CO7216 - Semantic Web

- the technology behind Web 3.0

PART 1

Yi Hong
Department of Computer Sciences
yh37@le.ac.uk

UNIVERSITY OF LEICESTER

*with thanks to Monika Solanki

Aims and Objectives of the Course

- Goals of the Semantic Web
- Foundations of the Semantic Web
- Standards for the Semantic web
- Ontological Engineering
- Programming the Semantic Web
- Semantic Web Applications
- Principles of Linked Data

Topics to be covered

- Introduction to Semantic Web
- Ontology Languages
  - Resource Description Framework (RDF) and RDFS
  - Web Ontology Language (OWL)
  - A Query Language for RDF: SPARQL
- Ontology Engineering
- DL/Deductive Reasoning
- Semantic Web Application
  - Linked Data, RDFa, Microformat and Microdata
  - Jena and OWL API

Logistics

- Module Assessment
  - Coursework and class tests (40%)
    - 1 group assignments on Ontology Modeling
    - 1 individual assignment on Ontology modeling
    - 2 Class Test
    - 1 Individual Programming Exercise
  - Midsummer examination – 2 hours (60%)
    - Contains 3 Questions
    - Each of the questions is worth 50 marks (only the best two answers will be considered).

External Resources

- W3C website: http://www.w3.org/standards/semanticweb
- Additional material on the Blackboard
- Some books
  - Grigori Antoniou, Frank van Harmelen, ***A Semantic Web Primer***, MIT Press, USA, 2008.
  - Hitzler, Pascal, ***Foundations of Semantic Web technologies***, Chapman & Hall, 2009.
  - ***Semantic Web for the Working Ontologist***, Dean Allemang, Jim Hendler, Morgan Kaufmann, 2nd edition.
  - Free online book: ***Linked Data: Evolving the Web into a Global Data Space***.
  - More can be found at: http://www.w3.org/wiki/SwBooks

Prerequisites

- Good understanding of XML and XML Schema

Introduction to the Semantic Web

The "Current" Web

- **Tim Berners Lee** (amongst others) invented the WWW.
  - Launched in early 1990s, exponential growth in mid-1990s
  - Simplest way to share information.
  - Essential: The idea of an open community
  - HTML is typically the language used to code information.

The “Current” Web

- Three major components:
  - Protocol to request data: ***HTTP***
    - GET /path/to/file/index.html HTTP/1.0
  - Address where data can be found: ***URL***
    - http://www.somehost.com/path/index.html
  - Markup language which displays data in a form “humans” can interpret it: ***HTML*** (often styled with CSS)

```
<html>
 <head>
   <title>Title of page</title>
 </head>
 <body>
   This is my first homepage.
   <b>This text is bold</b>
 </body>
</html>
```

The “Current” Web

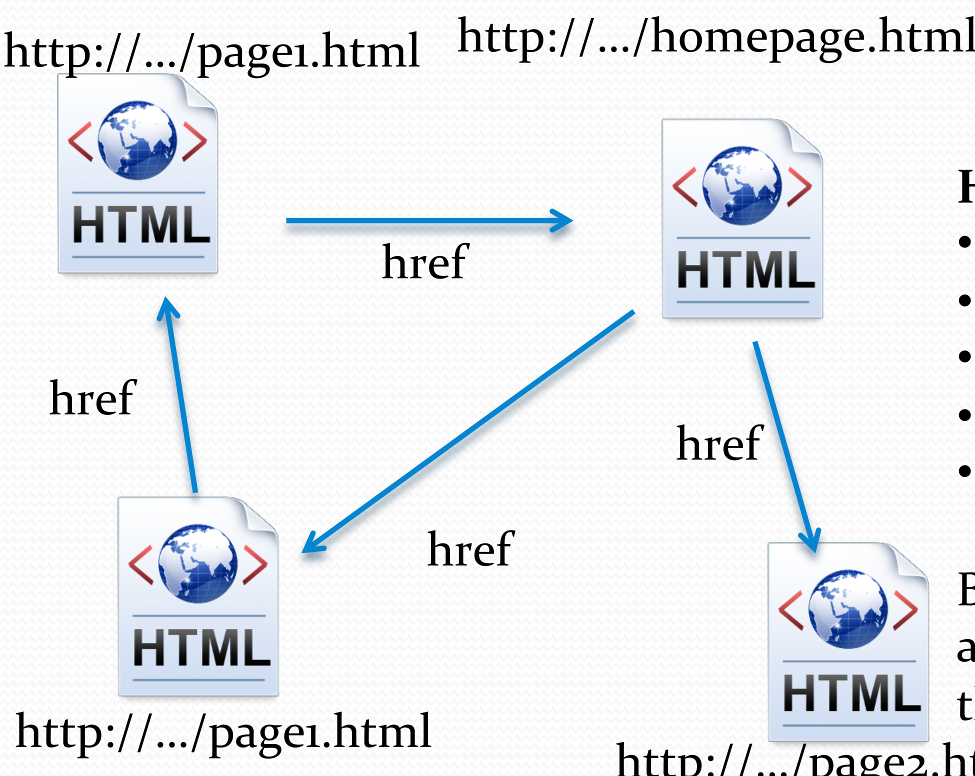
- A syntactic web.
- A library of documents (web pages) connected by links. Most web pages are designed for human consumption using:
  - natural languages-English, Chinese,.... Graphics,
  - multimedia, text, links... renderization (page layout).
- Computers: **Presentation**-easy.
- Human: **Interpretation**-hard.
- Tools are available to
  - author web pages (HTML editors).
  - retrieve web pages (search engines).
  - interpreting web pages (??????)

The Linked "Syntactic" Web

- The hyperlinks defines a relationship between the current page and the target
- link text describes the destination content

Page 1

```
<html>
 <head>
   <title>page1</title>
 </head>
 <body>
 <a href="http://../homepage.html">
visit homepage</a>
 </body>
</html>
```



Humans

- can read and understand link text
- can process web based information easily
- can deduce facts from partial information
- can create mental associations
- are used to various sensory information

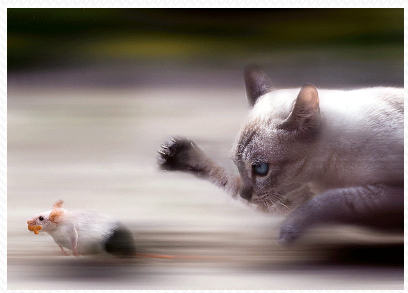
But we can do this only if there is a small amount of information that is available to them

Searching the Syntactic Web: Hard Work!

