FOUNDATIONS OF SEMANTIC WEB TECHNOLOGIES

Overview and XML

Sebastian Rudolph
Agenda

• Introduction of Lecturer
• Organizational Matters
• What is the “Semantic Web”? 
Agenda

- Introduction of Lecturer
- Organizational Matters
- What is the “Semantic Web”? 
Introduction of Lecturer

Prof. Dr. Sebastian Rudolph

Since Apr 2013 Full Professor for Computational Logic at the Computer Science Department, TU Dresden

2006 – 2013 Research Assistant → Project Leader → Privatdozent at the Chair of Knowledge Management, Institute AIFB, University of Karlsruhe → Karlsruhe Institute of Technology

2003 – 2006 Research Assistant at the Chair of Psychology of Teaching and Learning, TU Dresden

2000 – 2003 PhD Scholarship Holder Graduate School, TU Dresden

1995 – 2000 Studies for high-school teaching (Math, Physics, CS), TU Dresden
Introduction of Lecturer

scientific interests

- logic-based knowledge representation and reasoning
- semantic technologies
- complexity and decidability problems
- computational linguistics
- theory of databases
- (and much more)
Agenda

- Introduction of Lecturer
- Organizational Matters
- What is the “Semantic Web”?
Organizational Matters: Time and Place

- Wednesdays, 09:20 – 10:50 (2.DS) and 11:10 – 12:40 (3.DS)
- Mondays, 09:20 – 10:50 (2.DS)
- exact schedule see webpage
- INF E005
- accompanying web page:
Literature

Hitzler, Krötzsch, Rudolph, Sure
“Semantic Web Grundlagen”
Springer-Verlag

Hitzler, Krötzsch, Rudolph, Sure “Foundations of Semantic Web Technologies”
CRC Press
Agenda

- Introduction of Lecturer
- Organizational Matters
- What is the “Semantic Web”?
The Web

The Web is at the heart of the transition from industrial to information society, providing the infrastructure for a novel quality of handling information in terms of retrieval and provision

- high availability
- high up-to-date-ness
- low cost
The Web

Commercialization on all levels
The Web

Commercialization on all levels
Further aspects of daily life are being “webized”:

- authorities, administration (eGovernment)
- education (eLearning, eEducation)
- social contacts (social networking platforms, dating sites)
- everyday life?
What means “Semantic”?

Syntax vs. Semantics

Syntax (from greek συνταξις composition, sentential structure) denotes the (normative) structure of data, i.e., it characterizes what makes data “well-formed”

Semantics (greek σηµαυτικος belonging to the sign) denotes the meaning of data, i.e., it characterizes what conclusions can be drawn from it.

\[
\begin{align*}
4+3 &= 12 & 3+4 &= 7 \\
\text{syntactically wrong} & & \text{syntactically correct} & & \text{syntactically correct}
\end{align*}
\]
Problems of the Web

- plethora of information
- targeted at human end user
Problems of the Web

- plethora of information
- targeted at human end user
Problems of the Web

- plethora of information
- targeted at human end user
Problems of the Web

- plethora of information
- targeted at human end user

Guilin Qi

Professor
School of Computer Science and Engineering
Southeast University
China

Tel: +86 (0) 25 52090910
Fax: +86 (0) 25 52090880
E-mail: gqi@seu.edu.cn

Curriculum Vitae (PDF)

Brief Introduction:

Dr. Guilin Qi is a professor working at Southeast University in China. His research topics include knowledge representation and reasoning, semantic Web, uncertainty reasoning. His current research interests include the areas of

- Knowledge representation: belief merging, belief revision, inconsistency handling, nonmonotonic reasoning, information fusion, argumentation, paraconsistent logic
<h1>Ian Horrocks</h1>
<table>
<tr>
<td class="personImg">
    <img src="IH005-1.jpg" alt="Photo Ian Horrocks"/>
</td>
<td>
    <div class="personinfo">
        <div>Professor Ian Horrocks FRS</div>
        <div>Professor of Computer Science</div>
        <div>Fellow, <a href="http://www.oriel.ox.ac.uk">Oriel College</a></div>
        <div>ian.horrocks@cs.ox.ac.uk</div>
        <div>+44 1865 273939</div>
        <div>+44 1865 273839 (fax)</div>
        <p>Wolfson Building, Parks Road, Oxford OX1 3QD</p>
    </div>
</td>
</tr></table>
Problems of the Web

- localizing information problematic
- today’s search engines good but mostly keyword-based
- desirable: search for content → semantic search
Problems of the Web

