The Semantic Web

(some of) what you need to know

Andrea Kosavic
Walter Giesbrecht
OLA Superconference 2009.01.30
Today's Web

- most of today’s Web is suitable for human consumption, (i.e., no artificial assistance required)
- typical uses (involving seeking and making use of information) are not well supported by existing software tools (e.g., search engines)
Problems with search engines

- high recall, low precision
- low or no recall
- results are highly sensitive to vocabulary
- results are single Web pages
Problem

- main obstacle to providing better support to Web users is that, at present, the meaning of Web content is not machine-accessible.

- can a machine distinguish between the differences in meaning of
  - I am a librarian.
  - I am a librarian, you might think. Well, ..
More problems

- most information is currently available in a weakly structured form

- current technology suffers from limitations:
  - searching information
  - extracting information
  - maintaining information
  - uncovering information
  - viewing information
The Semantic Web

- propagated by the World Wide Web Consortium (W3C)
- driving force is Tim Berners-Lee

He expects from this initiative the realization of his original vision of the Web, a vision where the meaning of information played a far more important role than it does in today’s Web.
The Semantic Web

- early web is a "Web of documents"

- Semantic Web is a "Web of data"

- Semantic web vision:
  - data should be related to one another just as documents (or portions of documents) are today
### Web linking: traditional vs. semantic

<table>
<thead>
<tr>
<th><strong>Traditional Web</strong></th>
<th><strong>Semantic Web</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>hyperlinks define a relationship between current page and target</td>
<td>no notion of current page</td>
</tr>
<tr>
<td>hyperlinks are not explicitly named, their role is deduced by a human reader</td>
<td>relationships between any two resources can exist</td>
</tr>
<tr>
<td></td>
<td>the link ITSELF (between objects) is named</td>
</tr>
<tr>
<td></td>
<td>the naming of the link allows for better interchange of data</td>
</tr>
<tr>
<td></td>
<td>RDF gives a formal definition for links</td>
</tr>
</tbody>
</table>
Aims of the Semantic Web

- much more advanced knowledge management systems:
  - Knowledge will be organized in conceptual spaces
  - Automated tools to check for inconsistencies and extract new knowledge
  - Keyword-based search replaced by query answering: requests will be retrieved, extracted, and presented in a human-friendly way
  - Query answering over several documents
  - Restricting access to information will be possible.
What’s it all about?

- The Semantic Web is about two things:
  - common formats for integration and combination of data drawn from diverse sources, where on the original Web mainly concentrated on the interchange of documents
  - language(s) for recording how the data relates to real world objects
The layered Semantic Web
Welcome to the home page of the Agilitas Physiotherapy Centre.
Do you feel pain? Have you had an injury? Let our staff Lisa Davenport, Kelly Townsend (our lovely secretary) and Steve Matthews take care of your body and soul.

Consultation hours
Mon 11am - 7pm
Tue 11am - 7pm
Wed 3pm - 7pm
Thu 11am - 7pm
Fri 11am - 3pm
But note that we do not offer consultation during the weeks of the State Of Origin games.
Content + Metadata

<company>
<treatmentOffered>Physiotherapy</treatmentOffered>
<companyName>Agilitas Physiotherapy Centre</companyName>
<staff>
  <therapist>Lisa Davenport</therapist>
  <therapist>Steve Matthews</therapist>
  <secretary>Kelly Townsend</secretary>
</staff>

</company>
Resource Description Framework (RDF)

- RDF is a standard model for data interchange on the Web.