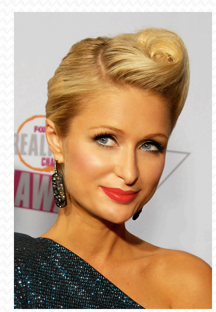
- What you want may not be “exactly” what you get.

Keyword-based searching: ambiguous keyword..

mouse, cat



Paris Hilton



Searching the Syntactic Web: Hard Work!

- Success of the Web -> Search Engines, ***but*** ….

- Results of Web search have to be interpreted by humans.
- Information overload: The complexity and volume of search results is too large to be handled by humans.
- High recall, low precision: e.g. 20 relevant pages for every 100 irrelevant ones. Too much is as bad as too little.
- Results are sensitive to vocabulary: a big problem with keyword search.
- Most searches target only single web pages.
  Results of Web searches are not readily accessible by other software tools;

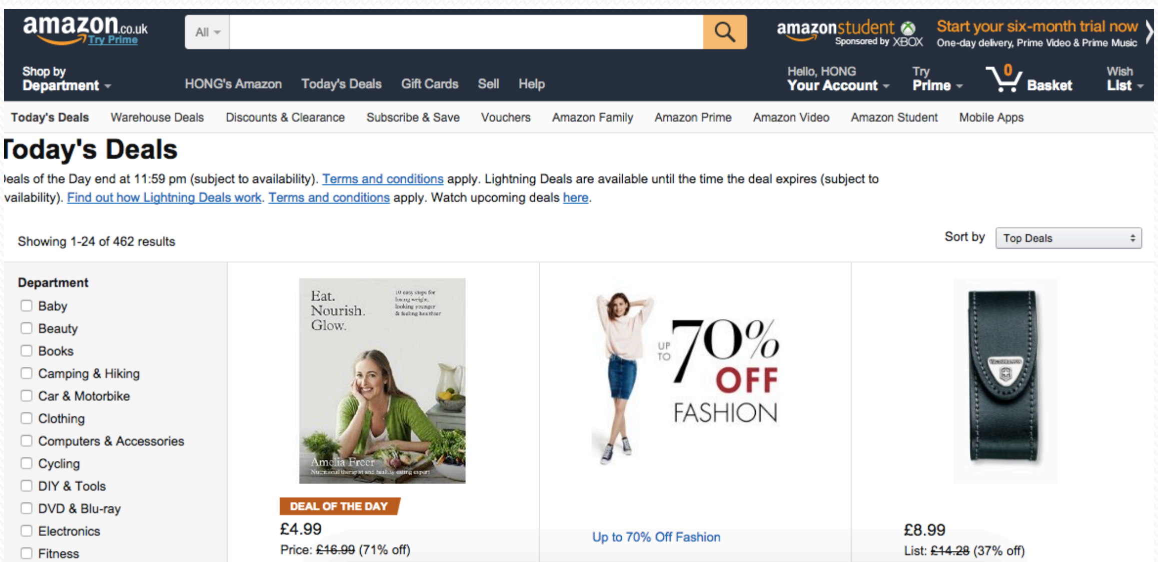
How feasible are some other tasks?

- Making complex queries:
  - “Find a restaurant that serves only vegetarian meals from Greece, is close to a central London tube station and where main meals cost less than £10”
  - Find a picture with two animals in it, one is chasing the other.
- Extracting data from repositories:
  - Online shopping, job hunting etc.
- Discovering and using Web services: e.g. booking a holiday
  - Flight Booking, Accommodation Booking
    Tourist attraction search etc.

Problems with the “current” web

- Web pages do not contain information about their content.
- The meaning or **semantics** of Web content can be interpreted only by humans

Amazon as seen by humans:



Problems with the "current" web

- Web pages do not contain information about their content.
- The meaning or **semantics** of Web content can be interpreted only by humans

Amazon as seen by computer:

%£$^%^$£&$%^$^&%&^&%&^%))(@*£&(&£(*&££@%£
$^%^$£&$%^$^&%&^&%&^%))(@*£&(&£(*&££@%£$^
%^$£&$%^$^&%&^&%&^%))(@*£&(&£(*&££@%£$^%^$
£&$%^$^&%&^&%&^%))(@*£&(&£(*&££@%£$^%^$£&$
%^$^&%&^&%&^%))(@*£&(&£(*&££@
%£$^%^$£&$%^$^&%&^&%&^%))(@*£&(&£(*&££@%£
$^%^$£&$%^$^&%&^&%&^%))(@*£&(&£(*&££@%£$^
%^$£&$%^$^&%&^&%&^%))(@*£&(&£(*&££@

XML: A possible solution?

- Developed within W3C
- Allows for extensibility
  Accompanied by much related work:
  - XML Schemas /cf DTDs
  - XML Namespaces
  - XLink and XPointer:Better hyperlinking
  - XSLT: Transform XML resources XPath: query XML documents
- Markup documents with “meaningful” tags !!!

Problems with XML

John Smith's homepage is http://www.cs.le.ac.uk/johnsmith

This information could be typically be represented in XML according to different XML schemas:

1

```
<author>
   <uri>http://www.cs.le.ac.uk/johnsmith</uri>
   <name>John Smith</name>
</author>
```

2

```
<homepage href="http://www.cs.le.ac.uk/johnsmith">
  <author>John Smith</author>
</homepage>
```

3

```
<webpage>
  <author>
   <uri href="http://www.cs.le.ac.uk/johnsmith"></uri>
   <details>
      <name>John Smith</name>
   </details>
   </author>
</webpage>
```

4

```
<document href="http://www.cs.le.ac.uk/johnsmith" author="John Smith" />
```

To a person reading these, these representations mean the same thing

Problems with XML

John Smith's homepage is http://www.cs.le.ac.uk/johnsmith

This information could be typically be represented in XML according to different XML schemas:

```
<webpage>
   <author>
    <uri href="http://www.cs.le.ac.uk/johnsmith">
   </uri>
    <details>
        <name>John Smith</name>
    </details>
    </author>
</webpage>
```

The above XML document as seen by an XML parser parser

```
   <x>
      <y>
       <z p="text2"></z>
       <q>
           <r>text1</r>
       </z>
       </y>
   </x>
```

Human can read and understand the meaning of this document because the **element names** are a big hint for a human reader

Even if computer knows the document schema (a big ***if***), the meaning of the document is still not clear.

Problems with XML

- XML only describes syntax (syntactic structure).
- Does not provide semantics (what does `<name>` mean?)
- XML is commonly depicted as “self-documenting” but this depiction ignores critical ambiguities.
- Not flexible : does not allow for extensibility.
- The hierarchical model for representation is limited.
- Expressing overlapping (non-hierarchical) node relationships requires extra effort
- Similar content may be described using different XML Schema/DTDs: e.g. is `<document>` the same as `<webpage>`?

