Semantic Web Technologies
Introduction and Course Outline

Heiko Paulheim
Hello

• Prof. Dr. Heiko Paulheim
• Assistant Professor
• Research Interests:
  – Semantic Web and Linked Open Data
  – Data Mining with Linked Open Data
  – Ontology Matching
  – Data Quality and Data Cleaning
  – Outlier Detection
• Room: B6 – C1.08
• Consultation: by appointment
• Heiko will teach the lectures
Hello

• André Melo
• Researcher
• Research Interests:
  – Semantic Web Mining
  – Hierarchical Classification
  – Frequent Itemset Mining
• Room: B6 – C1.06
• Consultation: by appointment
• André will teach the exercises
Introduction and Course Outline

• Administration
• Introduction
  – Vision of the Semantic Web
  – Building blocks of the Semantic Web
  – Technical foundations
Course Organization

• Lecture
  – Semantic Web standards and languages
  – Programming for the semantic web
  – Creating semantic web data

• Exercise
  – Understand semantic web principles, play with real data

• Project Work
  – teams of three students build a Semantic Web application
  – teams may choose their own data sets and tasks
    (in addition, we will propose some pointers for ideas)
  – write summary about project, present project results

• Final exam
  – 50 % written exam
  – 50 % project work
Waiting List

• Please register via ILIAS
  – if you've not yet done so
• This year, there is a waiting list
• Maximum participants: 30
• We apply a three-strikes-out rule
  – People not showing up in the first three lectures are assumed not to take part
  – We will reassign places after three weeks
• If you know you will not be attending (or just attending to listen)
  – please let us know
Exercise Timeslot

• There is a clash of the exercise with Model Driven Development
  – possible alternative: move exercise to 12-13.30
  – provided that we get a room

• Quick Poll
Course Contents and Schedule

- Today: Introduction
- 18.09.15: Knowledge Representation with RDF
- 25.09.15: Simple ontologies with RDF Schema
- 02.10.15: Linked Open Data, Programming the Semantic Web
- 09.10.15: Querying the Semantic Web with SPARQL
- 16.10.15: Work on project proposals
- 23.10.15: Complex ontologies with OWL
- 30.10.15: Reasoning with complex ontologies
- 06.11.15: Ontology engineering, top level ontologies
- 13.11.15 – 4.12.15: Project work, no lectures, no exercises
- 11.12.15: Final project presentation
Deadlines

• Submission of project work proposal
  – Wednesday, October 21\textsuperscript{st}, 23:59

• Submission of final project work report
  – Wednesday, December 9\textsuperscript{th}, 23:59
Course Organization

• Lecture Webpage: Slides, Announcements, Web Links
  – hint: look at version tags!

• Additional Material

• Time and Location
  – Lecture: Friday, 12.00 - 13.30, Room B6 23-25, A101
  – Exercise: Tuesday, 13.45 - 15.15, Room A5,6, C013
    (starting next week!)
Further Reading and Software

• Follow the links on the website
  – Most material is available online

• Programming environment
  – Eclipse (Java)
  – JENA framework (introduction will be given in the lecture)

• Ontology engineering environment
  – Protégé
  – http://protege.stanford.edu/
Warning

• This lecture contains
  – cartoons
  – interactive teaching elements
  – Java code
  – some weird philosophy

• ...have fun! :-}
Questions?
What is the Semantic Web?

• 2001 article by Tim Berners-Lee, Jim Hendler, and Ora Lassila:

„The Web is the killer app of the Internet. The Semantic Web is another killer app of that magnitude.“

Web vs. Internet?

The “Classic” Web

• Uses HTTP protocol and URLs
• HTML as a markup language
  – plus CSS, JavaScript, …
  – plus a few other, more or less standardized formats (GIF, JPEG, Flash, …)
• Browser as a universal client
The “Classic” Web

- Hypertext: linked documents

The World Wide Web

The World Wide Web was Established in the 90s by Tim Berners-Lee at CERN.

Tim Berners-Lee

Tim Berners-Lee (born 1955) is one of the inventors of the World Wide Web.

