Introduction to the Semantic Web
Questions

• What is the Semantic Web?
• Why do we want it?
• How will we do it?
• Who will do it?
• When will it be done?
“XML is Lisp's bastard nephew, with uglier syntax and no semantics. Yet XML is poised to enable the creation of a Web of data that dwarfs anything since the Library at Alexandria.”

“The web has made people smarter. We need to understand how to use it to make machines smarter, too.”

-- Michael I. Jordan, paraphrased from a talk at AAAI, July 2002 by Michael Jordan (UC Berkeley)
“The Semantic Web will globalize knowledge representation, just as the WWW globalized hypertext”

-- Tim Berners-Lee
Web is our greatest knowledge source
But it has limitations
Designed for people, not machines
Designed for people, not machines

- Content is mostly text, spoken language, images and videos
- Easy for people to understand
- But hard for machines

Machines need access to this knowledge too
Access via information retrieval

Vannevar Bush envisioned a hypertext/IR system in 1945
Access is primarily via information retrieval
• Key-word queries → ranked document list
• We still need to read the documents or watch the videos
• We often want an answer to a question: where is the Census Big Data Day event
And so do our machines and apps

Vannevar Bush envisioned a hypertext/IR system in 1945
We need to add knowledge graphs
We need to add knowledge graphs

• High quality semi-structured information about entities and relations
• Represented and accessed via Web standards
• Easily integrated, fused and reasoned with
State of the Art?

Google is a good example, but Microsoft, IBM, Apple and Facebook all have similar capabilities

- 2010 Google acquired MediaWeb and its **Freebase** KB
- 2014: Freebase: 1.2B facts about 43M entities
- 2015+: Google knowledge graph, updated by text IE

**DBpedia** open source RDF KB is another

- 800M facts about 4.6M subjects from English **Wikipedia**, data also available in 21 other languages
- Helps integrate 90B facts from 1000 RDF datasets in the linked data cloud
Ask: When was Tom Sawyer written?

The Adventures of Tom Sawyer

1876
Apple Pie by Grandma Ople

"This was my grandmother's apple pie recipe. I have never seen another one quite like it. It will always be my favorite and has won me several first place prizes in local competitions. I hope it becomes one of your favorites as well!"

Ingredients

1 recipe pastry for a 9 inch double crust pie
1/2 cup white sugar
1/2 cup unsalted butter
Domino Pure Cane Granulated Sugar

On Sale

What's on sale near you.
Almost all commercial recipe sites embed **semantic data** about their recipes in an RDF-compatible form using terms from the **schema.org** ontology.

Search engines read and use this data to better understand the semantics of the page content.
Conversational Bots

Voice-driven conversational systems like Amazon Echo and Google Home use knowledge graphs to help understand our requests.
Where does the knowledge come from?

- Initial knowledge graphs like DBpedia and Freebase started with data from Wikipedia and encoded it in custom ontologies.
- Semantic Web technologies are an open source way to encode the knowledge.
- They are and will continue to evolve.
- A current focus is on extracting information from text of source documents, e.g., journal articles, Newswire, social media, etc.
Why Web Based?

• The web is like a universal acid, eating through and consuming everything it touches.
  - Web principles and technologies are equally good for wireless/pervasive computing

• The semantic web is our first serious attempt to provide semantics for XML sublanguages

• It will provide mechanisms for people and machines (agents, programs, web services) to come together.
  - In all kinds of networked environments: wired, wireless, ad hoc, wearable, etc.
Who invented the Web?
Who invented the Web?
Origins

Tim Berners-Lee’s original 1989 WWW proposal described a web of relationships among named objects unifying many info. management tasks.

Capsule history

- Guha’s MCF (~94)
- XML+MCF=>RDF (~96)
- RDF+OO=>RDFS (~99)
- RDFS+KR=>DAML+OIL (00)
- W3C’s SW activity (01)
- W3C’s OWL (03)

http://www.w3.org/History/1989/proposal.html
W3C’s Semantic Web Goals

Focus on machine consumption:

"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."

TBL’s semantic web vision, circa 2000
Semantic web stack 2006

User Interface & applications

Trust

Proof

Unifying Logic

Query: SPARQL

ontology: OWL

Rules: RIF

RDF-S

Data interchange: RDF

XML

URI

Unicode

Crypto
Why is this hard?

after Frank van Harmelen and Jim Hendler
What a web page looks like to a machine...

after Frank van Harmelen and Jim Hendler
OK, so HTML is not helpful

Maybe we can tell the machine what the different parts of the text represent?

title

speaker

time

location

abstract

biosketch

host
XML to the rescue?