Scenario: Renting a room

- Search for information in data is difficult (XQuery).

Requirements: You want to rent a room in a house or flat in Leicester, close to the university and you can afford up to £400 pm including all bills paid. Sharing only with students.

Estate Agents Web site: Estate Agent uses the following to describe what is available.

```
<Accomodation>
    <Room>
       <Scope>Rent</Scope>
       <Place>Leicester</Place>
       <DistanceFromUni>10 mins</DistanceFromUni>
      <RentLimit>400</RentLimit>
      <Sharing>Students</sharing>
      <BillsPaid>Yes</BillsPaid>
  </Room>
</Accomodation>
```

Scenario: Renting a room

- Search for information in data is difficult.

```
<Accomodation>
    <Room>
       <Scope>Rent</Scope>
       <Place>Leicester</Place>
       <DistanceFromUni>10 mins</DistanceFromUni>
       <RentLimit>400</RentLimit>
       <Sharing>Students</sharing>
       <BillsPaid>Yes</BillsPaid>
  </Room>
</Accomodation>
```

Difficult to answer the question using XQuery, although structured information is provided.
Humans understand the semantics of the tags, machines don't! Valid match for humans - machines don't care !

What's missing? Semantics!!

Web content understood by humans

```
<h1>Accommodation</h1>
<ol>
  <li>Type: Rent</li>
  <li>Leicester</li>
  <li><10 mins from Univ/li>
  <li>Sharing with students</li>
  <li>400pcm</li>
  <li></li>
</ol>
```

Web content understood by software with background "knowledge"

```
<Accomodation>
  <Room>
   <Scope>Rent</Scope>
   <Place>Leicester</Place>
   <DistanceFromUni>10 mins
    </DistanceFromUni>
   <RentLimit>400</RentLimit>
   <Sharing>Students</sharing>
   <BillsPaid>Yes</BillsPaid>
  </Room>
</Accomodation>
```

We still need a mechanism that allows equivalent resources to be identified and **understood** by machines without programming this knowledge into software.

What's missing? Semantics!!

> "The bane of my existence is doing things that I know the computer could do for me.."
>
> -- ***Dan Connolly, The XML Revolution***

- Realising the complete "vision" is too hard for now (probably)
- But we can make a start by adding semantic annotation to web resources

History of the Semantic Web

- 2001: Historical article in the Scientific American – look at your notes..
- Tim Berners Lee - two part vision for the Web
  - to make the web more collaborative.
  - to make the the web understandable and processable by machines
- Data about the data currently on the Web: Metadata!!!
- Metadata captures a part of the meaning of the data.
- Historically data was locked away behind proprietary applications.
- With the Web, XML and Semantic Web: focus has shifted from applications to data.
- Machine processable data has to be smart data

What is Semantic Web

Semantic Web is a logical extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.

- The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners.

Semantic Web vs Current Web

- **Current web**
    - Document-centric
    - Human readers
    - Syntax (Schema)
    - HTML, XML etc.
- **Semantic web**
    - Knowledge representation
    - Machine readable and understandable
    - Semantics (Ontology)
    - RDF, OWL etc

The Semantic Web is an extension of the current Web and not its replacement.

What is Semantic Web

- Semantic Web is
- The “next” generation web
- Machine processable data.
- Solving the Application Integration problem.
- The infrastructure for the next IT revolution.
- A connected web - a smarter web.
- A Distributed web of data

Main Goals

- The vision of the Semantic Web is to extend principles of the Web from documents to data
- Data should be accessed using the general Web architecture
- A common framework that allows data to be shared and reused
- Allow data to be processed automatically by tools
- Revealing possible new relationships among pieces of data.

Semantic Web Applications

- Data integration; describing collections of pages that represent a single logical “document”
- Allows intelligent software agents to facilitate knowledge sharing and exchange
- Annotating content and describing content relationships
- Resource discovery and classification
- Semantic search engine

Semantic Web Uptake

- Semantic Web is NOT just for research community.
  - Semantic Web is increasingly used by governments, small and large business such as Oracle, IBM, Google, Adobe, Software AG, or Yahoo!
  - Some are already selling tools as well as complete business solutions.
    - Oracle Spatial and Graph RDF Semantic Graph
    - Microsoft Bing Satori knowledge base
    - Google Knowledge Graph
    - UK Government project to make available non-personal UK government data as open data (Data.gov.uk)

  Many EU/UK organisations and research projects have already used Semantic Web technologies. For example, British Museum is the first UK arts organisation to publish its collection semantically.

Summary

- Semantic Web is not a replacement of the current Web, it is an evolution of it
- Semantic Web is about:
  - Annotation of data on the Web data linking on the Web
  - Data Integration over the Web
  - Semantic Web aims at automating tasks currently carried out by humans
- Semantic Web is becoming real (maybe not as we originally envisioned it, but it is)

