

Introduction to the Semantic Web and Linked Open Data

Dr Nicholas Gibbins
nmg@ecs.soton.ac.uk

Goals

- Introduce the concepts and technologies behind the Semantic Web and the Web of Linked Data
- Show the range of uses to which SW/LOD is currently put
- Explain the dotAC approach to linked data and research information management
- Demonstrate the linked data support in Eprints 3.2

Non-Goals

- Detailed tutorial on the finer points of:
 - RDF
 - RDF Schema
 - OWL
 - SPARQL
 - ...
- Too much for one day!

Dramatis Personae



Nick Gibbins



Chris Gutteridge



Hugh Glaser



Ian Millard



Les Carr

History of the Semantic Web

- The World Wide Web was invented by Tim Berners-Lee (amongst others) while he was working at CERN
- TBL's original vision of the Web was much more ambitious than the reality of the existing (syntactic) Web
- TBL (and others) have since been working towards realising this vision, which has become known as the Semantic Web

History of the Semantic Web

“... a goal of the Web was that, if the interaction between person and hypertext could be so intuitive that the **machine-readable information space** gave an accurate representation of the state of people's thoughts, interactions, and work patterns, then **machine analysis could become a very powerful management tool**, seeing patterns in our work and facilitating our working together through the typical problems which beset the management of large organizations.”



T. Berners-Lee, The World Wide Web: Past, Present and Future, 1996

What is the Semantic Web?

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

It is the idea of having data on the Web defined and linked in a way that it can be used **for more effective discovery, automation, integration and reuse** across various applications.

The Web can reach its full potential if it becomes a place where **data can be processed by automated tools as well as people**.



THE
SEMANTIC
WEB

Making the Web Semantic...

<http://www2002.org>

WWW 2002

**THE ELEVENTH INTERNATIONAL
WORLD WIDE WEB CONFERENCE**



Sheraton Waikiki Hotel
Honolulu, Hawaii, USA
7-11 May 2002



HAWAII

1 LOCATION. 5 DAYS. LEARN. INTERACT.

Registered participants coming from:

Australia · Canada · Chile · Denmark · France · Germany · Ghana · Hong Kong · India · Italy · Ireland · Japan · Malta · New Zealand · The Netherlands · Norway · Singapore · Switzerland · The United States · Vietnam · Zambia

REGISTER NOW

On 7-11 May 2002, Honolulu, Hawaii will provide the backdrop for The Eleventh International World Wide Web Conference. This prestigious series of the International World Wide Web Conference Committee (IW³C²) attracts participants from around the world, and it provides a public forum for the World Wide Web Consortium (W3C) through the annual W3C track.


The conference is being organized by the [International World Wide Web Conference Committee \(IW³C²\)](#), the [University of Hawaii](#) and the [Pacific Telecommunications Council \(PTC\)](#).

FEATURED SPEAKERS (CONFIRMED)

	Tim Berners-Lee, inventor of the World Wide Web and Director of the W3C who now holds the 3Com Founders chair at the Laboratory for Computer Science (LCS) at the Massachusetts Institute of Technology (MIT).		Richard A. DeMillo, vice president and chief technology officer for Hewlett-Packard Company.
	Ian Foster, guru of "Grid Computing", associate		McArthur Prize Winner,

[Conference Proceedings](#)
[Call for Participation](#)
[Program](#)
[Registration Information](#)
[Hotel Accommodation](#)
[Conference Committee](#)
[Sponsorship/Exhibition Opportunities](#)
[Volunteer Information](#)
[Information about Hawaii](#)
[Previous & Future WWW Conferences](#)

...in a way that machines can deal with....



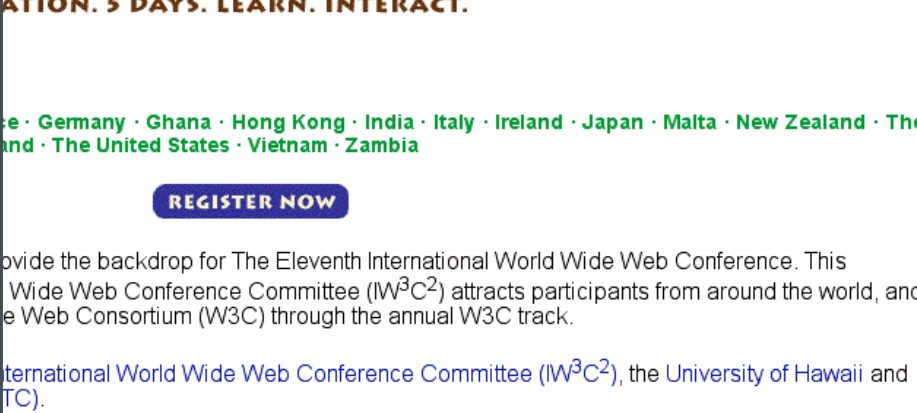
The banner features a URL <http://www2002.org> in a red box, the title **WWW 2002** in a red box, and the subtitle **THE ELEVENTH INTERNATIONAL WORLD WIDE WEB CONFERENCE** in a red box. It also includes the location: Sheraton Waikiki Hotel, Honolulu, Hawaii, USA, 7-11 May 2002, and the logo of the International World Wide Web Conference Committee.

This is a type of object event and this is its title

This is the URL of the web page for the event

This is a type of object photograph and the photograph is of Tim Berners-Lee

Tim Berners-Lee is an invited speaker at the event



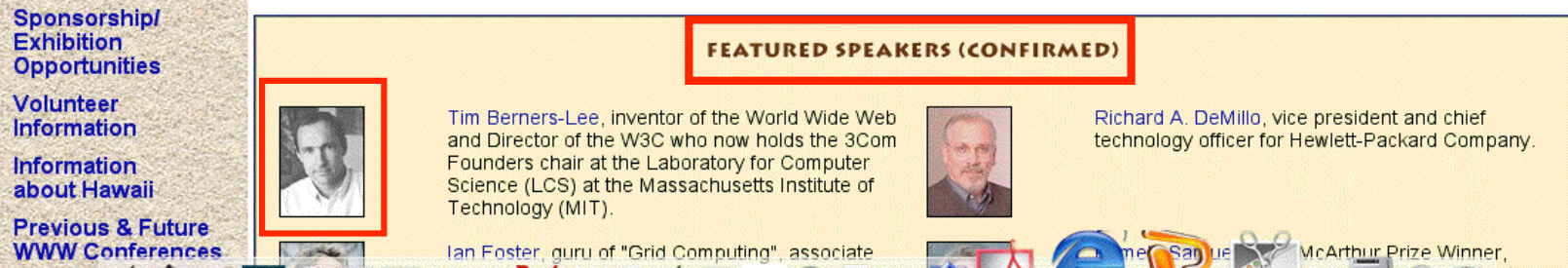
LOCATION. 5 DAYS. LEARN. INTERACT.

