The Semantic Web

Next Generation information representation, retrieval and processing?

Owen.Conlan@scss.tcd.ie
Athanasios.Staikopoulos@scss.tcd.ie
Agenda

- Syntax vs Semantics
- Data on the Web
- Semantic Web
- Semantic Web Technologies
- Vocabularies
- Linked Data
Semantics is …

Provide a well defined *meaning*, that computers can process

But for us:

**Semantics** = “a representation of the link between a *term* in a statement to the entity in the world that the term refers to”

- p31, Semantic Web for the Working Ontologist
Semantics vs Syntax …

**Semantics** = a way of encoding **meaning** (link between term and a model of the world).
⇒ Good for building applications

**Syntax** = a way of encoding terms so that they can be distinguished, structured, grouped and related to each other in a grammar. (Symbolic representation)
⇒ Good for building parsers

Note! We need a syntax (or syntaxes) for expressing a machine-readable semantics. (RDF is the candidate syntax for the Semantic Web)
Limitations of current approaches

**Structured** Information (files, databases …)
- interrelationships in structure is “implicit”
- easier to deal with computers, designed primarily for human interpretation
- e.g. patient table: name, surname, age, prescription

**Unstructured** Information (web documents …)
- information retrieval/querying based on “clever pattern matching” and human interpretation
- interrelationships of information based on “context” of document placement
- e.g. retail website
Evolution of World Wide Web
Traditional Web

• **Web of documents** – processed by humans
• Typical uses of the Web are information seeking, publishing, searching for people and products, shopping
• Dynamic pages generated based on information from databases but without original information structure found in databases.
Limitations of the Web Search

- Currently, users search for data on the Web asking questions like “which documents contain these words or phrases”

Limitations
- The Web search results are of low precision.
- Results are highly sensitive to vocabulary.
- Results are single Web pages.
- Most of the publishing contents are not structured to allow logical reasoning and query answering.
Data on the Web

- Web has made Data available
  - Easy publication
  - An infrastructure for retrieving and representing documents
  - An infrastructure for accessing data
- There are more and more data on the Web
  - government data, health related data, general knowledge, company information, flight information, sports, weather, news, restaurants,…
- More and more applications rely on the availability of that data
Data on the Web is not enough…

- **Next step is semantic interoperation**
  - Understanding what the data means
  - Linking in insightful ways
  - Automated support for data integration
  - Develop Smart applications
    - Sharing data ⇒ Sharing meaning

- **Need a proper infrastructure for a real Web of Data**
  - data is available on the Web - accessible via Web technologies and standards
  - data are interlinked over the Web
  - data are integrated over the Web

- **This is where Semantic Web technologies come in**
Interconnected web of data
the need for a knowledge driven approach

Increasingly the boundary between data in enterprise systems, personal devices and on web is becoming blurred.
Semantic Web

“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 co-operation.”

[Berners-Lee et al, 2001]

Semantic Web is a web of data that machines can “understand” too.
Difficulties for the Semantic Web

• Current Web represents information using natural language, graphs, pictures, tables, multimedia, …

• Humans can process and combine these information easily

• But, machines:
  • cannot use partial (or incomplete) information
  • have difficulties combining several sources of information
  • can read but cannot “understand” information
Example: Organising a trip

• Imagine you want to organise a trip using the web
  • You try to find a proper flight
  • You have to find a hotel – cheap, luxury
  • You have to trust the specialised site
  • You may want to know something about the place (photographs, maps, itineraries)

• Usually there is a need to
  • Consult a large number of sites
  • You have to mentally integrate all these information
  • It is a long and tedious process
Semantic Web Technologies

• A collection of standard technologies to realise the Web of Data – make the integration possible
  • Structured Web Documents (XML, XSD)
  • Describe Web Resources (RDF)
  • Web Ontology Languages (OWL)
  • Rule Languages (RIF, RuleML, SWRL)
  • Reasoning on the Semantic Web: reasoning tools (e.g. Jena)
  • Searching - Query language (SPARQL)
  • Storing the Semantic Web: Repositories (e.g. Sesame)
  • Semantic Web Services (OWL-S, WSMO)
  • Intelligent Software Agents
  • Trust and Belief
  • Social Web
  • Applications
Stack Architecture for Semantic Web

