COMP3220:
Document Processing and Semantic Technologies
The Semantic Web Architecture

Rolf Schwitter

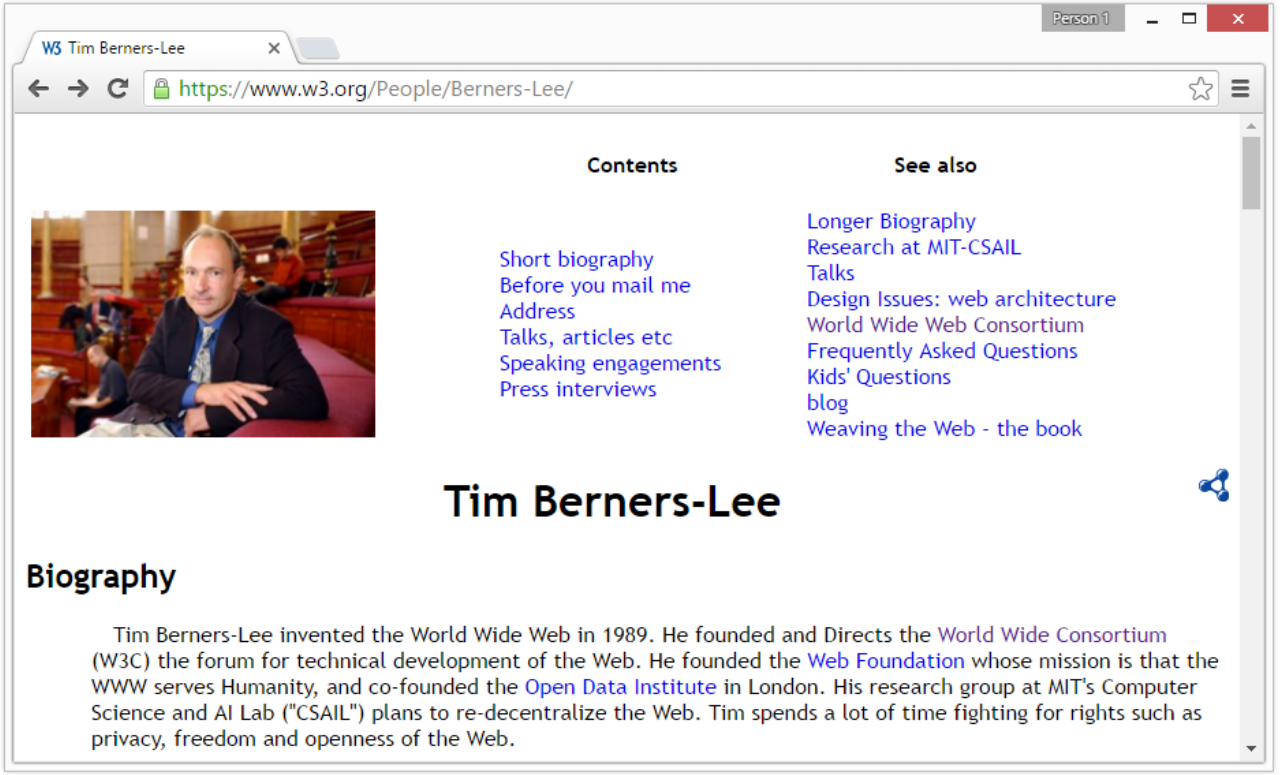
Rolf.Schwitter@mq.edu.au

Today's Agenda

- Problems with the existing Web
- The Semantic Web
- Semantic Web Architecture
- Four Key Standards

Problems with the Existing Web

- Consider a typical Web page of the existing Web:



Problems with the Existing Web

- HTML markup
  - for rendering information (e.g. headers, paragraphs, etc.)
  - hypertext links to related content.
- Most of the Web content is designed
  - for humans to read
  - not for computers to manipulate meaningfully.

Problems with the Existing Web

- Humans do the hard things:
  - linking and interpreting information.
- Computers do the easy things:
  - rendering information.
- Let computers do more of the hard work.
- Prerequisites:
  - information needs to be machine-readable and machine-processable
  - a language is required that links data.