"RDF is to data as HTML is to documents"

- RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a “triple”)

RDF triples

- based on the idea that the things being described have properties which have values, and that resources can be described by making statements that specify those properties and values
  - SUBJECT
    the part that identifies the thing the statement is about
  - PREDICATE
    the part that identifies the property or characteristic of the subject
  - OBJECT
    the part that identifies the value of that property
RDF Format

- essentially a simple sentence to describe an object, person ... anything

- example: Andrea is a speaker at this conference
  - Andrea = subject
  - speaker = predicate
  - conference = object
RDF requirements

- Triples need to be interpretable by machines

- Two things are needed:
  - A system of machine-processable identifiers for identifying a subject, predicate, or object in a statement without any possibility of confusion with a similar-looking identifier that might be used by someone else on the Web.
  - A machine-processable language for representing these statements and exchanging them between machines.
URIs and XML

- **URI (Uniform Resource Identifier)**
  - more generic form of URL (Uniform Resource Locator)
  - can be used to represent anything
    - network-accessible things
    - things that are not network-accessible, such as human beings, corporations, and bound books in a library.
    - abstract concepts that do not physically exist, such as the concept of a "creator".
Example of triplet

- http://www.example.org/index.html has a creator whose value is John Smith could be represented by an RDF statement having:
  - a subject http://www.example.org/index.html
  - a predicate http://purl.org/dc/elements/1.1/creator
  - and an object http://www.example.org/staffid/85740
RDF graph

objects can be represented by URIs or literals
More on RDF

- by defining a variety of URIs, different people working in RDF on different sets of information can merge knowledge bases by pointing to defined URIs

- URIs can be agreed on by standards organizations or communities or assigned by individuals.
The layered Semantic Web
Syntax vs. Semantics

I love reading Harry Potter novels.

owl:equivalentProperty

I ❤️ reading Harry Potter novels.
Ontologies

- Promote shared understanding
- Define the concepts and relationships used to describe and represent an area of knowledge
- Are knowledge organization systems
A Taxonomy of Knowledge Organization Systems

For more background: http://nkos.slis.kent.edu/KOS_taxonomy.htm
Comparing Components and Functions...

Presenting explicit semantic relationships:
- Eliminating ambiguity
- Controlling synonyms or equivalents

Hierarchical relationships
- Hierarchical and other associate relationships

Presenting properties and attributes of concepts

For more background: http://nkos.slis.kent.edu/KOS_taxonomy.htm
Ontology for a bug and enhancement tracking language

source: http://code.google.com/p/baetle/
Ontologies

- ontologies may include information such as
  - properties (X teaches Y)
  - value restrictions (only faculty members can teach courses)
  - disjointness statements (faculty and general staff are disjoint)
  - specification of logical relationships between objects (every department must include at least ten faculty members)
Ontologies

- Ontologies can be very complex (with thousands of terms) or very simple.

- Standard language to define ontologies:
  - Web Ontology Language (known as OWL)
  - OWL is compatible with and can be expressed in RDF
How ontologies help web searches

- Accuracy is improved
  - Search engine looks for pages that refer to a precise concept in an ontology
  - This is much better than collecting a group of pages where the same keyword may occur
- Browser can help the user by looking at the ontology hierarchy and suggest narrower terms and broader terms
The layered Semantic Web
Other SW technologies

- **logic**
  - one level above ontologies
  - for making inferences, decisions
  - can provide explanations for discovered relationships

- **agents**
  - software to collect and organize information, and present choices for the users to select from
Intelligent agents

- **today**
  - USER
  - present in web browser
  - search engine
  - WWW documents

- **tomorrow?**
  - USER
  - personal agent
  - intelligent infrastructure services
  - WWW documents
Intelligent agents

- collect and organize information, presenting it to users to choose from
- metadata is used to identify and extract info from Web sources
- ontologies assist in Web searches to interpret information and to communicate with other agents
- logic is used to process information and to draw conclusions
Other Technologies

- **SPARQL**
  - query language
  - allows applications to search for specific information within RDF data

- **GRDDL**
  - makes Web pages, microformat tags, XML documents, and data more readable to Semantic Web applications
  - opens more data to new possibilities and creative reuse

For more, see [www.w3.org/2001/sw](http://www.w3.org/2001/sw)
Putting it all together

- Make data available in RDF
- Characterize and classify relationships between objects in detail (OWL, SKOS, RDF schemas)
- Define logical relationships among resources (OWL)
- Query information (SPARQL)
- Extract from and bind to traditional data sources to insure data interchange (GRDDL, RDFa)
Putting it all together