- Heterogeneity of present information on diverse levels:
  - character encoding (e.g. ASCII vs. Unicode)
  - used natural languages
  - positioning of information on webpages

- desirable: cross-web information integration
Problems of the Web

- **implicit knowledge**, i.e. many pieces of information are not provided explicitly, but follow from the combination of the given data
- requires methods from formal logics
- automated deduction
Problems of the Web

Approaches toward a solution:

1. Ad hoc: Deployment of AI methods (most notably NLP techniques) to evaluate existing unstructured information on the Web

2. A priori: Structure information on the Web at authoring time in a way facilitating later automated deployment
Problems of the Web

Approaches toward a solution:

1. Ad hoc: Deployment of AI methods (most notably NLP techniques) to evaluate existing unstructured information on the Web

2. A priori: Structure information on the Web at authoring time in a way facilitating later automated deployment

⇒ Semantic Web
Problems of the Web

two essential prerequisites for the implementation:

1. open standards for describing information
   - clearly defined
   - flexible
   - extendable

2. methods for eliciting information from such descriptions
1994  First public presentation of the Semantic Web idea
1998  Start of standardization of data model (RDF) and a first ontology languages (RDFS) at W3C
2000  Start of large research projects about ontologies in the US and Europe (DAML & Ontoknowledge)
2002  Start of standardization of a new ontology language (OWL) based on research results
2004  Finalization of the standard for data (RDF) and ontology (OWL)
2008  Standardization of a query language (SPARQL)
2009  Extension of OWL to OWL 2.0
2010  Standard Rule Interchange Format (RIF)
Agenda

- XML – Motivation/Idea
- XML – Syntax
- IRIs
- Name Spaces
Agenda

- XML – Motivation/Idea
- XML – Syntax
- IRIs
- Name Spaces
Annotation with Mark-up Languages

- basic idea of mark-up: endow (unstructured) text with additional information (or structure)
- synonym: annotate text

\[
\begin{align*}
\text{text} & \quad = \quad \text{data} \\
\text{additional information} & \quad = \quad \text{metadata}
\end{align*}
\]
Annotation with Mark-up Languages

- common strategy: include to-be-annotated text in so-called tags:

```
<tag_name> ...Text... </tag_name>
```

- opening tag
- closing tag

- Additional information is read and interpreted by processing software
Annotation with Mark-up Languages

- most prominent example: HTML tags encode visual presentation information:
  
  `<i>This book</i> has the title `<b>Foundations of Semantic Web Technologies</b>`.

- Output of web browser:
  
  `This book` has the title `Foundations of Semantic Web Technologies`.

- Strategy also suited for annotation of content, e.g.:
  
  `<firstname>Sebastian</firstname>` `<lastname>Rudolph</lastname>` works in `<city>Dresden</city>`.
Annotation with Markup-Languages

- nesting of tags is permitted

```xml
<lecture>
  <title>
    Deduction Systems
  </title>
  <lecturer>
    <title>
      Prof. Dr.
    </title>
    <firstname>
      Sebastian
    </firstname>
    <lastname>
      Rudolph
    </lastname>
  </lecturer>
</lecture>
```
Annotation with Markup-Languages

<lecture>
  <title>
    Deduction Systems
  </title>
  <lecturer>
    <title>
      Prof. Dr.
    </title>
    <firstname>
      Sebastian
    </firstname>
    <lastname>
      Rudolph
    </lastname>
  </lecturer>
</lecture>

- nesting of tags is permitted
- multiple usage of tags is permitted
• nesting of tags is permitted
• multiple usage of tags is permitted
• XML tags constitute a tree structure
Agenda

- XML – Motivation/Idea
- XML – Syntax
- IRIs
- Name Spaces
XML

- eXtensible Markup Language
- Origin: structured text (HTML4.0 ∈ XML ⊂ SGML)
- web standard (W3C) for data exchange:
  - input and output data can be described by means of XML
  - industry only has to agree on standardized tag names (the vocabulary)
- complementary language for HTML:
  - HTML describes presentation
  - XML describes content
- database perspective: XML as a data model for semi-structured data
XML-Syntax – Preamble

- XML document is a text document
- starts with declaration
  - contains version number of the standard used
  - optional: character encoding information

`<?xml version="1.0" encoding="utf-8"?>`
XML-Syntax – XML element

XML element:
- description of an object enclosed by matching tags
- content of an elements: text and/or further elements (arbitrary nesting possible)
- empty elements: `<year>`</year> short: `<year/>
- “outermost” element is called root element (and there can be only one per document)

```
<author>
  <firstname> Serge </firstname>
  <lastname> Abiteboul </lastname>
  <email> sab@abc.com </email>
  email address may be wrong!
</author>
```
XML-Syntax – XML attributes