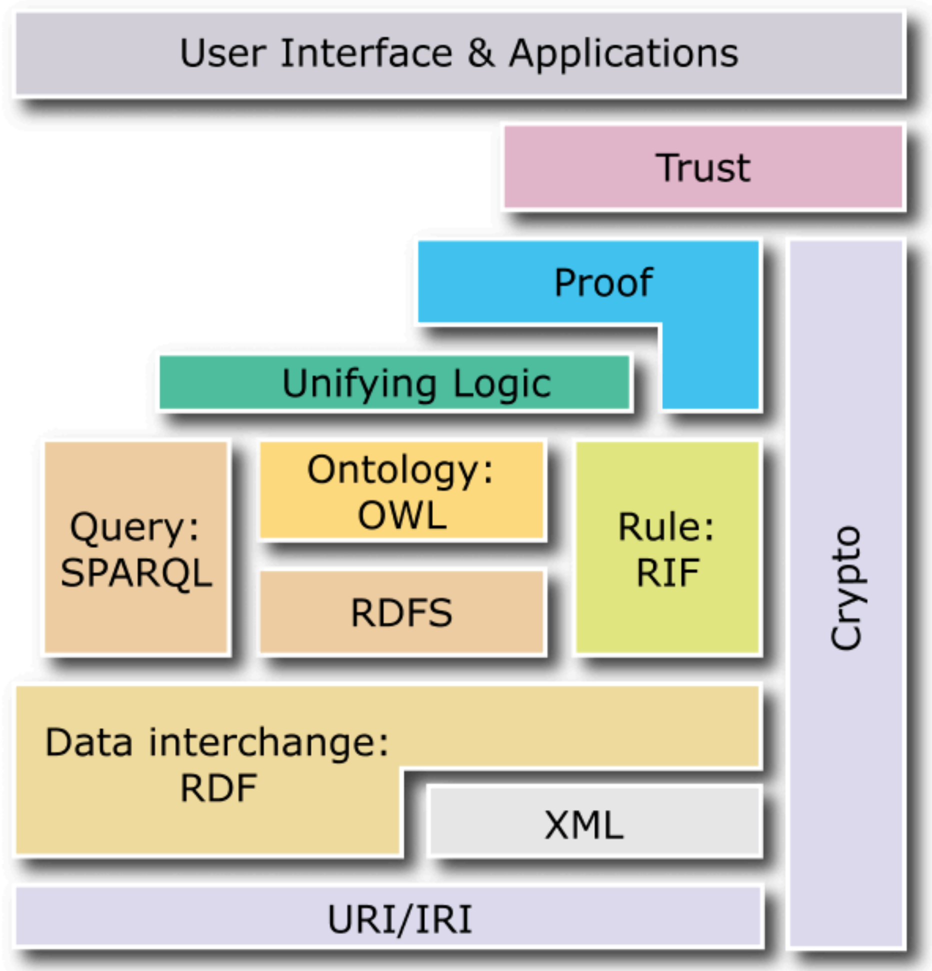
Essential Reading

- Chapter 1 of the recommended books (Lecture Notes, page 5)
- W3C Semantic Web Frequently Asked Questions: http://www.w3.org/RDF/FAQ
- Scientific American article (Blackboard -> Web Links)

Infrastructure for the Semantic Web

Infrastructure for the Semantic Web

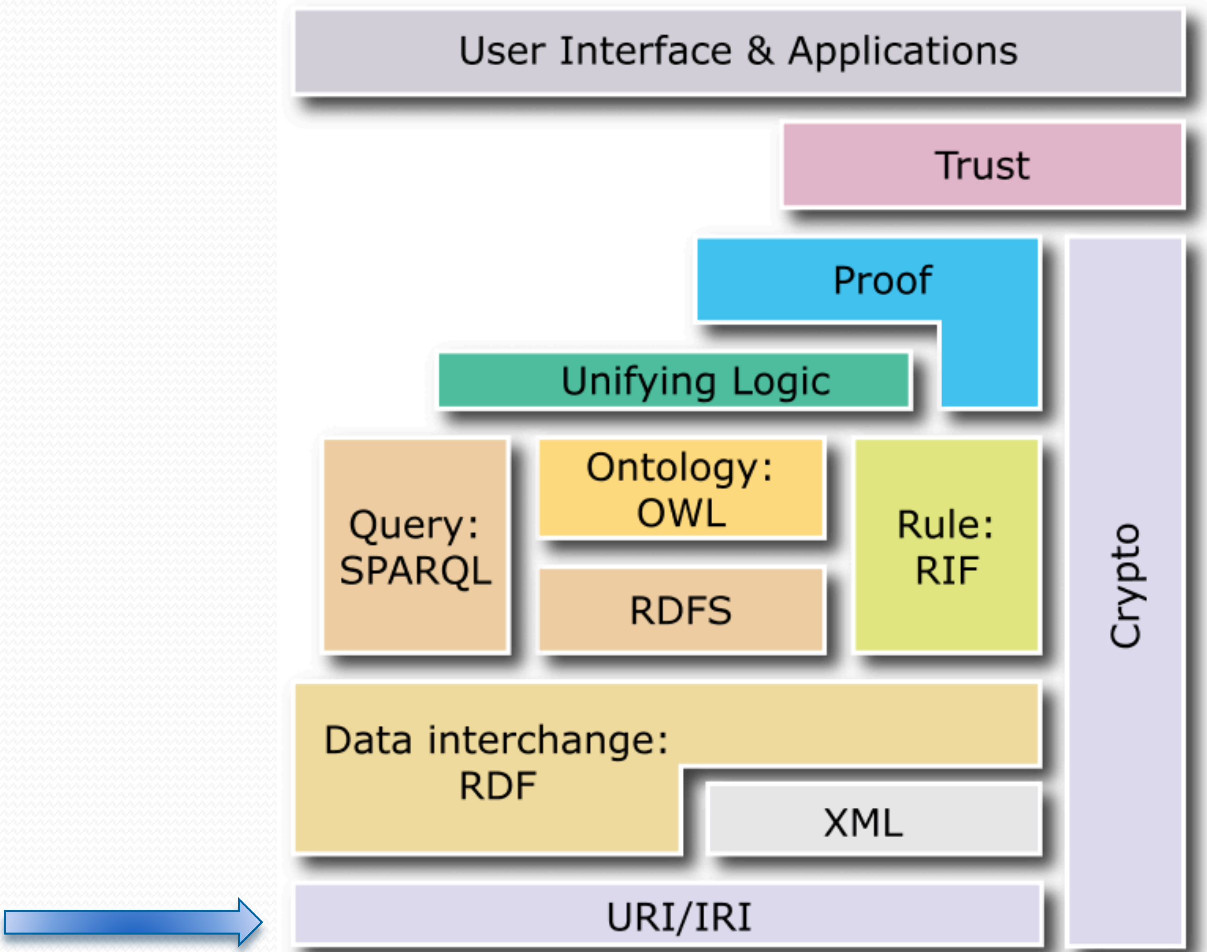
Semantic Web Layer Cake



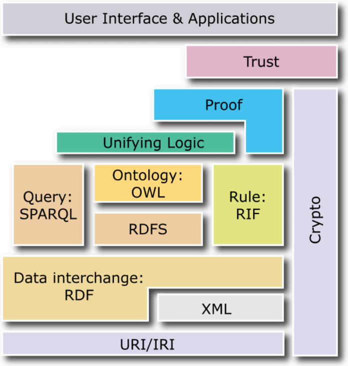
Semantic Web Technologies

- Explicit **Metadata** (about "structuring" the data, It describes characteristics about the resource in a structured way)
- **Ontologies** to standardise concepts and relations between them
- **Logic and Inference**: languages founded in various flavours of logic
- **Software Agents**: make use of all the above to help us in our tasks

Semantic Web Layer Cake



URL vs URI vs IRI



- **URL (Uniform Resource Locator)**
  - A reference (an address) to a resource on the Internet.
- **URI (Uniform Resource Identifier)**
  - A reference to identify a resource.
  - The most common form of URI is the URL.
  - URIs identify and URLs locate so every URL is also a URI, but there are URIs which are not URLs.
- **IRI (Internationalized Resource Identifier)**
  - IRI is an extension to URI, URIs are limited to a subset of the ASCII character set, but IRIs may contain characters from the Universal Character Set (Unicode/ISO 10646)

On the Semantic Web, IRIs/URIs identify not only Web documents, but also real-world objects

Introduction to Ontologies

What is an Ontology?