Tim Berners-Lee's View

- "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." [Berners-Lee et al. 2001]
- The existing Web is a web of documents.
- The Semantic Web is a web of data.
- It enables people to create data stores on the Web, build vocabularies, and write rules for handling data.

The Semantic Web

SCIENTIFIC AMERICAN.COM

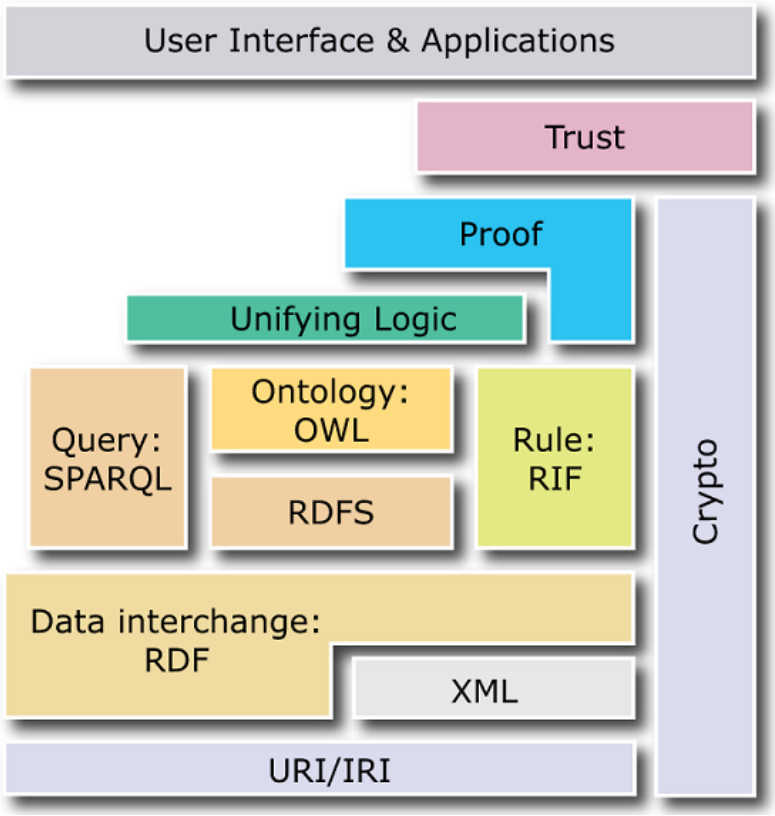
May 17, 2001

The Semantic Web

A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities

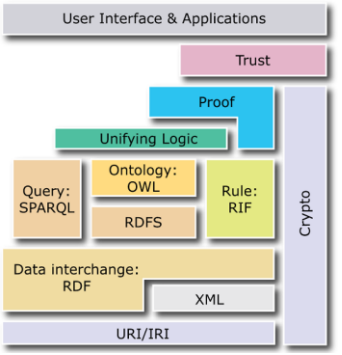
By Tim Berners-Lee, James Hendler and Ora Lassila

Semantic Web Architecture



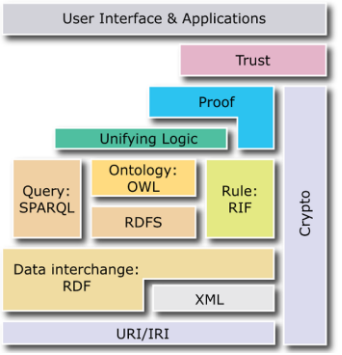
http://www.w3.org/2007/03/layerCake.svg

URI/IRI and Unicode



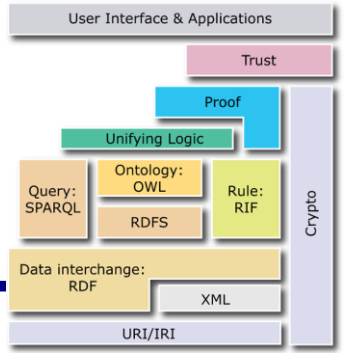
- URI/IRI and Unicode are already used in the existing web.
- A URI/IRI is a string of a standardised form.
- A URI/IRI allows to uniquely identify a resource (e.g., a document).
- Unicode is a standard of encoding international character sets.
- Unicode allows that human languages can be used on the web.