Registered participants coming from:
Australia · Canada · Chile · Denmark · France · Germany · Ghana · Hong Kong · India · Italy · Ireland · Japan · Malta · New Zealand · The Netherlands · The United States · Vietnam · Zambia


REGISTER NOW


provide the backdrop for The Eleventh International World Wide Web Conference. This prestigious series of the International World Wide Web Conference Committee (IW³C²) attracts participants from around the world, and the Web Consortium (W3C) through the annual W3C track.


The conference is being organized by the International World Wide Web Conference Committee (IW³C²), the University of Hawaii and the Pacific Telecommunications Council (PTC).




FEATURED SPEAKERS (CONFIRMED)


 Tim Berners-Lee, inventor of the World Wide Web and Director of the W3C who now holds the 3Com Founders chair at the Laboratory for Computer Science (LCS) at the Massachusetts Institute of Technology (MIT).

 Richard A. DeMillo, vice president and chief technology officer for Hewlett-Packard Company.

 Ian Foster, guru of "Grid Computing", associate

 me, San, ue, McArthur Prize Winner

...via meta content



<http://www2002.org>

WWW 2002
THE ELEVENTH INTERNATIONAL
WORLD WIDE WEB CONFERENCE

Sheraton Waikiki Hotel
Honolulu, Hawaii, USA
7-11 May 2002

```
<owl:Class rdf:ID="Conference">
<rdfs:subClassOf rdf:resource="#Meeting-Taking-Place"/>
<rdfs:subClassOf rdf:resource="#Publication-Type-Event"/>
-<rdfs:subClassOf>
-<owl:Restriction>
<owl:onProperty rdf:resource="#published-proceedings"/>
<owl:allValuesFrom rdf:resource="#Conference-Proceedings-Reference"/>
</owl:Restriction>
```

This is a type of object event and this is its title

This is the URL of the web page for the event

This is a type of object photograph and the photograph is of Tim Berners-Lee

Tim Berners-Lee is an invited speaker at the event

LOCATION. 5 DAYS. LEARN. INTERACT.

Registered participants coming from:
Australia · Canada · Chile · Denmark · France · Germany · Ghana · Hong Kong · India · Italy · Ireland · Japan · Malta · New Zealand · The United States · Vietnam · Zambia

REGISTER NOW

Honolulu, Hawaii will provide the backdrop for The Eleventh International World Wide Web Conference. This prestigious series of the International World Wide Web Conference Committee (IW³C²) attracts participants from around the world, and it provides a public forum for the World Wide Web Consortium (W3C) through the annual W3C track.

The conference is being organized by the International World Wide Web Conference Committee (IW³C²), the University of Hawaii and the Pacific Telecommunications Council (PTC).

FEATURED SPEAKERS (CONFIRMED)



Tim Berners-Lee, inventor of the World Wide Web and Director of the W3C who now holds the 3Com Founders chair at the Laboratory for Computer Science (LCS) at the Massachusetts Institute of Technology (MIT).



Richard A. DeMillo, vice president and chief technology officer for Hewlett-Packard Company.



