associated with it.
  – The identity relationship, degree of persistence, and whether reuse is possible or allowed depends on the URI scheme and the owner.

• **Note 1:** The definition of “owner” may vary from URI scheme to scheme.

• **Note 2:** Two URIs are the same if (and only if) they are the same character for character. There is no definite canonicalization except for a few conversions (hex-encoding of special characters etc.).

URI Axioms (3)

- **Axiom 3 (non unique):** URI space does not have to be the only universal space.
  - Any new space of identifiers or address space can be represented as a subset of URI space.

- **Axiom “Opacity of URIs”:** The only thing one can use an identifier for is to refer to an object.

URI: Resources

- A URI represents a **resource**. A "resource" is a conceptual entity (a little like a Platonic ideal).
- A resource may be **generic** in that "as a concept it is well specified but not so specifically specified that it can only be represented by a single bit stream". As an example, successively specific resources might be:
  - The Bible
  - The Bible, King James Version
  - The Bible, KJV, in English
  - A particular ASCII rendering of the KJV Bible in English
- **The authority which allocates the URI is the authority which determines to what it refers**: Therefore, that authority determines to what extent that resource is generic or specific.
- Not all resources are network "retrievable"; human beings, corporations, and bound books can also be considered resources (see RFC2396).

URI: Fragment Identifier

- Fragment identifiers (#) are not part of a URI
  - http://www.ibm.com/aboutus.html#address
- A fragment identifier is only meaningful when a URI reference is intended for retrieval and the result of that retrieval is a document for which the identified fragment is consistently defined.
  - (I disagree 😊)

Q: RFC2396, p. 14f.
URI – Summary

- URI, URN, URL
- URI Schemes ≠ Protocols
- URN Namespace Identifiers (NIDs)
- Generic URIs, Persistence, and Identity
Cool URIs don’t change

• Cool
  http://www.uni-wuerzburg.de/faktultaeten/wifak
  http://www.uni-wuerzburg.de/studienordnungen/wifak/diplom-bwl/2004-07-31

• Uncool
  http://www.uni-wuerzburg.de/meier/showFak.asp?id=6

T. Berners-Lee: Cool URIs don’t change, http://www.w3.org/Provider/Style/URI
Data Model: RDF

Peter knows Mary
Mary works at Siemens
Siemens produces computers.
Data Model: RDF

Peter knows Mary
Mary works at Siemens
Siemens produces computers.
Fundamental Idea:
Link Data by Unique Identifiers for Entities
... on at the Web scale...
The W3C RDF Validator

RDF Validation Results

Source | Triples | Messages | Graph | Feedback | Back to Validator input

Validation Results
Your RDF document validated successfully.

The original RDF/XML document

```xml
1: <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
2:   xmlns:dc="http://purl.org/dc/elements/1.1/">
3:   <rdf:Description rdf:about="http://www.w3.org/">
5:   </rdf:Description>
6: </rdf:RDF>
```

Graph of the data model

http://www.w3.org/RDF/Validator/
Vocabulary

• We need to use a shared, machine-readable vocabulary.
• Ideally, a computer should be able to „understand“ part of the vocabulary, e.g.,
  – Spot inconsistencies and
  – Compute implicit facts
Vocabulary: Languages

• RDF-S
• OWL DLP
• OWL Lite
• OWL Full
• Rule Languages
• Useful Language Fragments
Data

• We have to augment existing Web data by references to such vocabularies etc.
• This is called „Annotation“.
• It can be done by humans, semi-automatically, or fully automatically.
• Problem: Data and annotation must be kept in sync.
Tools
Applications

• Life Sciences
• Social Network Tools
• Semantic Desktop
• Knowledge Management
• HR / Skills Matchmaking
• Semantic Search Engines
• Etc.
Components of the Semantic Web

External Resources
- urn:ISSN:1234-3456
- tel:+1-239-590-7311

Network-accessible Resources
- mailto:psmith@ibm.com
- ftp://www.ibm.de/disclaimer.txt
- http://purl.org/dc/elements/1.1/creator

URI Space

+ RDF statements
<urn:ISSN:1234-3456> <http://purl.org/dc/elements/1.1/creator> <mailto:psmith@ibm.com>

+ formal semantics: RDF-S $\rightarrow$ OWL:
  
  A creator is a subclass of person.
  isMarriedTo is a symmetric property.

Ontologies

Tools
- Jena (HP)
- Snobase (IBM)
- ICS-FORTH RDFSuite
- openRDF Sesame

APIs
- RDF Storage and Retrieval
- Business Applications!
...the Semantic Web is NOW!

- Technology is pretty mature
- Major corporations are preparing products and services
- Lots of open-source software
- Semantic Wikis and DBpedia
- Open Linked Data initiative
Thank you!

The slides of today’s class will be available at [http://www.heppnetz.de/teaching/img/](http://www.heppnetz.de/teaching/img/) shortly.