XML fans propose creating a XML tag set to use for each application.

For talks, we can choose <title>, <speaker>, etc.

after Frank van Harmelen and Jim Hendler
XML ≠ machine accessible meaning

But, to your machine, the tags still look like this...

The tag names carry no meaning.

XML DTDs and Schemas have little or no semantics.

after Frank van Harmelen and Jim Hendler
XML Schema helps

XML Schemas provide a simple mechanism to define shared vocabularies.

after Frank van Harmelen and Jim Hendler
But there are many schemas after Frank van Harmelen and Jim Hendler
There’s no way to relate schema XML Schema file 1

XML Schema file 42

Either manually or automatically. XML Schema is weak on semantics.
We need a way to define ontologies in XML
So we can relate them
So machines can understand (to some degree) their meaning
Semantic Web

Use Semantic Web Technology
to publish shared data &
knowledge

Semantic web technologies
allow machines to share
data and knowledge using
common web language and
protocols.

~ 1997

Semantic Web beginning
Semantic Web => Linked Open Data

Use Semantic Web Technology
to publish shared data &
knowledge

Data is inter-linked to support integration and fusion of knowledge

LOD beginning
The node in the center is DBpedia

About: University of Maryland, Baltimore County

An Entity of Type: National Space Grant College and Fellowship Program, from Named Graph: http://dbpedia.org, within Data Space: dbpedia.org

The University of Maryland, Baltimore County (often referred to as UMBC) is an American public research university, located in Baltimore County, Maryland, United States, mostly in the community of Catonsville, approximately 10 minutes (8.3 miles) from downtown Baltimore City and 30 minutes (33.5 miles) from Washington, D.C.

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbo:abstract</td>
<td>The University of Maryland, Baltimore County (often referred to as UMBC) is an American public research university, located in Baltimore County, Maryland, United States, mostly in the community of Catonsville,</td>
</tr>
</tbody>
</table>

http://dbpedia.org/page/University_of_Maryland,_Baltimore_County
Semantic Web => Linked Open Data

Use Semantic Web Technology to publish shared data & knowledge.

Data is inter-linked to support integration and fusion of knowledge.

LOD growing
Semantic Web => Linked Open Data

Use Semantic Web Technology to publish shared data & knowledge

Data is interlinked to support integration and fusion of knowledge

... and growing
Linked Open Data

Use Semantic Web Technology to publish shared data & knowledge

LOD is the new Cyc: a common source of background knowledge

Data is inter-linked to support integration and fusion of knowledge

2010

...growing faster
Linked Open Data

Use Semantic Web Technology to publish shared data & knowledge

LOD is the new Cyc: a common source of background knowledge

Data is interlinked to support integration and fusion of knowledge

2011: 31B facts in 295 datasets interlinked by 504M assertions on ckan.net
Traditionally, all languages are divided into three parts:

1. **Syntax**: legal forms that make up the sentences in a language

2. **Semantics**: mapping of sentences to meaning (perhaps truth theoretic)

3. **Pragmatics**: everything else (how to do things with language, knowledge of world, etc.)
1: Syntax

• Use URIs* to denote classes, properties, objects, relations
  - http://live.dbpedia.org/resource/Alan_Turing
  - http://schema.org/Person
  - http://www.w3.org/1999/02/22-rdf-syntax-ns#type

• Use strings for literals

• Use triples to make statements
  - dbpedia:Alan_Turing rdfs:type schema:Person.
  - “Alan Turing is a Person”

*URI = Uniform Resource Identifier
2: Semantics

- Semantics maps URIs to the things they denote in “the world”
- Some of this in your mind or in how you write your program
- The meaning of some URIs allow automatic inference
  - **Parent** relation is **inverse** of the **child** relation
  - `schema:parent owl:inverse schema:children`
3: Pragmatics

• Semantics is more than just about truth (statements that assert things)

• We also have to account for commands, requests, questions, context, etc.
  - Some of this is handled by Web protocols (GET, POST)
  - Some by special SQ protocols (e.g., SPARLQ for queries and updates)
  - Some by having reference KBs of the world (e.g., DBpedia) to help identify common entities
Where are we