- History and Origin ..
  - In Classical Philosophy.
    - The term “Ontology” is a compound word, originated in Greek, *onto* (i.e. being or the nature of things.) and *logia* (i.e. theory).
    - means “nature of existence
  - Definition in the Dictionary (Merriam-Webster)
    - *“A branch of metaphysics concerned with the nature and relations of being”*
    - *“A particular theory about the nature of being or the kinds of existents (the kind of things that exists) ”*

What is an Ontology?

Computer Science Definition

- **An ontology is an engineering artifact [Guarino98]:**
  - most prevalent use in AI
  - constituted by a specific **shared** vocabulary used to describe a certain reality, plus
  - a set of explicit assumptions regarding the intended meaning of the vocabulary.
  - usually has the form of a first order logical-theory
  - vocabulary words appear as unary and binary predicates names respectively called concepts and relations
- **An ontology describes a hierarchy of concepts related by subsumption relationships**
  - Suitable axioms are added in order to express relationships between concepts and to constrain their intended interpretation

What is an Ontology?

Computer Science Definition

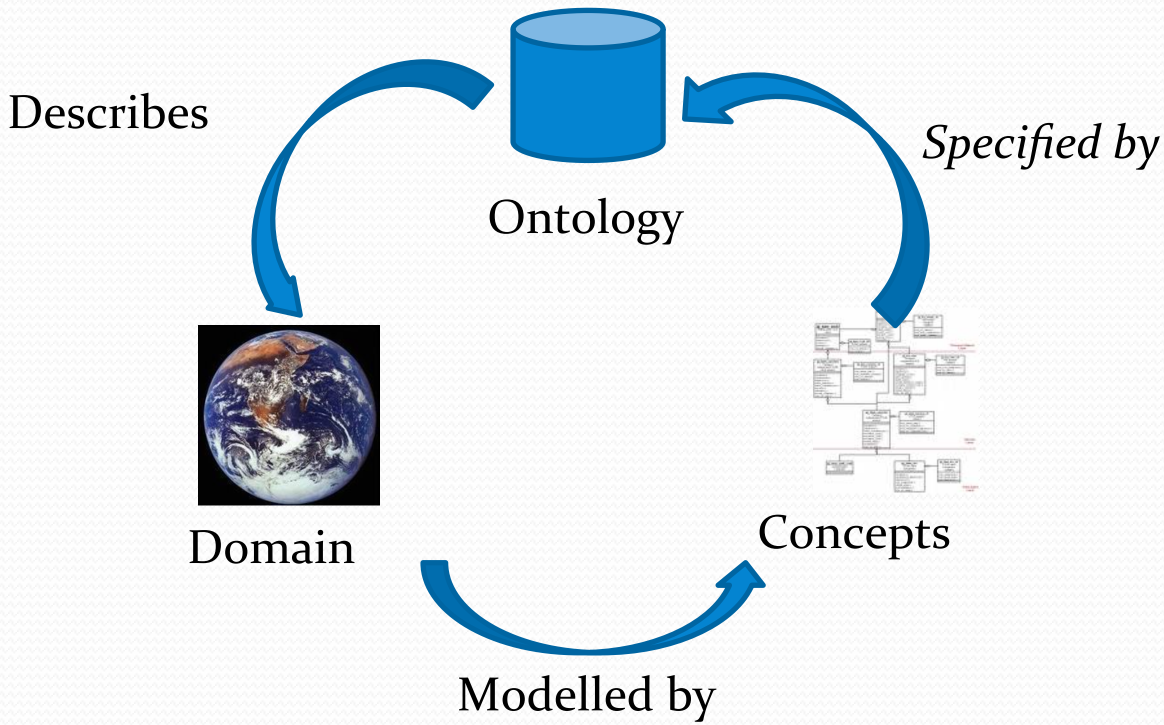
“An ontology is a formal specification of a conceptualization” – Thomas Gruber

- borrows from AI literature on Declarative Knowledge
  is a description (like a formal specification of a program) of the concepts and relationships
  Conceptualisation: abstract and simplified view of the domain of interest.
- objects or entities that are assumed to exist in the domain of interest
- relationships that hold between them(roles)
  formally, <D,R> where D is a domain and R is a set of relevant relations on D

What is an Ontology?

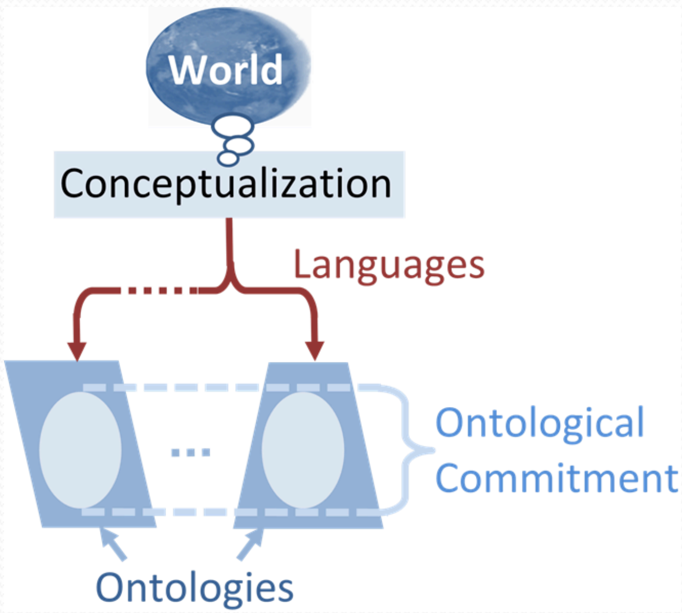
"An ontology is a formal specification of a conceptualization"

– *Thomas Gruber*



Conceptualization

- “An ontology is language dependent while a conceptualisation is language independent ”