XML attribute:
- pair of name and string-value in start or self-closing tag
- associated with one XML element
- alternative option for describing data

Further possible description of the same data:

```xml
<author email="sab@abc.com">
  <firstname>Serge</firstname>
  <lastname>Abiteboul</lastname>
</author>
```

```xml
<author firstname="Serge" lastname="Abiteboul" email="sab@abc.com"/>
```
HTML vs. XML

- HTML: fixed vocabulary (set of tags) and semantics (visual presentation of text)
- XML: free choice of names for describing application-specific syntax and semantics
- XML $\subset$ SGML

```
< Bib id="o1">
  < paper id="o12">
    < title > Foundations of Databases </ title >
    < author >
      < firstname > Serge </ firstname >
      < lastname > Abiteboul </ lastname >
    </ author >
    < year > 1997 </ year >
    < publisher > Addison Wesley </ publisher >
  </ paper >
... 
</ Bib>
```

```
<h1> Bib </h1>
<p>
  <i> Foundations of Databases </i>
  Serge Abiteboul
  <br> Addison Wesley, 1997
</p>
```

HTML  XML
Agenda

- XML – Motivation/Idea
- XML – Syntax
- IRIs
- Name Spaces
IRIs – Idea

- IRI = Internationalized Resource Identifier
- serve for denoting resources in a world-wide unique way
- a resource can be any object that has (in the context of a given application) a clear identity (e.g. books, cities, persons, publishers, relations between those, abstract concepts etc.)
- in certain domains, something similar already exists: ISBN number for books
IRIs – Syntax

- extension of the notion of URLs; not every IRI relates to a Web document but mostly a Web document is referred to by using its URL as IRI
- starts with the so-called IRI schema, which is separated by a colon (:) from the subsequent part (e.g.: http, ftp, mailto)
- IRIs often hierarchically structured
IRIs – Self-defined IRIs

- necessary, if for a certain resource no IRI exists or is known (yet)
- strategy in order to avoid unintentional double use of an IRI for different things: use http-IRIs of a webpage that you control
- allows for providing a documentation describing the IRI under this address
The Describing vs. the described

- Separation of IRIs for (non-information) resources and their documentation (information resources) by IRI references (appended fragments starting with “#”) or content negotiation

- e.g.: as a IRI for Shakespeare’s “Othello”, http://de.wikipedia.org/wiki/Othello should not be used, but rather http://de.wikipedia.org/wiki/Othello#IRI
Agenda

- XML – Motivation/Idea
- XML – Syntax
- IRIs
- Name Spaces
XML Name Spaces: Motivation

- in XML documents, element and attribute names ("markup vocabulary") have universal validity
- in an XML application, these names are interpreted uniformly
- if XML data from several sources is merged, name conflicts / clashes may occur
- name spaces help avoid such conflicts
XML Name Spaces

- XML name spaces are similar to the notion of modules in programming languages
- disambiguation of tag names through usage of different “prefixes”
- a prefix is separated from a local name by a colon (:), thereby `prefix:name` tags come into being
- name space bindings are ignored by some tools: so-called “shallow name spaces”
Name Space Bindings

• prefixes are associated with name space IRIs by inserting an attribute `xmlns:prefix` into the relevant element or some of its predecessor elements: `prefix:name_1, ..., prefix:name_n`
• the value of the attribute `xmlns:prefix` is an IRI, that may point to a description of the syntax of the name space
• an element can use bindings for several (different) name spaces by using separate attributes `xmlns:prefix_1, ..., xmlns:prefix_m`
Example: Without Name Spaces

```xml
<lecture>
  <title> Deduction Systems </title>
  <lecturer>
    <title> Prof. Dr. </title>
    <firstname> Sebastian </firstname>
    <lastname> Rudolph </lastname>
  </lecturer>
</lecture>
```

title is an ambiguous element name
Two Distinct Name Spaces

<lec:lecture xmlns:lec = "http://www.example.org/lectures"
             xmlns:per = "http://www.example.org/person">
  <lec:title> Deduction Systems </lec:title>
  <lec:lecturer>
    <per:title> Prof. Dr. </per:title>
    <per:firstname> Sebastian </per:firstname>
    <per:lastname> Rudolph </per:lastname>
  </lec:lecturer>
</lec:lecture>

**title** has been disambiguated by using the prefixes **lec** and **per**
Agenda

- XML – Motivation/Idea
- XML – Syntax
- IRIs
- Name Spaces