XML and Namespaces



- XML guarantees a common syntax for the Semantic Web.
- XML is a general purpose markup language for documents that contain structured information.
- XML documents contain elements that can be nested and that may have attributes and content.
- XML namespaces provide a method to avoid element name conflicts.
- XML Schema (or a document type definition) describes the structure of an XML document.

RDF (Resource Description Framework)



- RDF is based on the idea of
  - identifying things using IRIs and
  - describing resources[1] in terms of properties and values.
- RDF represents
  - simple statements about resources
  - as a graph of nodes and arcs
  - representing the resources, their properties and their values.

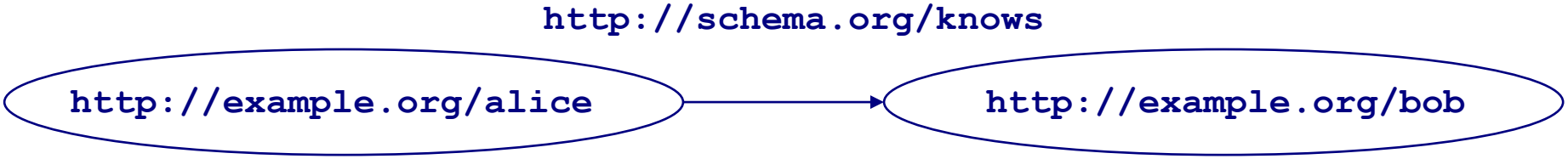
[1] Note that RDF describes a resource as any object that is uniquely identifiable by an URI/IRI.

RDF

- For example, the statement:

  Alice knows Bob.

  can be represented as an RDF graph:



- That means IRI references identify things referred to in RDF statements.

RDF

- RDF statements consist of three parts:
  - subject, predicate, object.
- For example, in the natural language statement:

  Alice knows Bob.

  the three RDF parts are:
  - subject: Alice
  - predicate: knows
  - object: Bob.
- RDF provides an XML-based exchange syntax for graphs.

RDF/XML

```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF xmlns:rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:schema = "http://schema.org/">

  <rdf:Description rdf:about = "http://example.org/alice">
    <schema:knows rdf:resource = "http://example.org/bob"/>
  </rdf:Description>

</rdf:RDF>
```

N-Triples Notation

```
<http://example.org/alice>
<http://schema.org/knows>
<http://example.org/bob> .
```

Turtle Notation

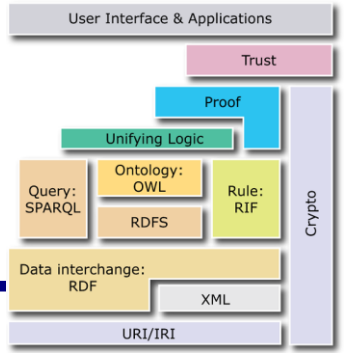
```
@prefix schema: <http://schema.org/> .

<http://example.org/alice> schema:knows <http://example.org/bob> .
```

Different Syntaxes

- A number of different serialisation formats exist for writing down RDF graphs:
  - RDF/XML (XML syntax for RDF)
  - Turtle family of RDF languages (N-Triples, Turtle, TriG, N-Quads)
  - RDFa (for embedding RDF into HTML and XML)
  - JSON-LD (JSON-based RDF syntax).