• The W3C version of the open semantic web has been growing steadily
• The languages and standards are being used in government and industry
  - BBC uses RDF to make up much of its content online
  - Google and Facebook detect AND MAKE USE OF (some) RDF embedded in html pages
  - Google, Yahoo, Microsoft and Yandex formed [Schema.org](http://schema.org) to develop useful vocabularies
  - Data.gov has many datasets in RDF
### Wikipedia data in RDF

**dbpedia:Alan_Turing**
- dbpedia:King's_College_Cambridge
- dbpedia:Princeton_University
- dbpedia:Royal_Society
- dbpedia:Order_of_the_British_Empire
- dbpedia:Fellow_of_the_Royal_Society
- dbpedia:Officer_of_the_Order_of_the_British_Empire
- 1912-06-23 (xsd:date)
- 1912-06-23 (xsd:date)
- Alan Mathison Turing
- dbpedia:Paddington
- dbpedia:Maida_Vale
- 1954-06-07 (xsd:date)
- dbpedia:Wilmslow
- dbpedia:Alonzo_Church
- dbpedia:Robinson_Gandy
- dbpedia:Computer_science
- dbpedia:Mathematics
- dbpedia:Cryptanalysis
- 118802976
- dbpedia:Turing_machine
- dbpedia:Cryptanalysis_of_the_Enigma
- dbpedia:Automatic_Computing_Engine
- dbpedia:Turing_test
**Wikidata**

- **Wikidata** aims to create a free rdf-like KB about the world that can be read/edited by humans & machines
  - Wikimedia project started in April 2012 with external funding
- Wikidata clients use the repository, e.g., to populate Web pages or Wikipedia infoboxes
- Based on ideas from [Semantic MediaWiki](https://www.semantic-mediawiki.org/) and [Freebase](https://developers.google.com/freebase/)
Semantic MediaWiki (SMW) is a free, open-source extension to MediaWiki – the wiki software that powers Wikipedia – that lets you store and query data within the wiki's pages.

Semantic MediaWiki is also a full-fledged framework, in conjunction with many spinoff extensions, that can turn a wiki into a powerful and flexible "collaborative database". All data created within SMW can easily be published via the Semantic Web, allowing other systems to use this data seamlessly.

More about Semantic MediaWiki
- Introduction to SMW
- FAQ
- Talks and publications
- Testimonials

Wiki of the Month - January 2013

Triple A is a site that provides IT architecture information for educational organizations. It supports MBO-scholien (Dutch secondary schools) wanting to innovate educational processes.
Freebase

“An entity graph of people, places and things, built by a community that loves open data”
Google Knowledge Graph

Google’s slogan for the knowledge graph: “things, not strings”
Who wrote Tom Sawyer?
Who wrote Tom Sawyer?

Mark Twain
The Adventures of Tom Sawyer, Author

The Adventures of Tom Sawyer - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/The_Adventures_of_Tom_Sawyer
The Adventures of Tom Sawyer by Mark Twain is an 1876 novel about a young boy growing up along the Mississippi River. The story is set in the fictional town of St. Petersburg, inspired by Hannibal, Missouri, where Twain lived.
List of Tom Sawyer characters - Tom Sawyer Abroad - Hannibal, Missouri - 2000 film
Who wrote Tom Sawyer?

Mark Twain
The Adventures of Tom Sawyer, Author

Samuel Langhorne Clemens, better known by his pen name Mark Twain, was an American author and humorist. He wrote The Adventures of Tom Sawyer and its sequel, Adventures of Huckleberry Finn, the latter often called "The Great American Novel". Wikipedia

Born: November 30, 1835, Florida, MO
Died: April 21, 1910, Redding, CT
Full name: Samuel Langhorne Clemens
Spouse: Olivia Langdon Clemens (m. 1870–1904)
Buried: April 24, 1910, Woodlawn Cemetery, Elmira, NY

Get updates about Mark Twain
Apple Pie by Grandma Ople

Recipe by: MOSHASMAMA

"This was my grandmother's apple pie recipe. I have never seen another one quite like it. It will always be my favorite and has won me several first place prizes in local competitions. I hope it becomes one of your favorites as well!"