Adapted from:
http://www.w3.org/2008/Talks/0924-Vienna-IH/Slides.pdf
Sample projects

- Linked Data
  - What is Linked Data?
  - DBPedia
  - Geonames
  - LCSH.info

- Standards and vocabularies
  - FOAF
  - OAI-ORE

- Applying semantic web technologies
  - RDF Book Mashup
  - Ingenta Connect
  - SIMILIE
Linked data

- A method of exposing, sharing, and connecting data on the Web via HTTP-based (dereferencable) URIs

- W3Cs Linking Open Data project is working to:
  - publishing various open datasets as RDF on the Web
  - creating links between different data sources
    - datasets consist of over two billion RDF triples (October 2007)
Diagram for the Linked Open Data datasets

Linked data

- Goal is improved information discovery on the Web

- For each resource:
  - Use HTTP URIs so that people can look it up
  - Provide useful information about the resource when its looked up (de-referenced)
  - Include links to other, related URIs
RDF representation of Wikipedia consists of 274 million pieces of information (RDF triples). The knowledge base currently describes more than 2.6 million things in RDF triples. It allows you to query Wikipedia in a more sophisticated way:

- "Give me all cities in New Jersey with more than 10,000 inhabitants"
- “Give me all Italian musicians from the 18th century”
The lcsh.info experiment

- Used the Simple Knowledge Organization System (SKOS)
  - SKOS is an RDF vocabulary for making thesauri, controlled vocabularies, subject headings and folksonomies available on the Web
  - Converted Library of Congress Subject Headings MARCXML to SKOS RDF
- Created URLs for each MARC authority record
  - used Library of Congress Control Number (LCCN) and converted it to a URL (http://lcsh.info/{lccn}#concep)

http://arxiv.org/abs/0805.2855
Why lcsh.info?

- The LCSH/SKOS dataset has multiple places where links could be created to external datasets, including:
  - GeoNames
  - CIA World Fact Book
  - RDF BookMashup
  - DBpedia (http://dbpedia.org)
  - potentially Library of Congress Classification, Name Authority File, and LCCN Permalink Service
A sad ending...

- On December 18th 2008, LC shut down lcsh.info

- Why? Some conjectures in the blogosphere...
  - SKOS data at lcsh.info was a derivative of the LCSH MARC data that is being sold
  - from Dan Chudnov:  
    *It was an important enough success that it was taken down. If it never gained notice, if it weren’t useful, if it didn’t promise something bigger, if it didn’t make sense, if nobody cared, it would still be up. Y’know?*

LC is still considering running a service like lcsh.info at loc.gov, but it’s not yet available
http://lcsh.info/comments1.html
Standards and vocabularies

- DCMI (Dublin Core Metadata Initiative)
- RDA (Resource Description and Access)
- OAI-ORE
- FOAF (Friend of a Friend)
Open Archives Initiative Object Reuse and Exchange (OAI-ORE)

- used to describe aggregations of Web resources
- used for exchange of aggregations of Web resources
- follows Linked Data guidelines; uses URIs
- adopts Semantic Web technologies, including an RDF implementation
Friend of a Friend (FOAF)

- Created mid 2000

- FOAF is a vocabulary that uses RDF and OWL languages

- designed to allow for integration of data across a variety of applications, Web sites and services, and software systems

- FOAF is a linked data system
FOAF Benefits

- enables sharing and use information about people and their activities (eg. photos, calendars, weblogs)
- finds common interests among friends and acquaintances, even if they do not belong to the same social-networking sites such as MySpace or Facebook
- is developing semantic trust networks—white lists of trusted senders—as a way to fight e-mail spam
FOAF supporters
How to explore FOAF for yourself

- Create your own FOAF description using FOAF-a-Matic

- Publish your FOAF description and link it to the FOAFBulletin Board
Introduction

FOAF-a-matic is a simple Javascript application that allows you to create a FOAF ("Friend-of-A-Friend") description of yourself. You can read more about FOAF in Edd Dumbill’s "XML Watch: Finding friends with XML and RDF" article, at the FOAF homepage on RDFWeb, and also the FOAF vocabulary description.