Ian Foster, guru of "Grid Computing", associate



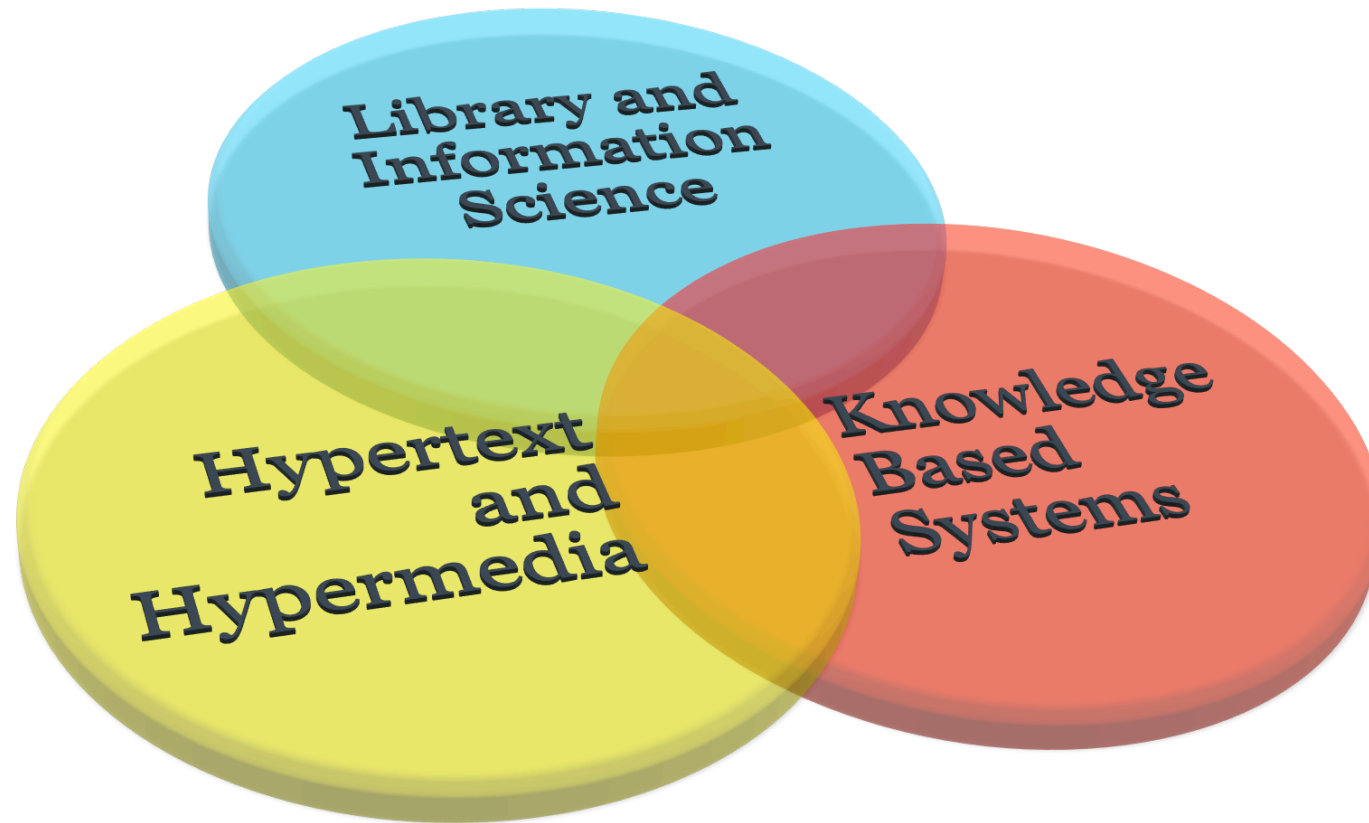
McArthur Prize Winner

The image is a close-up of the central portion of Michelangelo's famous fresco, 'The Creation of Adam'. It depicts two hands reaching toward each other: the hand of God on the right, extended from a reclining position, and the hand of Adam on the left, reaching out from a similar reclining position. The fingertips are just inches apart, creating a sense of tension and divine spark. The background is the aged, cracked, and yellowed plaster of the original artwork.

Origins of the Semantic Web

<http://www.flickr.com/photos/adebond1/3311255710/>

Interwoven Themes





3554 Industrial We
Industries, Location of - M

3566 Information Theory in B
Inga

3578 Insects (by place) B
Inside the H

3590

3555 Industries, Location of - M
Industry - Social Aspects - B

3567 Ingb
Inglir

3579 Inside the I
Institut C

3591

3556 Industry - Social Aspects - C
Industry and State

3568 Inglis
Inheritance and H

3580 Institut D
Institut Francais de B

3592

3557 Industry and State
(by place) H
Indyn

Inheritance and I
Injections, H

3581 Institut Francais de T
Institut Fa

3593

Library and Information Science

<http://www.flickr.com/photos/annarbor/4350629792/>

Metadata

- The origins of the Semantic Web lie in metadata
- Metadata is data about data
 - A webpage is data
 - A description of the webpage is metadata
 - Metadata for a webpage could include
 - author
 - date of publication
 - file size
 - ...
- Library cataloguing = metadata

Beyond metadata

- The scope of the modern Semantic Web goes beyond bibliographic metadata for webpages
- Metadata is still just data
- If we have an infrastructure for metadata, we can use it for data in general



Knowledge Based Systems

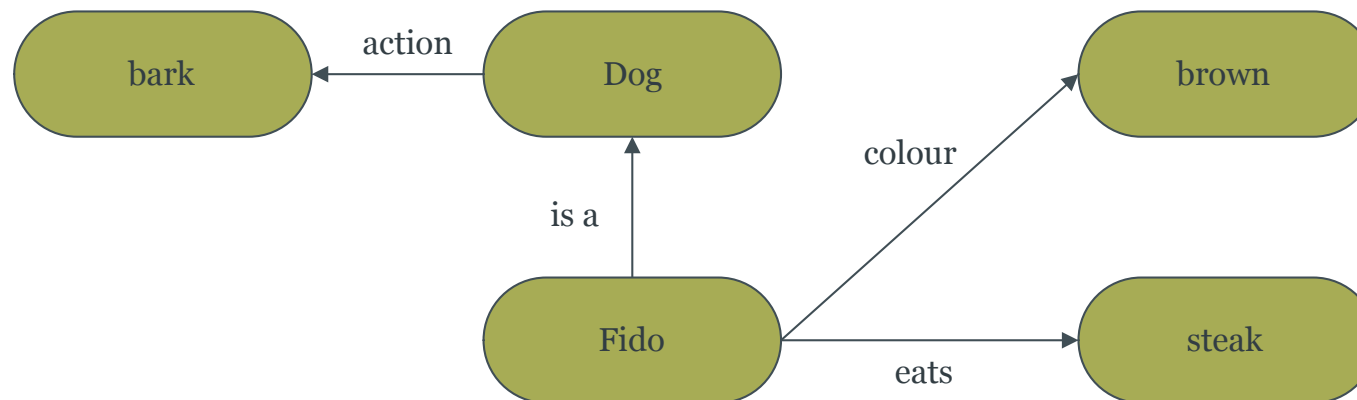
<http://www.flickr.com/photos/donsolo/3768623542/>

Knowledge representation

- Long-standing discipline within Artificial Intelligence
 - (the Semantic Web has a strong heritage!)
- Knowledge representation languages should:
 - Handle qualitative knowledge
 - Allow new knowledge to be inferred
 - Represent both the general and the specific
 - Capture complex meaning
 - Allow meta-level reasoning
- RDF, RDF Schema and OWL are knowledge representation languages

Network knowledge representation

- “Traditional” knowledge representation is formal logic
- Network knowledge representation originated in 1960s with psychologists and linguists
- Knowledge is represented as a graph
 - Nodes are objects or concepts
 - Edges are relations or associations



Vocabularies

- A knowledge representation language by itself is of little use
- We need to be able to tailor the language to our application domain
 - The bibliographic domain needs to be able to talk about works and authors
 - The e-commerce domain needs to be able to talk about orders and prices
 - ...
- We need domain-specific vocabularies (or **ontologies** – more on these later)



