Semantic Web Technologies
Introduction
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**.

... 

CERN

The CERN is a European research center, located close to Geneva.

... 

Tim Berners-Lee

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

...
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
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

In the eyes of a computer

Print in bold: „hmf298hmhuds“
Print in italics: „mj2i9ji0“
Print normal: “fdsah 02hfadsh0um2m0adsmf0ihm asdfjköfdsa298ndsfmij32mio lk2mjpoimjioidpmsajiomjm“

<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
Searching for Information on the Web

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

Homonym and Polysemeas
Untyped Links

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

```plaintext
what is the distance from reinheim to mannheim

Assuming "mannheim" is a city | Use as an airport instead

<table>
<thead>
<tr>
<th>Input interpretation:</th>
</tr>
</thead>
<tbody>
<tr>
<td>distance from Reinheim, Hesse to Mannheim, Baden-Württemberg</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>44.97 km (kilometers)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unit conversions:</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.94 miles</td>
</tr>
<tr>
<td>44.97 km (kilometers)</td>
</tr>
<tr>
<td>44973 meters</td>
</tr>
<tr>
<td>4.497x10^8 cm (centimeters)</td>
</tr>
<tr>
<td>24.28 nmi (nautical miles)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Direct travel times:</th>
</tr>
</thead>
<tbody>
<tr>
<td>car (55 mph)</td>
</tr>
<tr>
<td>sound</td>
</tr>
<tr>
<td>light in fiber</td>
</tr>
<tr>
<td>light in vacuum</td>
</tr>
</tbody>
</table>
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

WWW

Extract information from the Web

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

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

URI = scheme "::" hier-part [ "?" query ][ "#" 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
  - ...
The World Wide Web was initiated in the 90s by Tim Berners-Lee at CERN.

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

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

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

<table>
<thead>
<tr>
<th>Code Point</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>00-0F</td>
<td>Latin scripts and symbols</td>
</tr>
<tr>
<td>10-1F</td>
<td>Linguistic scripts</td>
</tr>
<tr>
<td>20-2F</td>
<td>Other European scripts</td>
</tr>
<tr>
<td>30-3F</td>
<td>African scripts</td>
</tr>
<tr>
<td>40-4F</td>
<td>Middle Eastern and Southwest Asian scripts</td>
</tr>
<tr>
<td>50-5F</td>
<td>Central Asian scripts</td>
</tr>
<tr>
<td>60-6F</td>
<td>South Asian scripts</td>
</tr>
<tr>
<td>70-7F</td>
<td>Southeast Asian scripts</td>
</tr>
<tr>
<td>80-8F</td>
<td>East Asian scripts</td>
</tr>
<tr>
<td>90-9F</td>
<td>East Asian scripts</td>
</tr>
<tr>
<td>A0-AF</td>
<td>Unified CJK Han</td>
</tr>
<tr>
<td>B0-BF</td>
<td>American scripts</td>
</tr>
<tr>
<td>C0-CF</td>
<td>Symbols</td>
</tr>
<tr>
<td>D0-DF</td>
<td>Diacritics</td>
</tr>
<tr>
<td>E0-EF</td>
<td>UTF-16 surrogates and private use</td>
</tr>
<tr>
<td>F0-FF</td>
<td>Miscellaneous characters</td>
</tr>
<tr>
<td></td>
<td>Unallocated code points</td>
</tr>
</tbody>
</table>

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:
    \(<\text{physician}>\ldots<\text{physician}>\)
  – …or empty element tags
    \(<\text{young} />\)

• Attributes:
  \(<\text{physician} \text{ location=“Smalltown“}>\)

• Tags are nested (with exactly one root element):
  \(<\text{physician}>
    \(<\text{address}>\ldots<\text{address}>\)
  \)</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>
    ...
  </hours>
</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></p>

• 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

```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>
```
XSLT: Transformation of XML Documents

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

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

```xml
<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

```
<physician xmlns="http://www.med.com/physician"
            xmlns:addr="http://www.med.com/addr">
  <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)

- Defines 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

```xml
<!ELEMENT address(street,no,line*,zip,city,state?)>
```

- ?, + and * mark optional and possible multiple elements

• Definition of attribute lists

```xml
<!ATTLIST person title CDATA>
```

- Allowed modifiers: #REQUIRED, #FIXED, #IMPLIED, “...“
- Enumerating allowed values: (dr|prof)

• Definition of entities:

```xml
<!ENTITY sw "Semantic Web"> &sw;
```

- 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" type="xs:string">
            ...<stree>"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>
            ...<monday>9-11 am</monday>
          </hours>
        </name>
      </sequence>
    </complexType>
  </xs:element>
</xs:schema>
```
XML Schema – Modular Schemas

```xml
<xs:schema elementFormDefault="qualified"
xmlns:xs="http://www.w3.org/2001/XMLSchema"
xmlns:addr="http://www.address.com/">

  <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>
```

```xml
<xs:schema elementFormDefault="qualified"
xmlns:xs="http://www.w3.org/2001/XMLSchema">

  <xs:element name="address">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="street"
          type="xs:string">
          ...
        </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

1 [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

```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>
```

(probably)
the private address

= ?

Yellow Pages

```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>
```

(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?