In short, FOAF is a way to describe yourself -- your name, email address, and the people you’re friends with -- using XML and RDF. This allows software to process these descriptions, perhaps as part of an automated search engine, to discover information about your and the communities of which you’re a member. FOAF has the potential to drive many new interesting developments in online communities. Ben Hammersely’s "Click to the Clique" article for the Guardian Unlimited website further explores these ideas.

The FOAF-a-Matic is being provided as a quick and easy way for you to create your own FOAF description. Simply work through the forms on this page and complete whichever details you’d like to add to your description. As a minimum you’ll need to supply your name and email address, and similarly for any friends you might add. It’s worth adding a few friends to your description (but feel free to add as many as you like) because then when FOAF harvesters index your FOAF description, they’ll be able to tie you all together as a network of individuals.

Note: none of the information you enter in this page is used or stored in any way. The processing is entirely client-side, so your privacy is assured.

If you have comments about this application, or further questions about FOAF, why not join the RDFWeb-dev mailing list?

Update: I’m currently writing the FOAF-a-Matic Mark 2 a desktop application for creating and managing your FOAF data.

The Forms

Personal

Some information about you, and how people can contact you.

- Title (Mr, Mrs, Dr, etc)
- First Name
- Last Name (Family/Given)
- Nickname
- Your Email Address
- Homepage
- Your Picture
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/"
    xmlns:admin="http://webservice.net/mvcb/"
>
    <foaf:PersonalProfileDocument rdf:about="">
        <foaf:maker rdf:resource="#me"/>
        <foaf:primaryTopic rdf:resource="#me"/>
        <admin:generatorAgent rdf:resource="http://www.1dodds.com/foaf/foaf-a-matic"/>
        <admin:errorReportsTo rdf:resource="mailto:leigh@1dodds.com"/>
    </foaf:PersonalProfileDocument>

    <foaf:Person rdf:ID="#me">
        <foaf:name>Andrea Kosavic</foaf:name>
        <foaf:givenName>Andrea</foaf:givenName>
        <foaf:familyName>Kosavic</foaf:familyName>
        <foaf:mbox_sha1sum>306685edc0a9b203ae09652b24a37b7f3943e8</foaf:mbox_sha1sum>
        <foaf:homepage rdf:resource="http://relog.wordpress.com"/>
        <foaf:workplaceHomepage rdf:resource="http://www.library.yorku.ca"/>
        <foaf:workInfoHomepage rdf:resource="http://www.library.yorku.ca/ccm/Biblio/staff/bibservicesstafflisting"/>
        <foaf:schoolHomepage rdf:resource="http://www.utoronto.ca"/>
        <foaf:knows>
            <foaf:Person>
                <foaf:name>Walter Giesbrecht</foaf:name>
                <foaf:mbox_sha1sum>fa3be82b646d7e0e2478e9706f3e945496796686ec</foaf:mbox_sha1sum>
            </foaf:Person>
        </foaf:knows>
    </foaf:Person>
</rdf:RDF>
FOAFBulletinBoard

A user maintained directory of links to FOAF descriptions. While one (great) way to get into the FOAF network is to build links between yourself and your friends, adding a link to your description on this page will make it visible to FOAF harvesting applications. (This page can be automatically turned into RDF suitable for processing by FOAF applications.)

Please read carefully when editing the page. Try not to delete anyone!

Just click the link to edit the text of this page, then add your name, and a link to your FOAF description in the bulleted list below. The format is very straightforward, but first click the UserPreferences link found at the upper right corner of the page and create a profile.

See FOAF Bulletin Board Proposal for more background, and info on how to view this page as RDF, if you're interested in harvesting the data. There's at least one other wiki doing this, too. Another option for resolving shal1's is to query SharedDd - which maintains a lookup service for retrieving foaf:urls that are definitive for a specified shal1.


As of June 2003, Leigh's original RDF-extraction tools no longer work (since we switched to using the main Wiki, whose HTML is too bad to be tidied). However Jim Ley has set up a different rdfview of this page.