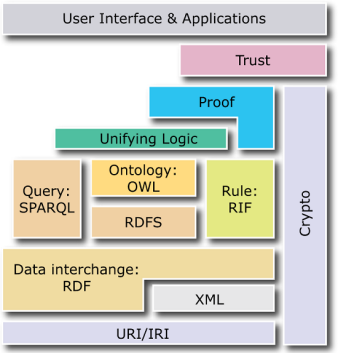
RDF Schema



- RDF provides a way to express statements about resources.
- But it offers no means for defining classes and properties.
- RDF Schema is a vocabulary for describing application-specific classes and properties of RDF resources.
- For example (in Turtle syntax):

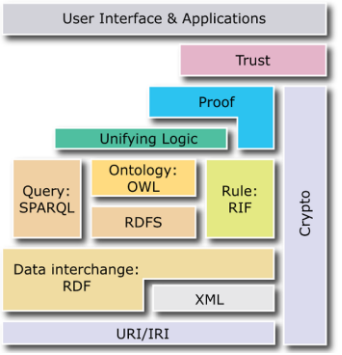
```
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ex:   <http://www.example.org/> .
   ex:AcademicStaff rdf:type rdfs:Class .
   ex:Professor rdfs:subClassOf ex:AcademicStaff .
   ex:inform rdf:type rdf:Property .
   ex:teach rdfs:subPropertyOf ex:inform .
```

Query: SPARQL



- SPARQL = Protocol and RDF Query Language
- SPARQL is an SQL-like query language for querying RDF data.
- But the RDF data model is different from relational databases:
  - RDF: open world assumption (incomplete knowledge)
  - databases: closed world assumption (complete knowledge).
- SPARQL consists of three main parts:
  - query language specification
  - query result XML format
  - data access protocol (for querying remote RDF databases).

Ontology: OWL



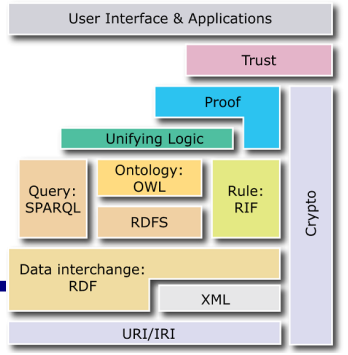
- RDF(S) is a very limited knowledge representation language.
- OWL extends RDF(S) by adding more advanced constructs to describe the semantics of statements.
- OWL is based on a version of description logic.
- OWL brings reasoning power to the Semantic Web.
- OWL 2 offers three different profiles (= fragments) tailored for different application domains.

OWL Ontology

Excerpt of an OWL ontology in Turtle notation:

```
:GraduateStudent rdf:type owl:Class ;
                 rdfs:subClassOf :Student ,
                                 [ rdf:type owl:Restriction ;
                                   owl:onProperty :hasDegree ;
                                   owl:someValuesFrom
                                     [ rdf:type owl:Class ;
                                       owl:oneOf ( :BA :BS )
                                 ].
```

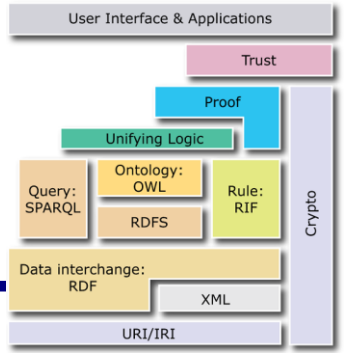
Rule



- Rules provide support for relations that cannot be expressed in OWL.
- For example:

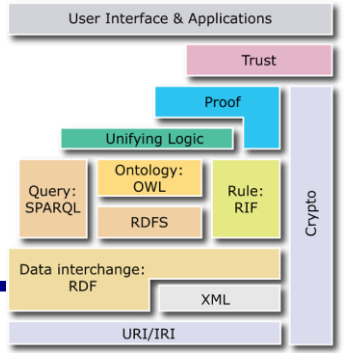
  If an actor plays a role in a film then the film stars the actor.
- Description Logic Programs (DLP), Answer Set Programming (ASP), and Semantic Web Rule Language (SWRL) are proposals for rule languages.
- RIF is an interchange format for different rule languages.
- RIF is based on the observation that many rule languages already exist.