CERN

The CERN is a European research center, located close to Geneva.
A Short History of the Web

• Let's see what you know...
• Try to find the correct chronological ordering of the following events:

1. First version of HTML
2. Wikipedia goes online
3. Foundation of Skype
4. First Web catalogue
5. Foundation of the W3C
6. First Search Engine
7. Foundation of Twitter
8. HTTP Standard
9. 500 Servers online
10. Foundation of Facebook
11. Dotcom bubble and stock market crash
12. First version of Internet explorer
13. Foundation of Google
14. First domain registered
15. First version of Firefox
16. TCP/IP Standard
17. 1,000 computers online
18. 1,000,000 computers online
19. 1,000,000,000 computers online
20. First multi user online game
A Short History of the Web

- 1974: TCP/IP Standard
- 1979: First Multi User Game
- 1985: First domain registered
  1,000 computers online
- 1989: Hypertext concept
  by Tim Berners-Lee
- 1991: First HTML version (20 elements)
- 1992: ~1,000,000 computers online
- 1993: Mosaic-Browser,
  around 500 web servers worldwide
- 1994: Full text search engines (WebCrawler, Lycos)
  Web catalogues (Yahoo!, AltaVista)
  Foundation of the W3C
A Short History of the Web

• 1995: Internet Explorer
• 1996: HTTP Standard
• 1998: Foundation of Google
• 2000: Dotcom Bubble, Stock Market Crash
• 2001: Foundation of Wikipedia
• 2003: Foundation of Skype
• 2004: Foundation of Facebook,
  First version of firefox
• 2006: Foundation of Twitter and WikiLeaks
• 2014: 1,000,000,000 computers online
The Dotcom Bubble and Stock Market Crash

Evolution of the Web

http://www.phdcomics.com/comics.php?n=1456
The “Classic” Web

In the eyes of a human

<html>
...<b>Dr. Mark Smith</b><i>Physician</i>
Main St. 14
Smalltown
Mon-Fri 9-11 am
Wed 3-6 pm
...
</html>

Dr. Mark Smith
Physician
Main St. 14
Smalltown
Mon-Fri 9-11 am
Wed 3-6 pm

Print in bold: "hmf298hmhusda"
Print in italics: "mj2i9ji0"
Print normal: "fdsah
02hfadsh0um2m0adsmf0ihm
asdfsdfdsaf298ndsfmij32mio
lk2mjpoimjiofdpmajiojmjm"
Full text search by keywords (e.g., Google):

• „Mark Smith“
• „Physician in Smalltown“
• „Doctor in Smalltown“
• „Physician in Smalltown with opening hours on Wednesday afternoon“
• „Somebody in Smalltown who can fix a broken leg“

→ “classic” Web is too inflexible for useful search
→ hard to use for intelligent agents
Problems of the “Classic” Web

• Finding information
  – Keyword based search instead of natural language questions
  – Different natural languages
  – Synonyms, homonyms and polysemes
  – Ambiguity of natural language

• Processing information
  – Formats and encodings

• Making use of information
  – Distributed across pages
  – e.g., a book's author on the publishers site, address on his/her personal page

Homonyms and Polysemes
Untyped Links

Bush Era Law Could Get You 20 Years in Prison For Clearing Your Browser History
Example: Wolfram Alpha
Example: Wolfram Alpha
Example: Google Knowledge Graph

- Paradigm shift in Web Search
  - “Things, not strings”
- Contains structured data for many entities
- Displayed to the user in a uniform way
- Connect entities via named links

note: these are typed links!
Solutions

Lectures: Web Mining, Information Extraction

Extract information from the Web

WWW

Create machine interpretable information

Lecture: Semantic Web Technologies
The Semantic Web Idea

• Provide information in machine interpretable form

• Make (semantic) links between (data) documents usable

• Allow reasoning

• Facilitate useful (!) complex queries
Semantic Web – Architecture

here be dragons...