Hypertext and Hypermedia

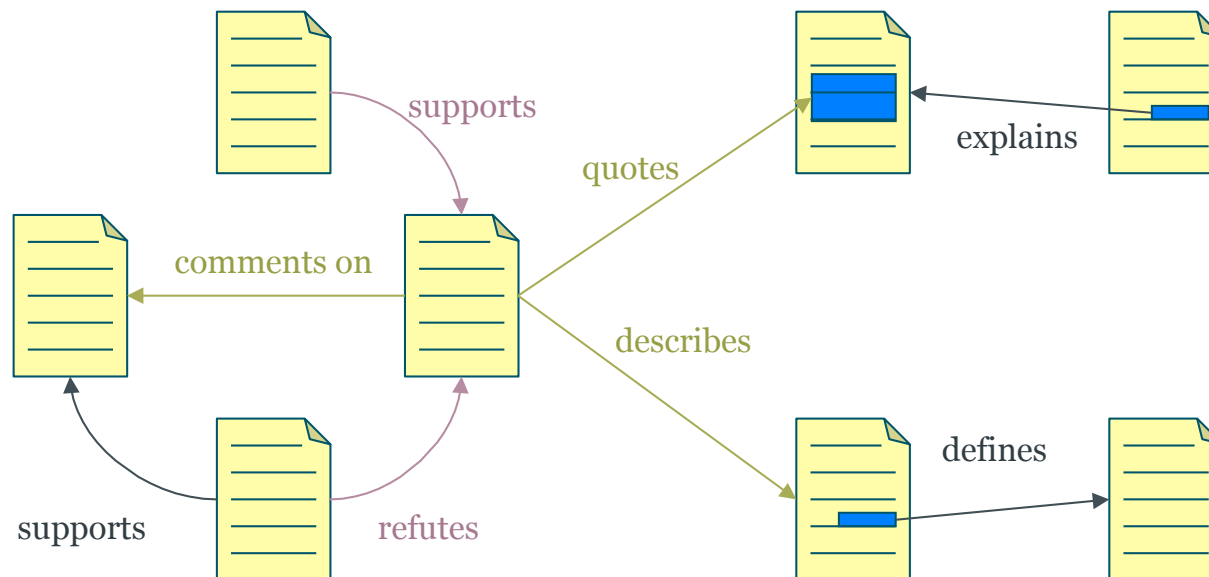
<http://www.flickr.com/photos/intherough/3470183543/>

Hypertext and hypermedia

- Non-linear writing
- Annotation and commentary
- Association of ideas
- Essence of hypermedia is connections
 - Relationships in an abstract domain
 - Implemented as navigable links
- Many kinds of relationships
 - Author-of, homepage-of, see-also, background-info, definition, more-detail
 - Typed links
- Links are complex structures
 - Multivalent, rich metadata
 - Not just simple GOTOs

Open Hypermedia

- Links should be
 - first-class objects
 - manipulated independently



Open hypermedia versus network KR

- Open hypermedia makes links between different bits of knowledge
 - Knowledge is expressed as text, images, etc
- Network knowledge representation makes links that are knowledge
- Are typed hypermedia links knowledge?
- Is a set of hypermedia link types an ontology?



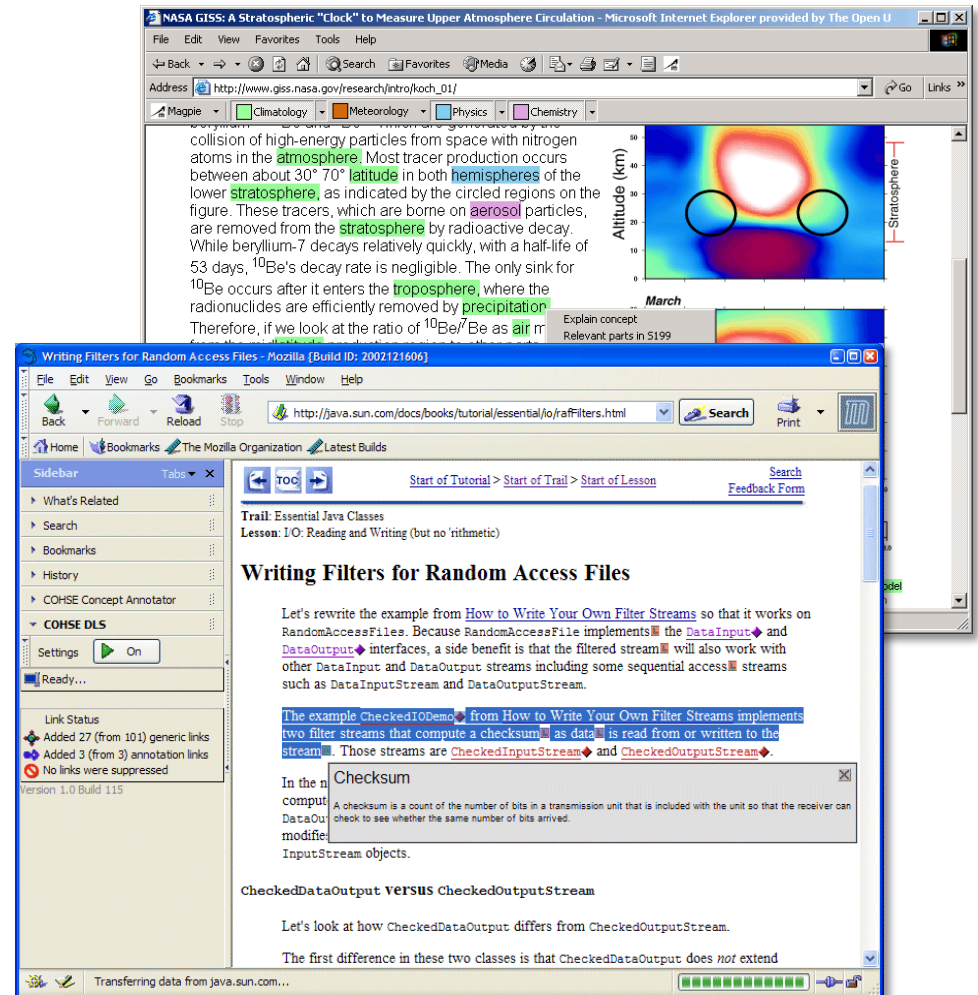
Semantic Web Foundations

<http://www.flickr.com/photos/kpc1975/3961856601/>

Which Semantic Web?