TODO: Some encodings on this page are broken.

People:

Live links

- A.M. Kochling
- Aaron Swartz
- Abdulaziz alNabhan
- Aditya K
- Adrian Walker
- Adrian Joy
- Akash Mahajan FOAF
- Akash Mahajan Site
- Alan Chieren
- Alastair Garnach
- Alexandre Alajette
- Alexandre Passant
- Ali Seiradain
- Amin Armanian
- Ananda Mathur
- Anders G Helin
- Andrea Kosac
- Andrea Splendiani
Andrea Kosavic

Knows:
Walter Giesbrecht

schoolHomepage:
http://www.yorku.ca/akosavic/www.ucoronto.ca
work info homepage:
http://www.library.yorku.ca/com/Biblio/staff/bilibrary/stafflisting
workplace homepage:
http://www.yorku.ca/akosavic/www.library.yorku.ca
homepage:
http://relog.wordpress.com/

sha1 sum of a personal mailbox URI name:
30668c66d49b0e80465db2452b7f29e43a

family name:
Kosavic

Given name:
Andreas

title:
MSc.

Referenced Vocabularies

http://www.w3.org/1999/02/22-rdf-syntax-ns#
The RDF Vocabulary (RDF)
http://xmlns.com/foaf/0.1/
http://xmlns.net/moby/

mooker

name: Arsenio Hadise Frederick
mbx: shakumathi@yorku.ca/+109/9792132ca334b
homepage: http://www.arsen.de/mhren/
mooker

start
Social Graph API

Build critical mass on your website

With so many websites to join, users must decide where to invest significant time in adding their same connections over and over. For developers, this means it is difficult to build successful web applications that hinge upon a critical mass of users for content and interaction. With the Social Graph API, developers can now utilize public connections their users have already created in other web services. It makes information about public connections between people easily available and useful.

Only public data

The API returns web addresses of public pages and publicly declared connections between them. The API cannot access non-public information, such as private profile pages or websites accessible to a limited group of friends.

How do I start?

1. Learn more about the Social Graph
2. Review the API Documentation
3. Try out the Example Applications

Video Introduction

Based on open standards

We currently index the public Web for XHTML Friend of a Friend (XFN), Friend of a Friend (FOAF) markup and other publicly declared connections. By supporting open Web standards for describing connections between people, websites can add to the social infrastructure of the web.
Applying Semantic Web technologies

- RDF Book Mashup
- Library vendors
- SIMILE Project
Assigns URIs to books, authors, reviews, online bookstores and purchase offers.

When a URI is dereferenced, the mashup queries the Amazon API for information about the book and the Google Base API for purchase offers from different bookstores that sell the book.

The query results are returned as an RDF description to the client.

http://www4.wiwiss.fu-berlin.de/bizer/bookmashup/
Library vendors...

Welcome to IngentaConnect
The home of scholarly research

Search
25,687,328 articles, chapters, reports and more...

Browse
31,964 publications

SIMILE project

- SIMILE project at MIT (simile.mit.edu) Semantic Interoperability of Metadata and Information in unLike Environments

- Seek adds faceted browsing features to Thunderbird for more effective browsing and searching of email
RDFizers

- RDFizers are utilities that generate an RDF representation of your data for you.

- Sample RDFizers from SIMILE:
  - JPEG -> RDF (converts folders of JPEG files)
  - MARC/MODS -> RDF
  - OAI-PMH -> RDF
  - OCW -> RDF (MIT Open CourseWare)
  - EMail -> RDF
  - BibTEX -> RDF

http://simile.mit.edu/wiki/RDFizers
Longwell and DWell

- Longwell is a faceted web browser that displays RDF

- DWell is a Longwell configuration
  - Harvests DSpace metadata into an RDF representation
  - Uses the Fresnel vocabulary to display RDF in Longwell

http://simile.mit.edu/wiki/Longwell
US Presidents

Here is the Exhibit JSON data file.