- Unicode
- URI
- XML + NS + xsmlschema
- RDF + rdfschema
- Ontology vocabulary
- Logic
- Proof
- Trust
- Rules
- Data
- Self-desc.doc.
Structured Web Documents

- XML is used to **encode** documents (e.g. knowledge bases) - provides the means to serialising structured documents
- It provides user definable and domain specific mark up (tags)
- XML Schema: determines the **syntax** (structure) of the document
- There is no commitment
  - On domain specific vocabulary to be used
  - Ontological modelling primitives (is a kind of)
<xml version='1.0' encoding='ISO-8859-1' standalone='yes' ?>
<doc type="book" isbn="1-56592-796-9" xml:lang="en">
  <title>A Guide to XML</title>
  <author>Norman Walsh</author>
  <chapter>
    <title>What Do XML Documents Look Like?</title>
    <paragraph>If you are [...]</paragraph>
    <ol>
      <item>
        <paragraph>The document begins [...]</paragraph>
      </item>
      <item>
        <paragraph>Empty elements have [...]</paragraph>
        <paragraph>In a very [...]</paragraph>
      </item>
    </ol>
    <section> [...] </section>
    [...] 
  </chapter>
  <chapter> [...] </chapter>
</doc>
Describing Web Recourses

• Resource Description Framework (RDF) is a framework for describing and interchanging metadata (data describing the web resources - anything on the Web)

• **Statements** are expressed as triples: a labelled connection between two resources or

  [subject=>predicate=>object]

• RDF can **integrate** information from multiple resources
  - URIs form the basis of identifying and joining graphs

• RDF graphs can be **serialised** in multiple ways
  - (most commonly XML)

• RDF provides **machine understandable semantics**
  - better precision in resource discovery than full text search
  - interoperability of metadata
Example RDF Triples as Graphs

• unv:Person1234 = http://www.scss.tcd.ie/owen.conlan
• dbpedia:Dublin = http://dbpedia.org/resource/Dublin
Ontologies

• Encoding data as graph covers only parts of the meaning of the data
• More elaborate constructs are needed
• An ontology is a specification of a conceptualization
• It describes the common concepts (vocabulary) and relationships between concepts - represents an area of knowledge (see RDFS and OWL)
• There should be a compromise between
  • rich semantics for meaningful applications
  • feasibility, implementability
Example
Logic and Inference

- Logic is the study of systems of reasoning - drawing conclusions
  - First-order logic: the logic of individual things
  - Second-order logic: the logic of types and relationships – can be complex and computational intensive

- Logic plays many different roles for the Semantic Web
  - Applying and evaluating rules
  - Inferring facts that haven’t been explicitly stated
  - Explaining why a particular conclusion has been reached (trace)
  - Detecting contradictory statements and claims
  - Key role in the statement of queries
Logic and Inference

• The rules take the form
  IF <logical conditions are met>
  THEN <perform specified actions>
  (this kind of rules used by so-called expert systems)

• Evaluating the truth of the logical conditions involves logic. Rules are often chained together

• A processor can work backward from one condition to work out what had to happen to get there

• What is needed
  • A web compatible language for expressing rules (standard)
  • Be able to specify among rules relationships and constraints
  • Tools/Engines to handle the rules and reason about the data
Rules

• Some conditions may be complicated in ontologies (ie, OWL) - For example combine predicates and rules
• Enhance expressivity
• Easier to read and write rules with a rule language
  \[ \text{Person}(?p) \land \text{hasSibling}(?p, ?s) \land \text{Man}(?s) \rightarrow \text{hasBrother}(?p, ?s) \]
• RuleML – is a family of XML rule-languages for publishing and sharing rules on Web
  • Focus on interoperation between standards
• SWRL (Semantic Web Rule Language) is a rule language for the Semantic Web, combines ontologies and rules
  • Rules are expressed in terms of OWL concepts
  • SWRL rules have the form of an implication between an antecedent (body) and consequent (head)
SPARQL is a query language of the Semantic Web - get information from RDF graphs
- Is a declarative query language (similar to SQL)
- Based on pattern matching against the RDF graph
- Extract information - e.g. triples, URIs, plain and typed literals
- Construct new RDF graphs from the queried graphs
- Different types of graph patterns are supported
  - Basic, Group, Optional, Alternative, Named, Constraints
- Matching a triple pattern (subject, predicate, object) to a graph: bindings between variables and RDF Terms