The Annotated Web

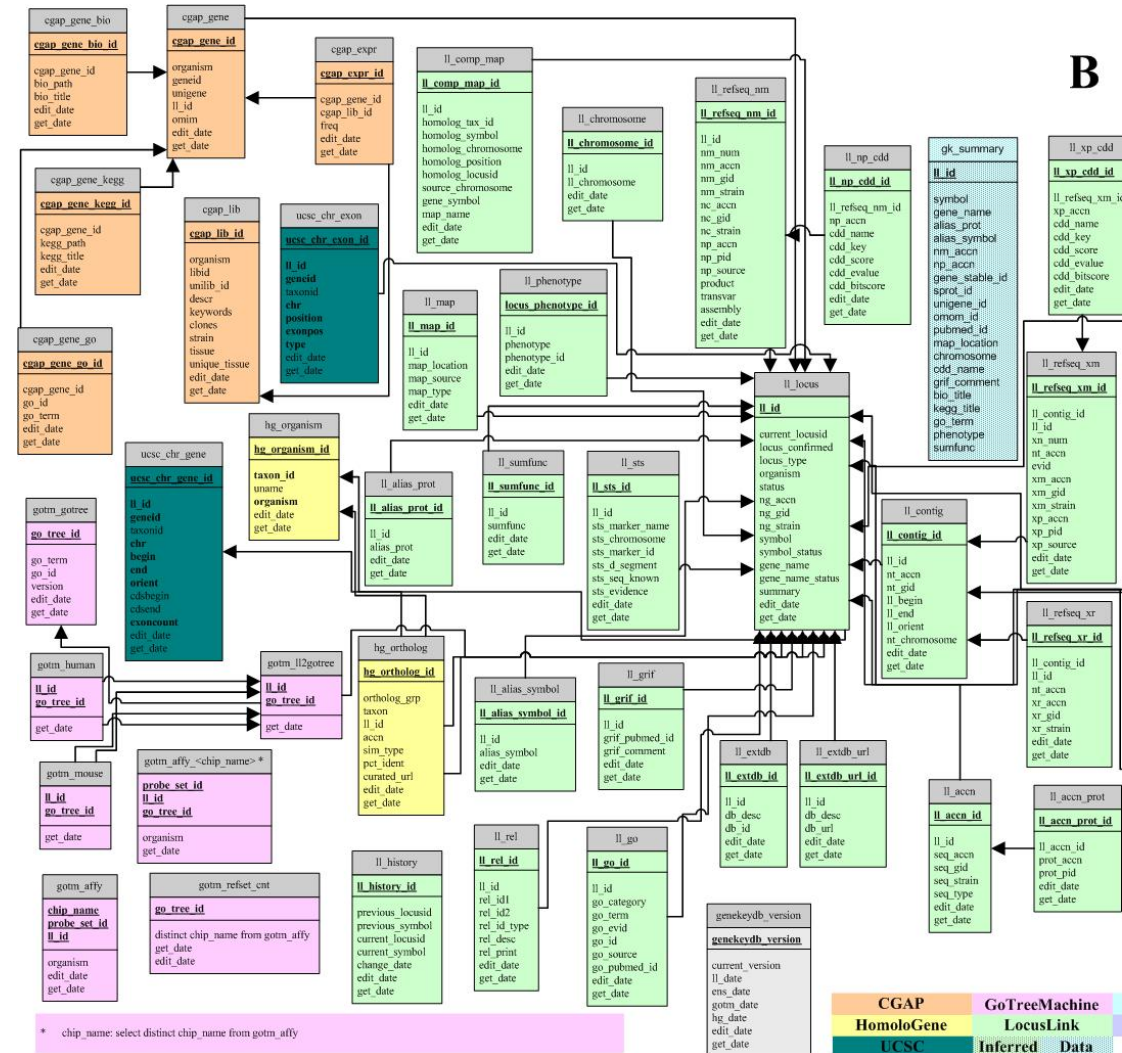
- Enrich existing web pages with annotations
- Classify web pages
- Use natural language techniques to extract information from web pages
- Annotations enable enhanced browsing and searching



Which Semantic Web?

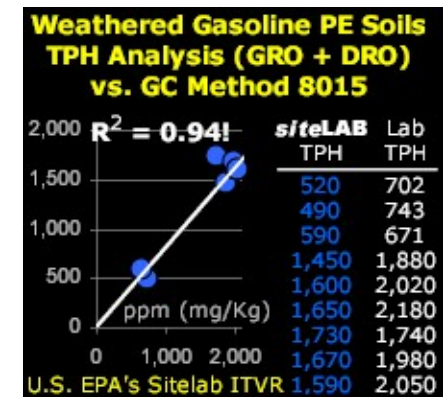
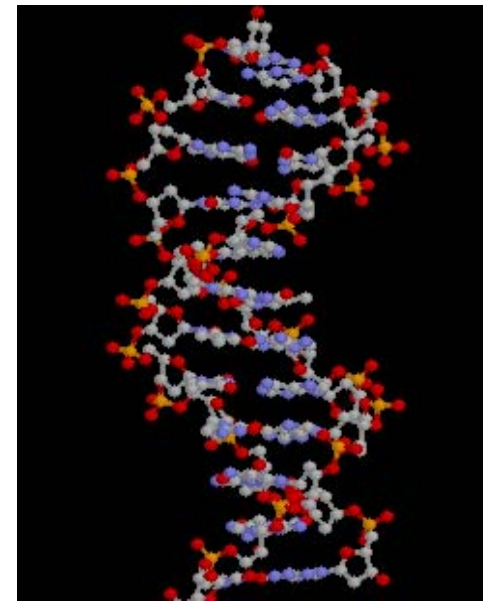
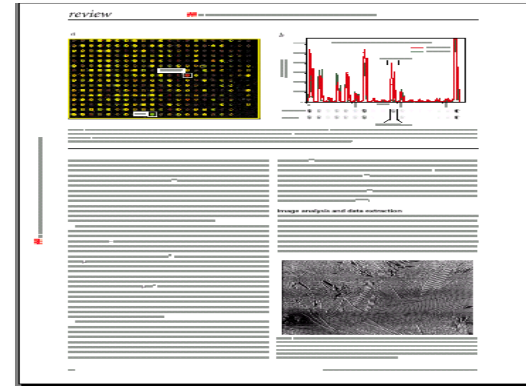
The Web of Data