Semantic Web Technologies (This lecture)

Technical Foundations

User Interface and Applications

Trust

Proof

Unifying Logic

Query: SPARQL

Ontology: OWL

Rules: RIF

Schema: RDF-S

Data Interchange: RDF

Data Interchange: XML

URI

Unicode

Berners-Lee (2009): Semantic Web and Linked Data
Uniform Resource Identifiers (URIs)

- Proposed by Tim-Berners-Lee as „Universal Resource Identifier“ (IETF RFC 1630)
- Used for naming and finding resources on the Web

\[
\text{URI} = \text{scheme} \; ":\" \; \text{hier-part} \; [ \; "?" \; \text{query} \; ][ \; "#" \; \text{fragment} \; ]
\]

http://example.com:8042/over/there?name=ferret#nose
URIs vs. URLs

• Uniform Resource Locators (IETF RFC 1738, 1994) are a *subset* of URIs

• URIs can refer to *arbitrary* things
• A URL refers to a resource on the Web

• Typical URL prefixes
  – http
  – ftp
  – mailto
  – telnet
  – file
  – ...

URLs on the Web

- Most common usage: Hyperlinks in HTML documents
- Links usually do not carry any meta information

Tim Berners-Lee

Tim Berners-Lee (born 1955) is one of the inventors of the World Wide Web. ...

The World Wide Web

The World Wide Web was initiated in the 90s by Tim Berners-Lee at CERN. ...
Character Sets on the Web

• ASCII („American Standard Code for Information Interchange“)
  ISO 646 (1963), 127 characters, 95 of which are printable:

  !"#$%&'()*+,-./0123456789:;<=>?
@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`
abcdefghijklmnopqrstuvwxyz{|}~

• Extension to 8 Bit: ISO 8859-1 to -16
  (1998)
  – covers major European languages
  – most well known: 8859-1 („Latin-1“)

• The Web, however, speaks many more languages...
The Multilingual Web

TOWER OF BABEL

HOW DARE YOU!!! FROM NOW ON YOU’LL TALK....

...UNICODE!

HE WAS NOT AMUSED

Unicode

- ISO 10646
  - first version 1991 (Europe, Near East, India)
  - Unicode 8.0 (June 2015)
  - defines ~120,000 characters
  - covers even very exotic languages
Unicode

Source: Wikimedia Commons
Information Representation in XML

XML (eXtensible Markup Language)

- A W3C standard since 1998
- Universal format for data exchange

```xml
<physician>
  <name>Dr. Mark Smith</name>
  <address>
    <street>Main St.</street>
    <number>14</number>
    <city>Smalltown</city>
  </address>
  <telephone>
    <number>+44 123 456789</number>
  </telephone>
  <hours>
    <monday>9-11 am</monday>
    <tuesday>9-11 am</tuesday>
  ...
  </hours>
</physician>
```
XML: Basic Concepts

• Tags (arbitrarily definable):
  – Form pairs:
    <physician> … </physician>
  – ...or empty element tags
    <young />

• Attributes:
  <physician location="Smalltown"> 

• Tags are nested (with exactly one root element):
  <physician>
    <address> … </address>
  </physician>
XML: Well-formed Documents

<physician>
  <name>Dr. Mark Smith</name>
  <address>
    <street>Main St.</street>
    <number>14</number>
    <city>Smalltown</city>
  </address>
  <telephone>
    <number>+44 123 456789</number>
  </telephone>
  <hours>
    <monday>9-11 am</monday>
    <tuesday>9-11 am</tuesday>
  ...
</physician>
HTML and XML

• HTML documents look like XML documents
  – ...but they are usually not well-formed!

  `<p>Look at this! <img src=smiley.gif> <br>`

• XHTML: HTML as well-formed XML documents
• A W3C standard since 2000

  `<p>Look at this! <img src="smiley.gif"/> <br/> </p>`
XPath: Accessing Information in XML

- Query language for XML
- A W3C standard since 1999 (Version 2.0: 2010)

/physician[name='Dr. Mark Smith']/telephone/number
XSLT: Transformation of XML Documents

• Stylesheet based processing of XML documents
• A W3C standard since 1999
• Uses XPath

<physician>
  <name>Dr. Mark Smith</name>
  <address>
    <street>Main St.</street>
    <number>14</number>
    <city>Smalltown</city>
  </address>
  <telephone>
    <number>+44 123 456789</number>
  </telephone>
  <hours>
    <monday>9-11 am</monday>
    <tuesday>9-11 am</tuesday>
    ...
  </hours>
</physician>

<xsl:template match="/physician">
  <b><xsl:value-of select="name"/></b>
</xsl:template>