Unifying Logic



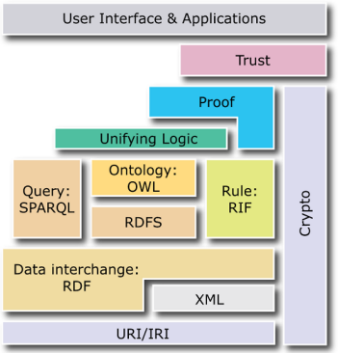
- Unifying Logic is the framework that includes the other languages as subsets: RDF(S), OWL, SPARQL, and rules.
- Each of these subsets is tailored for a specific kind of inference engine and/or a specific range of uses.
- What unifies them is the common model-theoretic semantics.
- The semantics enables all of them to interoperate on shared data and produce consistent results.
- Common Logic (CL): a framework for a family of logic-based languages: https://www.iso.org/standard/66249.html

Proof



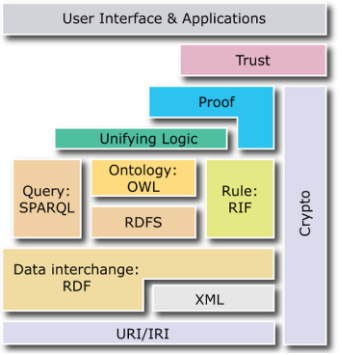
- How can I believe what the computer tells me?
- Agents (= information processors) will use logical operations to prove a range of assertions.
- Agents need to explain a human how they came to a particular conclusion.
- Once a proof has been carried out, it can be posted on the web.
- Other agents can use this proof without proving the results again.
- But, how do we know that a computer is not "lying"?
- Proofs must be augmented by trust (via cryptography).

Cryptography



- Cryptography ensures and verifies that semantic web statements are coming from trusted sources.
- This can be achieved via public-key cryptography and digital signature of statements.
- Digital signatures are encrypted messages (= data).
- Agents
  - can use this data to verify the source of information
  - have to check the sources of the information ($\rightarrow$ trust).

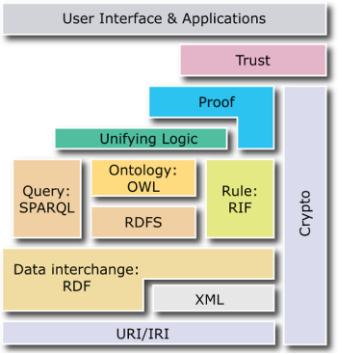
Trust



- A Web of Trust:

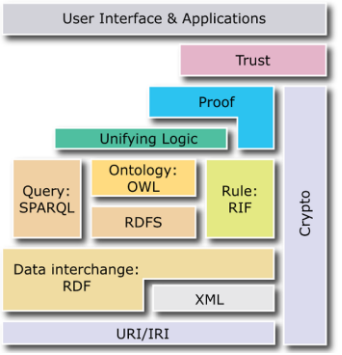
  If I trust Sue, and Sue trusts Bob and Jim, then I may decide to trust Bob and Jim, too.
- Distrust could also be propagated.
- We have to figure out a trust model.
- We have to formalise arrangements and write rules that encapsulate those arrangements.
- Again, these rules need to be machine-processable.

User Interface



- User interface guarantees cooperation between human and machine.
- High-level interface languages are required that are machine-processable as well as human-readable.
- Controlled natural languages can serve as high-level interface languages.
- User expresses statements and questions in controlled natural language.
- Machine translates these statements into a formal target notation.

Four Key Standards



- So far, the Semantic Web consists of four key standards:
  - RDF: the data modelling language of the Semantic Web.

    https://www.w3.org/TR/rdf11-primer/
  - SPARQL: the query language of the Semantic Web.

    http://www.w3.org/TR/sparql11-query/
  - OWL: the knowledge representation language of the Semantic Web.

    https://www.w3.org/TR/owl2-primer/
  - RIF: Rule Interchange format.

    http://www.w3.org/TR/rif-overview/

Take-Home Messages

- The existing Web is about linked hypertext documents.
- The Semantic Web is about linked data and knowledge discovery.
- Important technologies are:
  - RDF and RDF Schema
  - SPARQL (Query Language)
  - OWL (Ontology Web Language)
  - RIF (Rule Interchange Format)