Ingredients

- 1 recipe pastry for a 9 inch double crust pie
- 1/2 cup white sugar
- 1/2 cup unsalted butter

On Sale

What's on sale near you.
Apple Pie by Grandma Ople

9K made it | 6969 reviews

Recipe by: MOSHASMAMA

"This was my grandmother’s apple pie recipe. I have never seen another one quite like it. It will always be my favorite and has won me several first place prizes in local competitions. I hope it becomes one of your favorites as well."

Featured in Allrecipes Magazine

Ingredients

- 1 recipe pastry for a 9 inch double crust pie
- 1/2 cup white sugar
- 1/2 cup unsalted butter

On Sale
Almost all commercial recipe sites embed **semantic data** about their recipes in an RDF-compatible form using terms from the **schema.org** ontology.

Search engines read and use this data to better understand the semantics of the page content.
Facebook Open Graph

The Open Graph protocol

Introduction

The Open Graph protocol enables any web page to become a rich object in a social graph. For instance, this is used on Facebook to allow any web page to have the same functionality as any other object on Facebook.

While many different technologies and schemas exist and could be combined together, there isn’t a single technology which provides enough information to richly represent any web page within the social graph. The Open Graph protocol builds on these existing technologies and gives developers one thing to implement. Developer simplicity is a key goal of the Open Graph protocol which has informed many of the technical design decisions.

Basic Metadata

To turn your web pages into graph objects, you need to add basic metadata to your page. We’ve based the initial version of the protocol on RDFa which means that you’ll place additional `<meta>` tags in the `<head>` of your web page. The four required properties for every page are:

- `<og:title>` - The title of your object as it should appear within the graph, e.g., “The Rock”.
- `<og:type>` - The type of your object, e.g., “video.movie”. Depending on the type you specify, other properties may also
Apple’s SIRI

SIRI needs lots of semantic data about entities in the world

SIRI engineers from AI/SW community

speech => text => entities => task

SIRI lets you use your voice to send messages, schedule meetings, place phone calls, and more. Ask Siri to do things just by talking the way you talk. Siri is so easy to use and does so much, you’ll keep finding more and more ways to use it.
IBM’s Watson

IBM used Semantic Web technology and data in Watson, see http://bit.ly/X44alE
A collection of useful ontologies

Embed in HTML using RDFa to make machine understandable statements

---

**Thing > Person**

A person (alive, dead, undead, or fictional).

<table>
<thead>
<tr>
<th>Property</th>
<th>Expected Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Properties from <strong>Thing</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>additionalType</td>
<td>URL</td>
<td>An additional type for the item, typically used for adding more specific types from external vocabularies in microdata syntax. This is a relationship between something and a class that the thing is in. In RDFa syntax, it is better to use the native RDFa syntax – the 'typeof' attribute – for multiple types. Schema.org tools may have only weaker understanding of extra types, in particular those defined externally.</td>
</tr>
<tr>
<td>alternateName</td>
<td>Text</td>
<td>An alias for the item.</td>
</tr>
<tr>
<td>description</td>
<td>Text</td>
<td>A short description of the item.</td>
</tr>
<tr>
<td>image</td>
<td>URL</td>
<td>URL of an image of the item.</td>
</tr>
<tr>
<td>name</td>
<td>Text</td>
<td>The name of the item.</td>
</tr>
<tr>
<td>sameAs</td>
<td>URL</td>
<td>URL of a reference Web page that unambiguously indicates the item's identity. E.g. the URL of the item's Wikipedia page, Freebase page, or official website.</td>
</tr>
<tr>
<td>url</td>
<td>URL</td>
<td>URL of the item.</td>
</tr>
<tr>
<td>Properties from <strong>Person</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>additionalName</td>
<td>Text</td>
<td>An additional name for a Person, can be used for a middle name.</td>
</tr>
<tr>
<td>address</td>
<td>PostalAddress</td>
<td>Physical address of the item.</td>
</tr>
<tr>
<td>affiliation</td>
<td>Organization</td>
<td>An organization that this person is affiliated with. For example, a school/university, a club, or a team.</td>
</tr>
<tr>
<td>alumniOf</td>
<td>EducationalOrganization</td>
<td>An educational organization that the person is an alumni of.</td>
</tr>
</tbody>
</table>
Summary

• The Web has made us smarter by sharing information and knowledge as text, audio and images.
• Machines should also be able to use the Web to publish & retrieve information & knowledge.
• Human forms of knowledge are hard for machines to understand and generate.
• The Semantic Web is a collection of languages, ontologies, software tools, services and KBs that are designed to support machines.