<b>Dr. Mark Smith</b>
Namespaces in XML

- Elements with the same name can occur in different places
  - ...but the contents and semantics may differ
- How can we tell them apart?

```xml
<physician>
  <name>Dr. Mark Smith</name>
  <address>
    <street>Main St.</street>
    <number>14</number>
    <city>Smalltown</city>
  </address>
  <telephone>
    <number>+44 123 456789</number>
  </telephone>
  <hours>
    <monday>9-11 am</monday>
    <tuesday>9-11 am</tuesday>
    ...
  </hours>
</physician>
```
Namespaces in XML

- Namespace definition using prefixes (Notation: prefix:name)
- Each namespace itself is a URI
- Default namespaces may be defined

```xml
  <name>Dr. Mark Smith</name>
  <addr:address>
    <addr:street>Main St.</addr:street>
    <addr:number>14</addr:number>
    <addr:city>Smalltown</addr:city>
  </addr:address>
  <telephone>
    <number>+44 123 456789</number>
  </telephone>
  <hours>
    <monday>9-11 am</monday>
    <tuesday>9-11 am</tuesday>
    ...
  </hours>
</physician>
```
XML: Document Type Definition (DTD)

• Definiert valid elements for a class of XML documents
  – Names
  – allowed attributes
  – allowed nested child elements

• DTD is a part of the W3C's XML specification

• XML documents matching a DTD are called “valid”
XML: Document Type Definition (DTD)

<!DOCTYPE physician [ 
  <!ELEMENT physician ( 
    name, 
    address*, 
    telephone?, 
    fax?, 
    hours)>

<!ELEMENT address ( 
  street, 
  number, 
  city)> 

<!ELEMENT street (#PCDATA)> 

... ]>

<!DOCTYPE physician SYSTEM "physician.dtd"> 

<physician> 
  <name>Dr. Mark Smith</name> 
  <address> 
    <street>Main St.</street> 
    <number>14</number> 
    <city>Smalltown</city> 
  </address> 
  <telephone> 
    <number>+44 123 456789</number> 
  </telephone> 
  <hours> 
    <monday>9-11 am</monday> 
    <tuesday>9-11 am</tuesday> 
    ... 
  </hours> 
</physician>
XML: Document Type Definition (DTD)

• Definition of child elements and their order
  `<!ELEMENT address (street, no, line*, zip, city, state?)>
  – ?, + and * mark optional and possible multiple elements`

• Definition of attribute lists
  `<!ATTLIST person title CDATA>
  – Allowed modifiers: #REQUIRED, #FIXED, #IMPLIED, “…”
  – Enumerating allowed values: (dr|prof)`

• Definition of entities:
  `<!ENTITY sw “Semantic Web”>
  – May be used as shortcuts in the XML document: &sw;`
XML Schema

• W3C-Standard (since 2004)
• XML schemas are XML files themselves

• More flexible than DTDs:
  – Minimum and maximum number of elements
  – Combinations of elements (either or, combinations without fixed order, …)
  – Data types (Numbers, dates, …), own types may be defined
  – Support for namespaces
  – Possibility to create modular schemas
XML Schema

```xml
<xs:schema elementFormDefault="qualified"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="physician">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="name" type="xs:string">
          <xs:element name="address">
            <xs:complexType>
              <xs:sequence>
                <xs:element name="street" type="xs:string">
                  ...
                </xs:sequence>
              </xs:complexType>
            </xs:element>
          </xs:element>
        </xs:element>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>

<physician xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="physician.xsd">
  <name>Dr. Mark Smith</name>
  <address>
    <street>Main St.</street>
    <number>14</number>
    <city>Smalltown</city>
  </address>
  <telephone>
    <number>+44 123 456789</number>
  </telephone>
  <hours>
    <monday>9-11 am</monday>
    <tuesday>9-11 am</tuesday>
    ...
  </hours>
</physician>
```
XML Schema – Modular Schemas

```
<xs:import
namespace="http://www.address.com/
schemaLocation="address.xsd"/>
<xs:element name="physician">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="name" type="xs:string">
        <xs:element ref="addr:address" />
      </xs:element>
      ... 
    </xs:sequence>
  </xs:complexType>
</xs:element>
</xs:schema>
```
Example: Modular Schemas in XHTML

```html
<html xmlns="http://www.w3.org/1999/xhtml"
     xmlns:svg="http://www.w3.org/2000/svg">
  <body>
    <h1>SVG embedded inline in XHTML:</h1>
    <svg:svg width="300px" height="200px">
      <svg:circle cx="150" cy="100" r="50"
                 fill="#ff0000"/>
    </svg:svg>
  </body>
</html>
```

https://developer.mozilla.org/En/SVG:Namespaces_Crash_Course
So, what does a DTD/Schema Define?