Storing the Semantic Web: Repositories

• Semantic Web creates a wealth of data. Keeping them in one big text (e.g. Turtle or RDF/XML) is not the most efficient option (e.g. data are not indexed)

• Need for semantic repositories to support the efficient manipulation of Semantic Web data

• An RDF store holds place for storing the RDF datamodel as a sequence of: s (subject), p (predicate), o (object)
  • tools that combine the characteristics of database management systems (RDBS) – efficient storage, querying, management
  • inference engines – allow reason about the data

• Example of a Semantic Repository Engine
  • Sesame: most popular semantic repository that supports RDF(S) and all the major syntaxes and query languages related to it
Semantic Web Services

- A **web service** is a **network accessible interface** that exposes the application functionality
  - Once it is deployed, other applications (and other Web services) can discover and invoke
  - It is implemented by using standard technologies (WSDL, REST)
  - Clients do not need to know how it is implemented
- Web Services connect computers and devices with each other using the Internet to exchange data and combine data in new ways.
- However, all these service descriptions are based on semi-formal natural language descriptions.
- There is a need to make Web Services an automated technology by adding semantic web technology
Semantic Web Services

- **Semantic Web Services** are Web Services with a formal description (semantics) that can enable a better description, discovery, selection, invocation, composition, monitoring, and interoperability.
  - see Semantic Markup for Web Services (OWL-S) http://www.w3.org/Submission/OWL-S/

- **Processes** are created from the composition of Web Services and/or other components and allow to carry out more complex tasks such as e-commerce business activities
Intelligent Software Agents

• An **agent** is a computer system that is situated in some **environment** and that is capable of some **autonomous** action in order to meet its design **objective**

> “An autonomous agent perceives its environment via sensors and acts upon that environment through its actuators”

• There are different classifications such as
  
  o Reactive agents
  o Belief-desire-intentions agents
  o Goal based agents
  o Learning agents
Intelligent Software Agents

- Agents are capable of *interacting* with other agents by *exchanging* data and they can engage with other agents in some social activities such as coordination, cooperation, negotiation etc.

- **Semantics** are needed to
  - Support Agent communication, negotiation
  - Seek information
  - Interpret Concepts/Vocabulary
  - Represent Logic
Proof & Trust

- Trust is largely confined to Identity
- Identity is usually established via digital certificates and authentication
  - A digital certificate is a digital form of identification. It provides information about the identity of an entity.

- **Proof**: that an answer found in the semantic Web is correct
  - How – derived from logic
  - By whom – chain of providers
Semantic Technologies for Unstructured Data

They are related to Natural-language processing, Information Retrieval and Extraction

- **Entity extraction** – (people, places, events, dates)
- **Cluster analysis** – group related information where relationships are unknown
- **Classification** – map to specific categories
- **Dependency identification** – rule generation
- **Coreference resolution** – two or more expressions in a text refer to the same entity
- **Automatic Summarization** – identify key concepts and key sentences

- Example Tools: GATE (General Architecture for Text Engineering)
Social Web

• Provides new structures and abstractions on top of the traditional Web allowing people to connect and communicate via the Internet

• They are characterised
  • Community – they allow people (contributors) to collaborate and share information easily. (Wikipedia, blogs)
  • Mashups - Integrating Web Resources in new ways. (housing + Google maps).

• Social Networking Sites (SNS)
  • For example Facebook, LinkedIn, Twitter, YouTube

• Allow us
  • Explore trending topics, discover what people are saying, analyse fans/followers, examine friendships, cluster colleagues, analyse who is talking to whom, how often, common interests, …
Examples of Semantic Applications

- Semantic Web search Engines
- eBusiness, eCommerce
- eGovernment
- Health-care and Life Sciences
- eLearning
- eCulture
- Media Management (e.g. BBC)
- Supply Chain Management
- ...
Lucy issues instructions

Her agent follows hyperlinks in the request to ontologies where key terms are defined. Links to ontologies are used at every step.

After getting treatment info from the doctor's computer and schedule info from Lucy's and Pete's computers, the agent goes to a provider finder service.

Lucy's agent and the finder service negotiate using ontologies and agree on payment for its service.

The finder service sends out its own agents to look at semantics-enhanced insurance company lists and provider sites.

The agent sends the appointment plan to Lucy and Pete at Pete's home (per Lucy's request) for their approval.