English: Cat

中文：猫

Deutsch: Katze

Arabic سنور قط

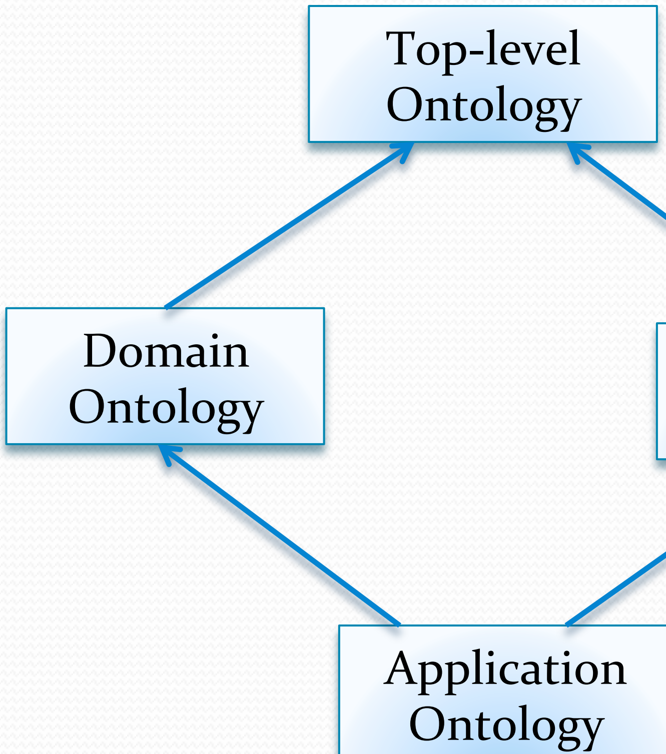
Translating the terms in the ontology from one language to the other (e.g. English to French) does not change the ontology conceptually

Ontological Commitment

- Ontology of a shared domain can be described by defining a set of representational ***terms***.
- Terms (lexical reference) > entities (non lexical referents) knowledge sharing amongst (agents)
- Agreement about the entities and the relations being talked about among agents
- Agreement with respect to semantics of entities and their relationships
- Agreement to use the shared vocabulary in a coherent and consistent manner
- Leads to a formalisation of the vocabulary: with the aim of mapping from terms to entities as exact as possible
- Bridges the semantic gap that exists between the lexical representation of information and its non-lexical conceptualisation

Types of ontologies

General concepts like space, time etc.
Domain independent



Generic domain e.g. Medicine, Education

Generic task
e.g. Stock trading
Online payment

Concepts corresponding to roles played by domain entities while performing certain tasks or activities

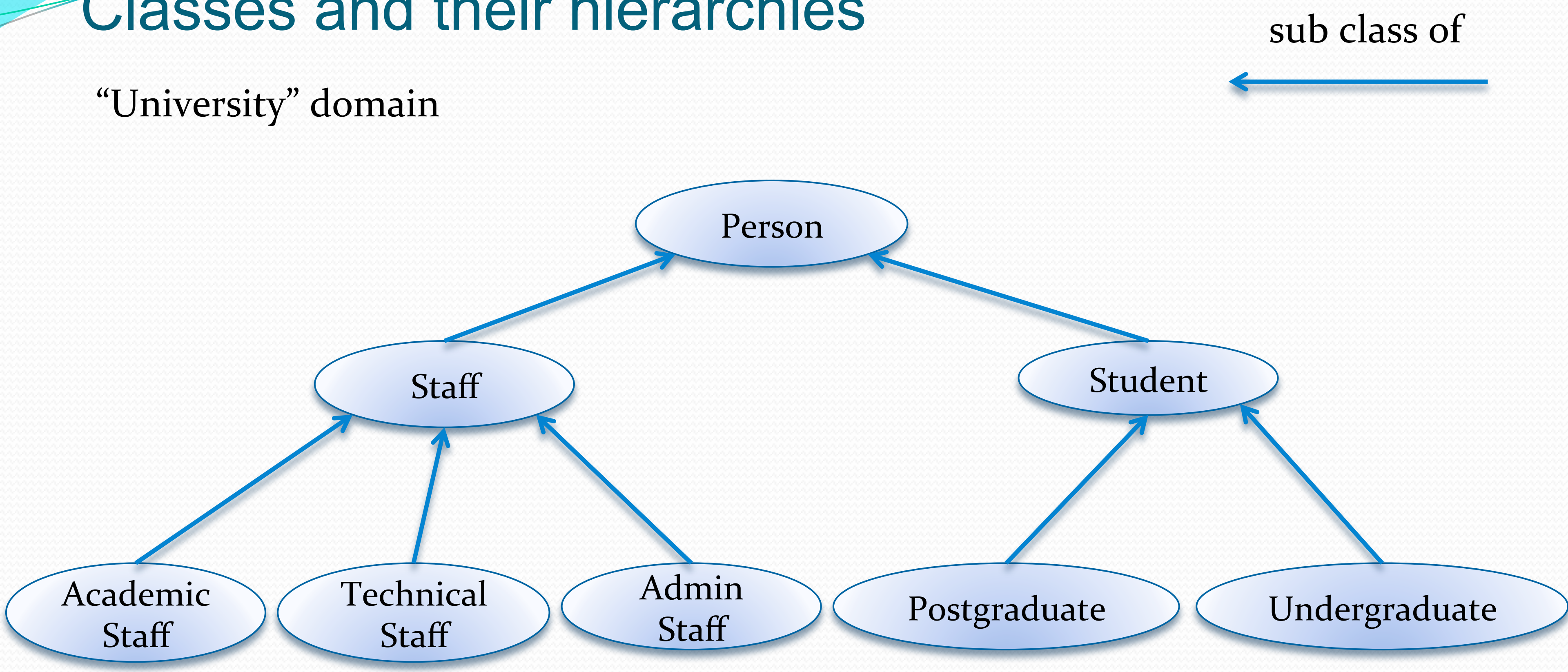
Structure of an Ontology

An ontology consists of a finite lists of terms and relationships between those terms.