• Syntax – σύνταξις ("together" + "order")
  – Which elements are there?
  – How are they arranged?
  – Which combinations are allowed?

• ...as opposed to: Semantics - σημαίνειν ("denote")
  – How to interpret the contents of an element?
  – What is their relation?
Syntax und Semantics: The Linguists' View

• Syntax: how are correct sentences formed?
  „This sentence no verb.“
  „The dreaming lamp give gives a freshly cut juices juice to the tire tired sink.“

• Semantics: what does a word and sentence mean?

• Notes
  – syntactic correctness does not guarantee semantic interpretability
  – semantic interpretability does not require syntactic correctness
    (for humans)
Syntax and Semantics: The Linguists' View

Definition of knowledge noun from the Oxford Advanced Learner's Dictionary

knowledge noun

[uncountable, singular] the information, understanding and skills that you gain through education or experience

• practical/medical/scientific knowledge
• knowledge of/about something He has a wide knowledge of painting and music.
• There is a lack of knowledge about the tax system.
→ See related entries: Teaching and learning

2 [uncountable] the state of knowing about a particular fact or situation

• She sent the letter without my knowledge.
• The film was made with the Prince's full knowledge and approval.
• She was impatient in the knowledge that time was limited.
• I went to sleep secure in the knowledge that I was not alone in the house.
• They could relax safe in the knowledge that they had the funding for the project.
• He denied all knowledge of the affair.

3 knowledge economy/industry/worker working with information rather than producing goods

• the emergence of consultancy as a knowledge industry
• the shift toward a knowledge economy
So, what does a DTD/Schema Define?

Employee catalog of the hospital

<physician>
  <name>Dr. Mark Smith</name>
  <address>
    <street>Main St.</street>
    <number>14</number>
    <city>Smalltown</city>
  </address>
  <telephone>
    <number>+44 123 456789</number>
  </telephone>
  <hours>
    <monday>9-11 am</monday>
    <tuesday>9-11 am</tuesday>
    ...
  </hours>
</physician>

(probably)
the private address

Yellow Pages

<physician>
  <name>Dr. Mark Smith</name>
  <address>
    <street>Main St.</street>
    <number>14</number>
    <city>Smalltown</city>
  </address>
  <telephone>
    <number>+44 123 456789</number>
  </telephone>
  <hours>
    <monday>9-11 am</monday>
    <tuesday>9-11 am</tuesday>
    ...
  </hours>
</physician>

(probably)
the work address
So, what does a DTD/Schema Define?

• XML Schema / DTD defines the **syntax** of an XML document, but no its **semantics**

• Tag names are not interpretable by machines
  
  – i.e., they do not ease the information retrieval process...
  
  – Semantics of the data is hidden – usually hard wired in the application

• The Semantic Web is meant as a remedy to that problem
  
  – **Semantic Web is/can do more than XML!**
A Note on Web Services

• Original vision
  – Describe functions of services as XML
  – e.g., stock market ticker, calculator, travel booking…

• ...so that an intelligent agent can combine them
  – and dynamically create a system for a given purpose

• Standards
  – WSDL, UDDI, SOAP, ...

• Problem
  – The semantics is missing!
Wrap Up

• Problems of the classic web
  – Not usable for machines / intelligent agents

• URIs
  – Unique identifiers for resource
  – URLs are dereferencable on the Web

• Unicode
  – A character set for all languages

• XML
  – XPath
  – XSLT
  – DTD
  – XML Schema
Semantic Web – Architecture

here be dragons...

Semantic Web Technologies (This lecture)

Technical Foundations

Berners-Lee (2009): Semantic Web and Linked Data
Questions?