Vision for "Man in street"
Vocabularies

• RDFS makes it possible to define *vocabularies*:
  • collection of properties and classes
  • relationships among those and to *terms in other vocabularies*

• Examples include
  • Dublin Core
  • FOAF
  • Organisations
  • Good Relations (ecommerce)
  • RSS (Rich Site Summary)
  • Vcard
The Dublin Core Metadata Initiative

• is an open forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models

Properties in the /terms/namespace

• abstract, accessRights, accrualMethod, accrualPeriodicity, accrualPolicy, alternative, audience, available, bibliographicCitation, conformsTo, contributor, coverage, created, creator, date, dateAccepted, dateCopyrighted, dateSubmitted, description, educationLevel, extent, format, hasFormat, hasPart, hasVersion, identifier, instructionalMethod, isFormatOf, isPartOf, isReferencedBy, isReplacedBy, isRequiredBy, issued, isVersionOf, language, license, mediator, medium, modified, provenance, publisher, references, relation, replaces, requires, rights, rightsHolder, source, spatial, subject, tableOfContents, temporal, title, type, valid
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/">

  <rdf:Description rdf:about="http://www.scss.tcd.ie/Owen.Conlan/">
    <dc:title>Dr. Owen Conlan’s Home Page</dc:title>
    <dc:creator>Owen Conlan</dc:creator>
    <dc:publisher>SCSS, University of Dublin</dc:publisher>
  </rdf:Description>

</rdf:RDF>
Friend of a Friend (FOAF)

• is a machine-readable ontology describing persons, their activities and their relations to other people

http://xmlns.com/foaf/spec/
Friend of a Friend (FOAF) - Example

```xml
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:foaf="http://xmlns.com/foaf/0.1/">

  <foaf:Person>
    <foaf:name>Peter Parker</foaf:name>
    <foaf:gender>Male</foaf:gender>
    <foaf:title>Mr</foaf:title>
    <foaf:givenname>Peter</foaf:givenname>
    <foaf:family_name>Parker</foaf:family_name>
    <foaf:mbox_sha1sum>cf2f4bd069302febd8d7c26d803f63fa7f20bd82</foaf:mbox_sha1sum>
  </foaf:Person>

</rdf:RDF>
```
Linked Data

- **Linked Data** lies at the heart of what Semantic Web is all about: *large scale* integration of, and reasoning on, data on the Web.

- **Goal**: “expose” datasets on the Web

- Set **links** among the data items from different datasets
1. put things on the Web through URIs and open license (any format)
2. use HTTP, URIs, so that people can look up these names, they are machine readable (not a scan)
3. provide useful information using standards – non proprietary formats (e.g. excel)
4. use open standards to identify things (e.g. RDF)
5. include links to other URIs, so people can discover more things

RDF is an ideal vehicle to realize these principles

“Linked Data” is also a set of principles:

Is your data 5 star ?
Linked Data or Open Data?

- Linked Data is actually linked only when data is rated with “5 star”

- The name “Linked Data” doesn’t make much sense for the lower rated data.

- The “3 star” data is thus interpreted as Open data (one based an open licence and in non--proprietary formats)
How To Link Data?

• Links happen at the instance level
  • cf. Hyperlinks in HTML
• OWL:SameAs -- equivalence
• RDF:SeeAlso -- associative

Bib:aBook OWL:SameAs DBPedia:aBook
DBPedia Berlin - Example

http://dbpedia.org/page/Berlin

- owl:sameAs
  - http://www4.wiwiss.fu-berlin.de/eurostat/resource/regions/Berlin
  - fbase:Berlin
  - http://linkededgeodata.org/triplify/node/240109189#id
  - http://sws.geonames.org/2950159/

- Can use one concept list to query another database:
  - ask Geonames about the concept known in DBPedia as Berlin
Linked Data: The WWW database
Linked Data: The WWW database
Some characteristics of Linked Data and its Applications

• The datasets are essentially read-only
  • they are curated “out of band”: regularly extracted from other databases, changed manually by data owners, etc

• The dominating paradigm is to extract data via SPARQL queries

• Applications use (very) large datasets via (RDF based) integration
Conclusions

• **Semantics** allow a common interpretation/meaning
• **Web Standards** facilitate interoperability
• **Data** on the Web is a major challenge
  • technologies are needed to use them, to interact with them, to integrate them
• **Semantic Web technologies, Linked Data** principles and practices, should play a major role in publishing and using Data on the Web