BIRTH PLACES • DEATH PLACES • TERMS • TABLE • DETAILS • PHOTOS

15 Presidents filtered from 42 originally (Reset All Filters)

sorted by: term; then by... • grouped as sorted

1. James Garfield
   Terms: 24, Republican
   Religious: Church of Christ
   Birth: Sat, Nov 19, 1831, 1:00 am, Moreland Hills, Ohio, USA
   Death: Sun, Sep 16, 1881, 11:00 pm, Elberon (Long Branch), New Jersey
   Link

2. Chester Arthur
   Terms: 24, Republican
   Religious: Episcopal
   Birth: Mon, Oct 5, 1830, 11:00 pm, Fairfield, Vermont, USA
   Death: Fri, Nov 15, 1886, 12:00 am, New York City, New York
   Link

3. Benjamin Harrison
   Terms: 24, Republican
   Religious: Presbyterian
   Birth: Mon, Aug 19, 1833, 1:00 pm, North Bend, Ohio, USA
   Death: Tue, Mar 11, 1901, 11:00 pm, Indianapolis, Indiana
   Link

4. William McKinley
   Terms: 25 and 29, Republican
   Religious: Methodist
   Birth: Fri, Jan 29, 1843, 1:00 am, Niles, Ohio, USA
   Death: Fri, Sep 14, 1901, 11:00 pm, Buffalo, New York
   Link

http://simile.mit.edu/exhibit/examples/presidents/presidents.html
**US Presidents**

Here is the Exhibit JSON data file.

BIRTH PLACES • DEATH PLACES • TERMS • TABLE • DETAILS • PHOTOS

15 Presidents filtered from 42 originally (Reset All Filters)

<table>
<thead>
<tr>
<th>Richard Nixon</th>
<th>Ronald Reagan</th>
<th>George H.W. Bush</th>
</tr>
</thead>
</table>

*Republican*
US Presidents

Here is the Exhibit JSON data file.

15 Presidents filtered from 42 originally (Reset All Filters)
Welcome.

Evergreen is an enterprise-class library automation system that helps library patrons find library materials, and helps libraries manage, catalog, and circulate those materials, no matter how large or complex the libraries. As a community, our development requirements are that Evergreen must be:

- **Stable**, even under extreme load.
- **Robust**, and capable of handling a high volume of transactions and simultaneous users.
- **Flexible**, to accommodate the varied needs of libraries.
- **Secure**, to protect our patrons' privacy and data.
- **User-friendly**, to facilitate patron and staff use of the system.

Evergreen is open source software, freely licensed under the GNU GPL.

Want to learn more? Visit the [Frequently Asked Questions](#) or search the site. Or chat with us live.
Most active domains of application...

- **Small and large business**
  - IBM, Oracle, Adobe, Software AG, Yahoo
  - complete business solutions being sold

- **Health Care and Life Sciences**
  - Gene Ontology (describe gene and gene products in organisms)
  - UniProt (protein sequence and annotation data)
W3C areas of application for Semantic Web technologies:

- Resource discovery and classification
  - Provide better search engine capabilities
  - Eg. ability to limit search to discrete domains

- Cataloguing
  - Describing the content and content relationships available at a particular Web site, page or digital library
W3C areas of application for Semantic Web technologies:

- **data integration**
  - data from various locations and various formats can be integrated into one seamless application
  - eg. Friend Feed (integrates social networking)

- **content rating**
  - describing collections of pages that represent a single logical "document"
W3C areas of application for Semantic Web technologies:

- **intelligent software agents**
  - facilitate knowledge sharing and exchange

- **intellectual property rights**
  - for describing IP rights of Web pages
    - eg. Creative Commons released a vocabulary for specifying license data in RDF. It is among the most widely used RDF vocabularies today
The future?

the semantic web

- is an EXTENSION of the current web
- will NOT REPLACE the current web
- may be adopted locally by major application areas

the current web will not be rebuilt!

- instead, RDF and related ontologies will develop incrementally
some parting thoughts

...
The End

conference materials and links to resources are at

www.yorku.ca/walterg/ola2009/