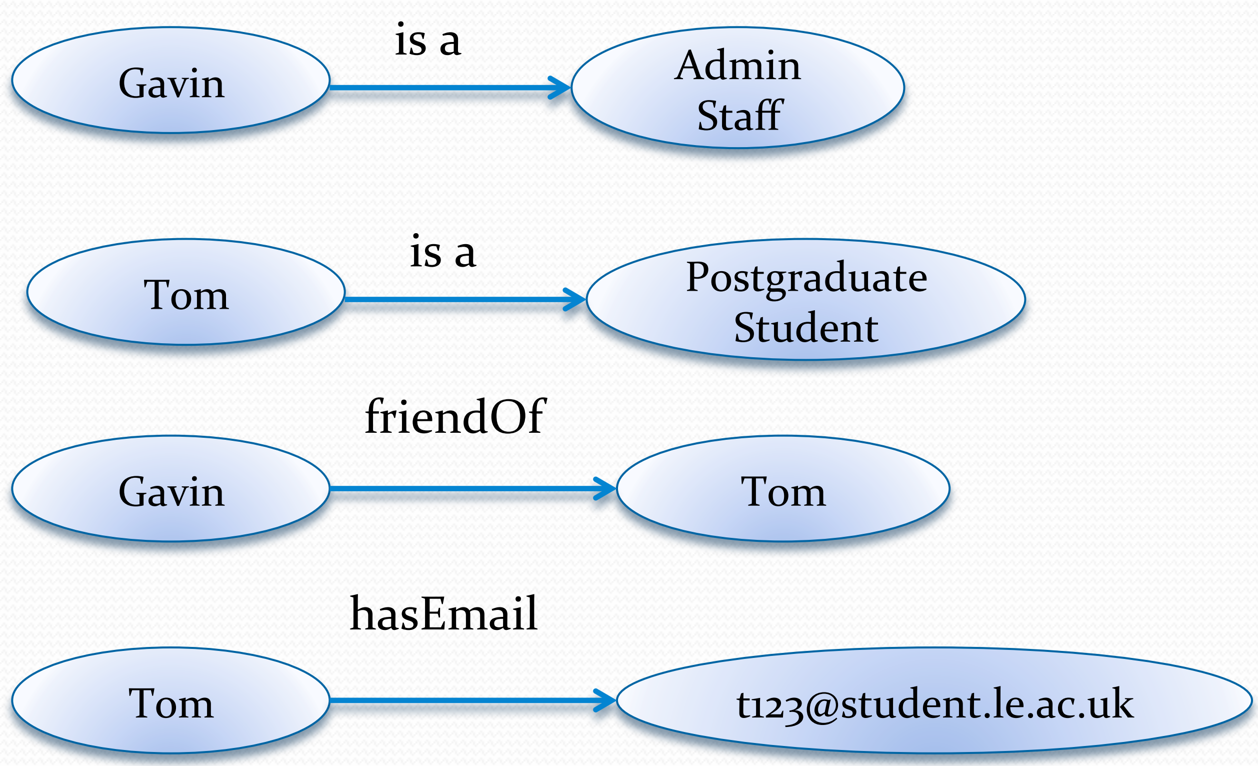
- Ontologies may include information such as:
  - Class-subclass relationship
    - Academic staff (child) is a subclass of University Staff (parent)
  - Class-instance (is-A) relation
    - X is an instance of class Staff
  - Property relationship
    - staff(X) teaches student(Y)
  - Logical Constraints
    - Undergraduate students and Postgraduate students are disjoint
    - Every academic staff can have at most 4 PhD students.

Classes and their hierarchies

“University” domain



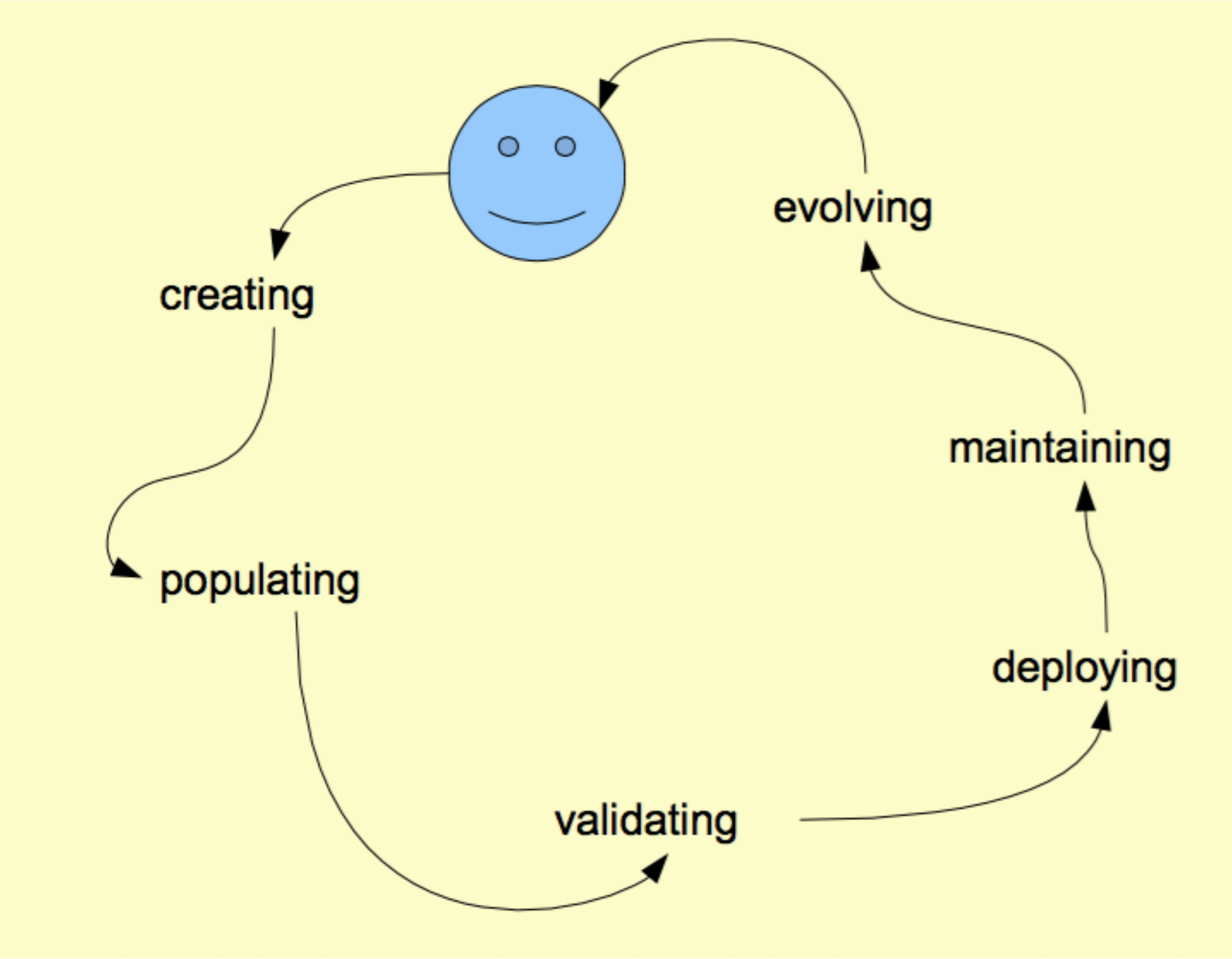
Relationships (Some examples)



Ontology Language for Semantic Web

- Classification
  - Logic types
    - Description Logics
    - First Order Logic (KIF)
    - Frame-based (F-logic)
  - Graphical Notation
    - Semantic Networks Topic Maps
    - UML Class/Object diagram
    - RDF Graph
  - Markup Ontology Languages
    - **RDF Schema**
    - OIL, DAML+OIL, **OWL** (Description Logics-based)