- Expose existing databases in a common format
- Express database schemas in a machine-understandable form
- Common format allows the integration of data in unexpected ways
- Machine-understandable schemas allow reasoning about data
- Make the most of the structure you already have



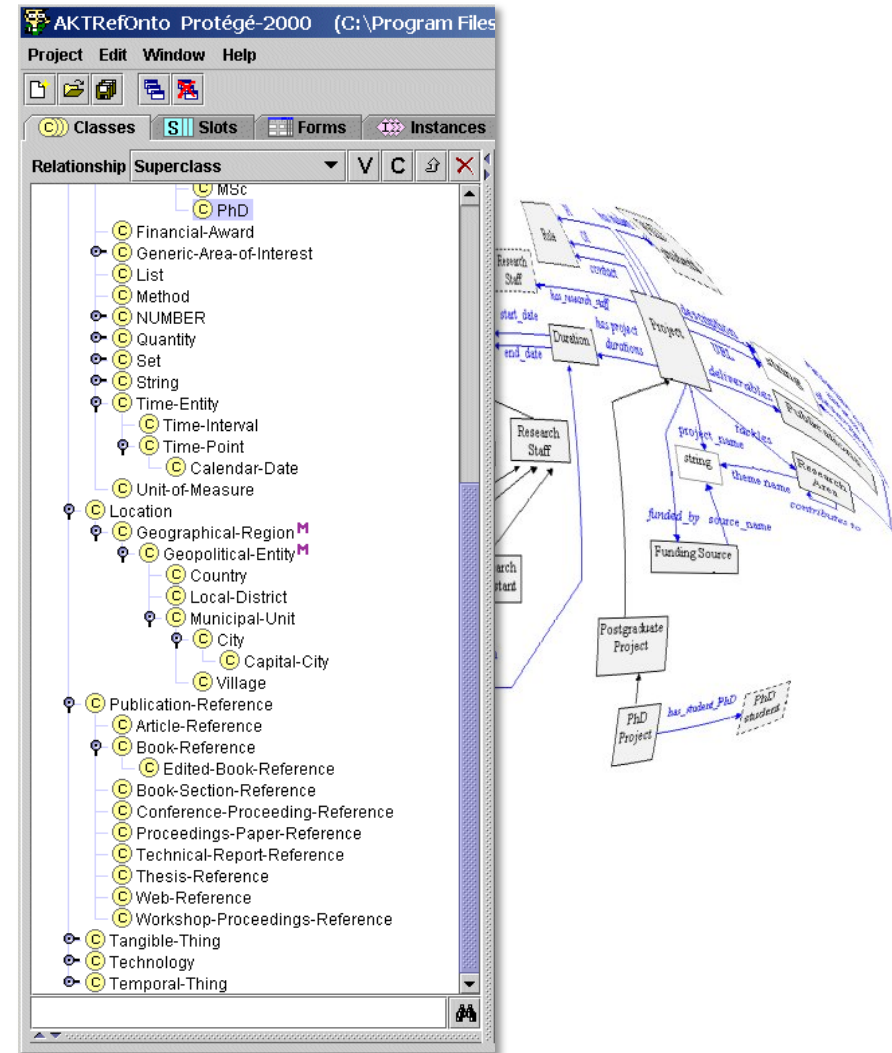
Can Annotate/Expose Anything

- Databases
- Scientific structures
- Workflows
- Publications
- People
- Product Descriptions
- Geographical Information
- Financial Data



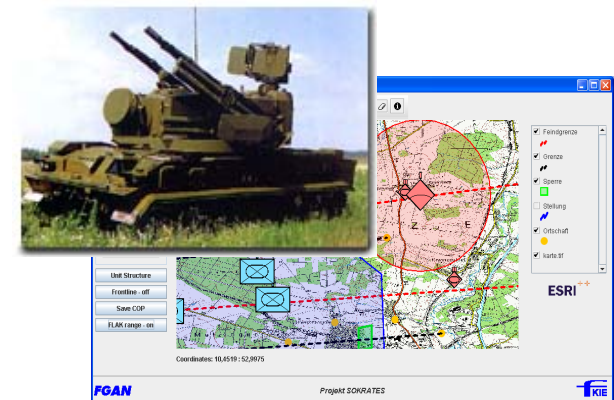
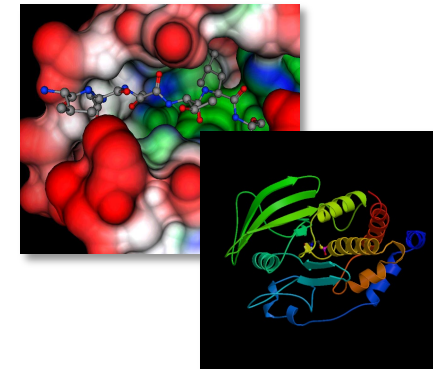
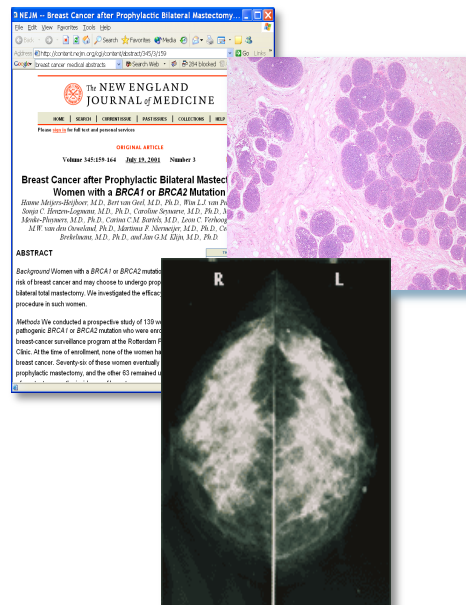
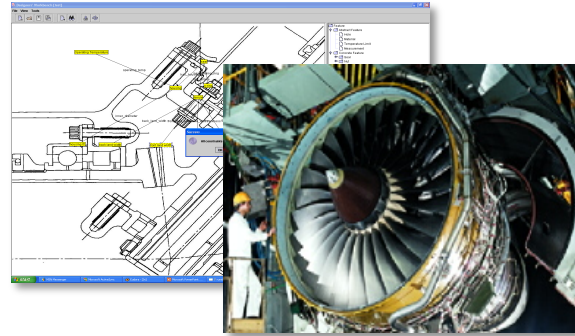
Building Shared Meaning

- To effectively share data, we need a shared language to describe objects in the domain of interest
- Need to agree a common viewpoint and terminology
- Key term: **ontologies**



Building Shared Meaning

- Benefits
 - Communication
 - Inter-operability and Integration
 - Sharing and Reuse
 - Control
- Examples
 - eScience
 - Genomics, proteomics...
 - Medicine
 - Manufacturing
 - Social networking
 - DoD/MoD
 - ...