Ontology life-cycle



Tools and Services

- Ontology Editors
  - e.g. Protégé, Fluent Editor, Topbraid Composer
- Validators:
  - e.g. OWL Validator
- Reasoners
  - e.g. HermiT, Racer, FaCT++, Pellet
- Visualisers:
  - e.g. Graphviz
- Storage
  - e.g. Jena TDB, OpenLinkVirtuoso, Sesame,
- Tools for aligning and integrating ontologies :
  - e.g. COMA
- Semantic Search engines
  - e.g. Swoogle

Vocabulary

- “Class” ≈ “Concept” ≈ “Category” ≈ “Type”
- “Instance” ≈ “Individual”
  “Entity” ≈ “object”, Class or individual
- “Property” ≈ “Slot” ≈ “Relation” ≈ “Relation type” ≈ “Attribute” ≈ Semantic link type” ≈ “Role”
  - but be careful about “role”
    - means “property” in DL-speak
    - means “role played” in most ontologies
      e.g. “doctor_role”, “student role” ...

https://www.w3.org/wiki/Lists_of_ontologies

Some Well-known Ontologies

- Dublin Core (Document publishing)
- Gene Ontology (Bioinformatics)
- WordNet (Lexical Database)
- CIDOC CRM (Culture & Heritage)
- FOAF (Friend-Of-A-Friend)
- Geopolitical ontology
- VCard RDF (business card)
- Pizza Ontology (an useful example comes with Protégé editor)

....

Semantic Web Languages - Expressiveness

- XML: surface syntax for structured documents but no semantic constraints
- XML Schema: restricts the structure of XML documents
- RDF: data model for objects(“resources”) and relations between them.
- RDF Schema: Vocabulary description language for describing properties and classes of RDF resources.
- DAML+OIL: DARPA Agent Markup Language. OIL stands for Ontology Inference Layer (superseded by OWL)
- OWL: a richer language for describing classes and am expressive suite of relationships between classes.
- **In this course; We will learn about RDF, RDF-S, OWL**

Semantic Web Languages: Expressiveness

expressiveness

XML

XML-S

RDF

RDF-S

DAML-OIL

OWL

Theoretical Background of Ontology languages -Logic

- A discipline that studies the principles of reasoning.
- Provides formal language for expressing knowledge.
- Provide well understood semantics
- Using logic, automated reasoners can deduce (infer) conclusions from given knowledge.

- For example:
  - OWL is an ontology language and W3C standard – Based on **Description Logic.**

Introduction to Description Logic (DL)

Symbol:		Description:
$C, D \longrightarrow A \mid$	(atomic concept)	
$\top \mid$	(universal concept)	top
$\bot \mid$	(bottom concept)	bottom
$C \sqcap D \mid$	(intersection)	the intersection of two concepts
$C \sqcup D \mid$	(disjunction)	the union of two concepts
$\neg C \mid$	(negation)	the complement of a concept
$\forall R.C \mid$	(value restriction)	the universal restriction of a concept by a role
$\exists R.C \mid$	(existential quantification)	the existential restriction of a concept by a role

Example:

$$
\begin{aligned}
\text{Woman} &\equiv \text{Person} \sqcap \text{Female} \\
\text{Man} &\equiv \text{Person} \sqcap \neg\text{Woman} \\
\text{Mother} &\equiv \text{Woman} \sqcap \exists\text{hasChild.Person} \\
\text{Father} &\equiv \text{Man} \sqcap \exists\text{hasChild.Person} \\
\text{Parent} &\equiv \text{Father} \sqcup \text{Mother} \\
\text{Grandmother} &\equiv \text{Mother} \sqcap \exists\text{hasChild.Parent} \\
\text{MotherWithManyChildren} &\equiv \text{Mother} \sqcap \geqslant 3\,\text{hasChild} \\
\text{MotherWithoutDaughter} &\equiv \text{Mother} \sqcap \forall\text{hasChild.}\neg\text{Woman} \\
\text{Wife} &\equiv \text{Woman} \sqcap \exists\text{hasHusband.Man}
\end{aligned}
$$

We'll get back to this later when studying OWL!!

Logical reasoning: example

- We know
  - All professors are faculty members
    - prof(X)-> faculty(X)
  - All faculty members are staff
    - faculty(X) -> staff(X)
  - Vegetarians are people who don't eat any kind of meat; vegetarians and non-vegetarians are disjointed; people are either vegetarians or non-vegetarians;
    - X.eatOnly(veg) -> vegetarian(X)
  - If X and Y are sibling, Z is the father of Y then Z is the father of X
    - sibling(X,Y) father(Z,Y) -> father(Z,X)
- And
  - John is a Professor
    - Prof(John)
  - John eat meats.
    - John.eatSome(meat)
  - Alice’s biological father is John, Tom is Alice’s brother
    - father(John, Alice) brother(Tom, Alice)

Logical reasoning: example

- We can deduce
  - John is a faculty
    - faculty(John)
  - John is also a member of Staff
    - Staff(John)
  - John is a non-vegetarian
    - non-vegetarian(John)
  - John is Tom's father
    - father(John, Tom)

Description Logic reasoning

Deductive rule-based reasoning

We'll get back to this later when studying reasoning over OWL ontologies!!

Logic: advantage

- Logic is used to uncover ontological knowledge which is implicitly given.
- It can help uncover unexpected relationships and inconsistencies
- Intelligent agents can use knowledge to make decisions and selecting courses of action
- Logic for semantic web must be machine processable.

Open World Assumption (OWA)

Yi eats
chicken tikka
boiled rice
potato salads
tofu

Alessandro eats
pasta
boiled rice
potato salads
hash browns

Open World Assumption
Knowledge is **incomplete**, hence we have to admit **undefined** answers

Vegetarian?	Yi	Alessandro
Closed world assumption	No	Yes
Open world assumption	No	???

Summary: Features of the Semantic Web

- *AAA* - Anyone can say Anything about Any topic.
- A good deal of information for the SW is already available on the web.
- Open world/close world: there is always more information that could be known.
- Nonunique naming: the same entity could be known by more than one name.
- The network effect: The value of joining in increases with the number of people who have joined.

Conclusions

- Current Web has limitations in terms of describing data that is published.
- Semantic Web endeavors to describe data on the web in a “meaningful” way to make it suitable for processing by machines.
- Key technologies for realising the Semantic Web include: metadata, ontologies, logic and inferencing, and software agents.

Acknowledgements

Lecture Notes based on slides provided by

Monika Solanki (Leicester)

Some slides inspired from:

Ian Horrocks (University of Oxford) and his working group

The intelligent agent image was inspired from a slide by

Charlie Abela (University of Malta)