Defining the ‘O’ word

- “A specification of a conceptualisation”
 - **Specification:** A formal description
 - **Conceptualisation:** The objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them
- More simply, an ontology consists of:
 - A specific vocabulary used to describe a certain reality
 - A set of explicit assumptions regarding the intended meaning of the vocabulary

Isn't <insert name here> an ontology?

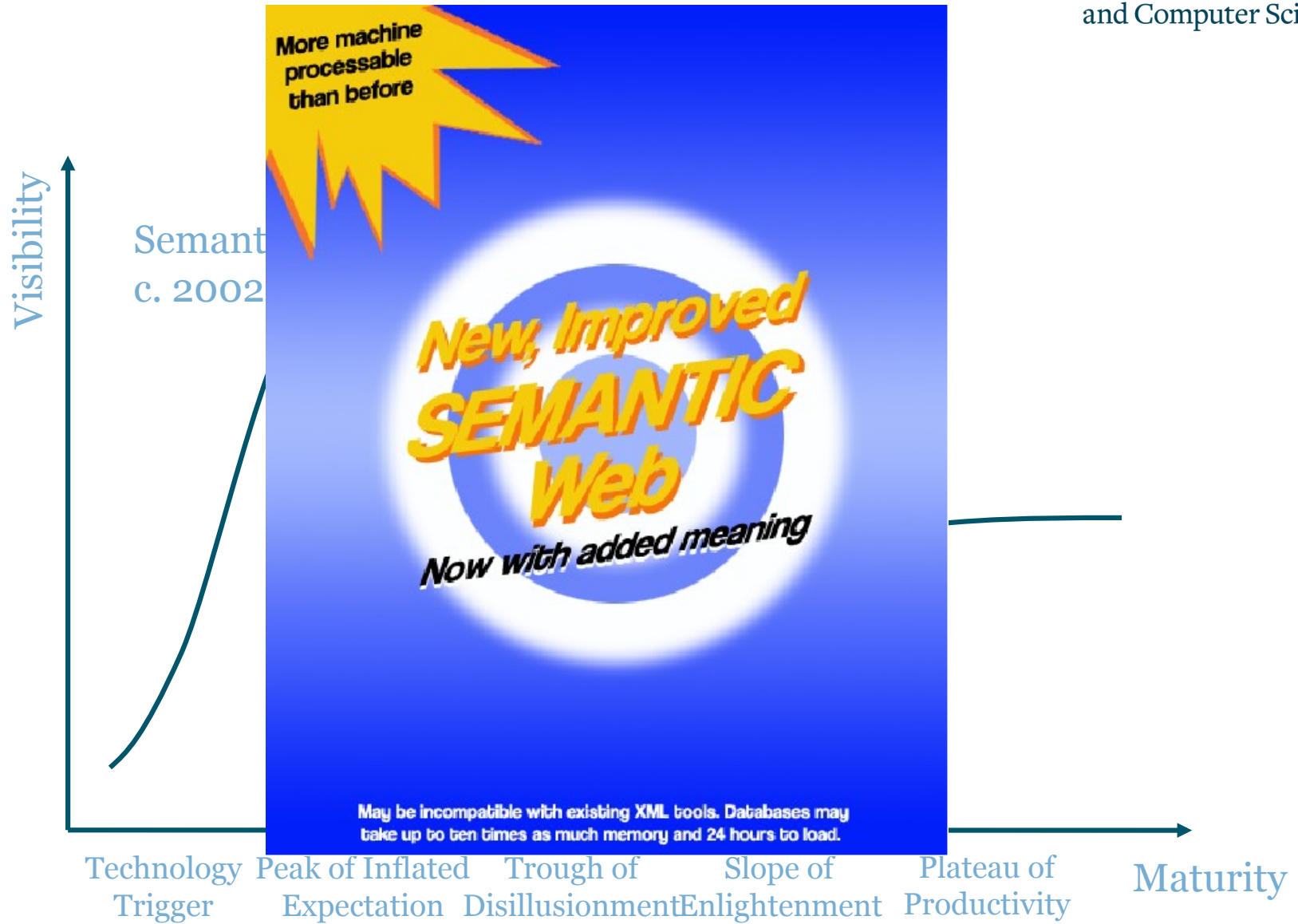
- Controlled vocabulary
 - LCSH, MESH
- Taxonomy (or other hierarchy)
 - LCC, Dewey, UMLS Metathesaurus
- Taxonomy + relations = ontology
 - Gene Ontology, GALEN, UMLS Semantic Network
- Ontology + rules
 - ...

The Cynic's View

The Semantic Web is just
old-fashioned artificial intelligence

Artificial intelligence hasn't delivered on its
previous promises, so why should it now?

The Semantic Web Hype Cycle



Rocket Science (not)

Is this rocket science? Well, not really. The Semantic Web, like the World Wide Web, is just taking well established ideas, and making them work interoperably over the Internet. This is done with standards, which is what the World Wide Web Consortium is all about. **We are not inventing relational models for data, or query systems or rule-based systems. We are just webizing them.** We are just allowing them to work together in a decentralized system - without a human having to custom handcraft every connection.

Tim Berners-Lee, Business Case for the Semantic Web,
<http://www.w3.org/DesignIssues/Business>

The Cynic's View, part 2

Don't we already have a machine-readable
interchange format, namely XML?

On the World Wide Web...

- The **World Wide Web** is the **Web for people**
 - Information is predominantly textual
 - Technologies include URI, HTTP, XML, HTML
 - Information needs humans to give it meaning
- XML is a **machine-readable** format
 - It can be parsed to give an unambiguous document structure

but

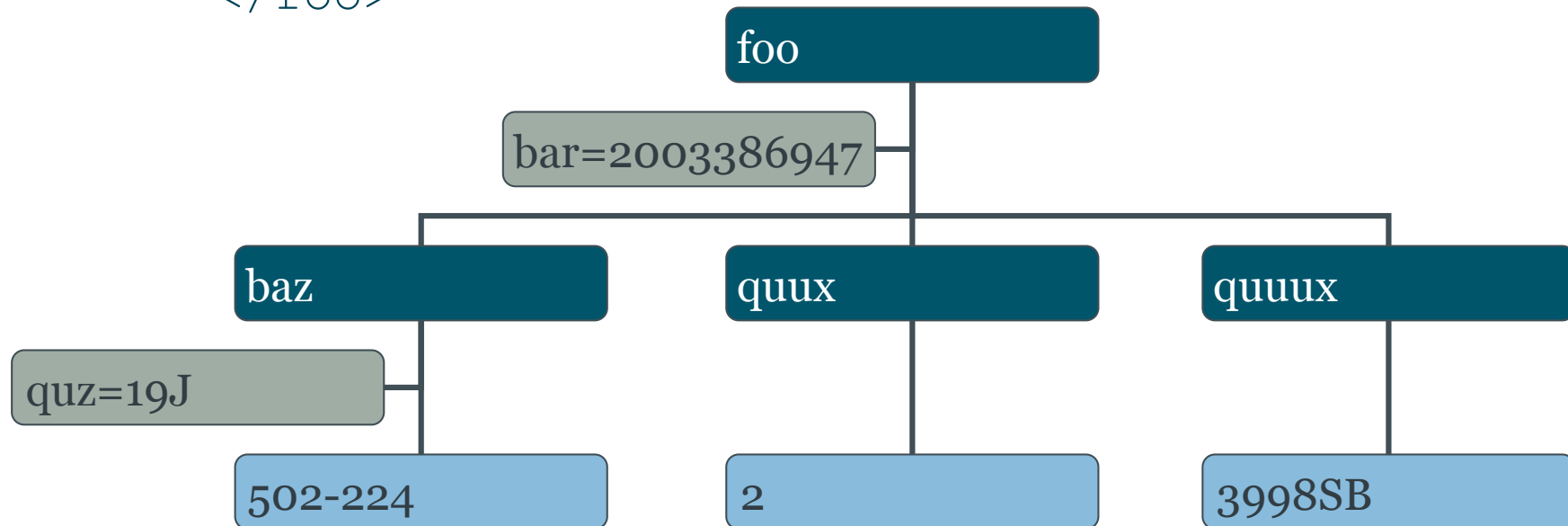
- It has no formal meaning
- Meanings of XML interchange formats must be explicitly agreed

On the Semantic Web...

- The **Semantic Web** is the **Web for machines**
 - Information is structured
 - Technologies include RDF, RDFS, OWL (in addition to those for the Web)
 - Information can be interpreted by machines
 - Humans need not interact directly with Semantic Web information
- RDF is a **machine-understandable** format
 - The structures generated by an RDF parser have a formal meaning
 - RDF is a framework for interchange formats that provides a base level of common understanding

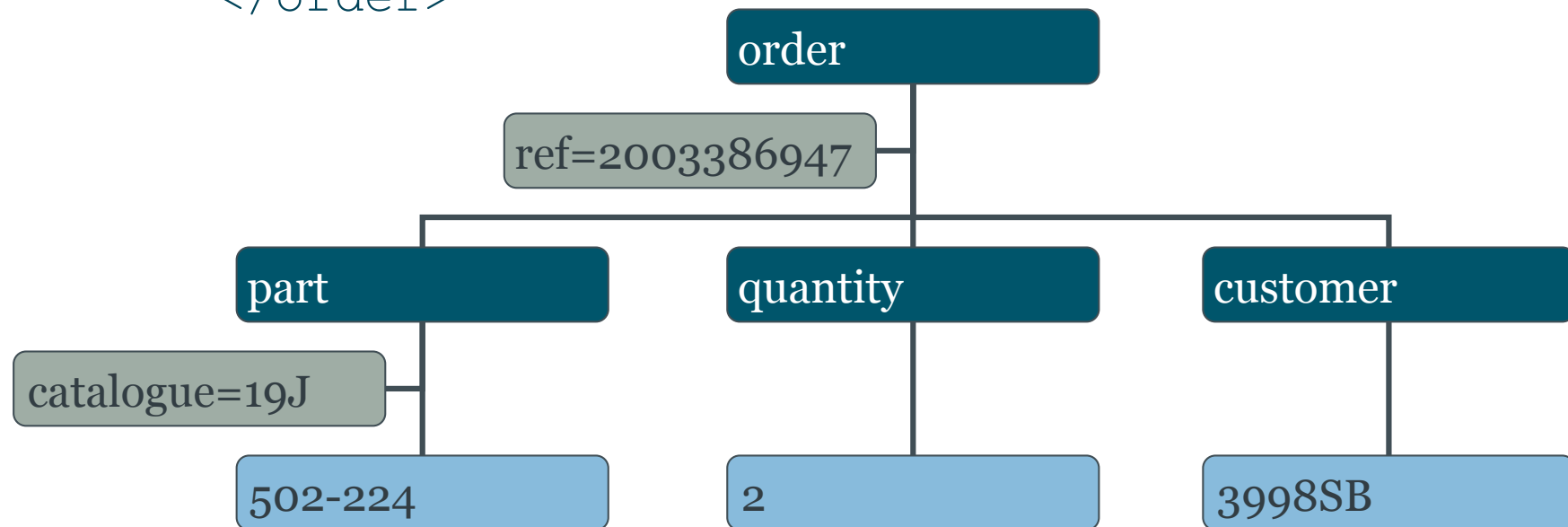
Machine readable: XML

```
<foo bar="2003386947">  
  <baz quux="19J">502-224</baz>  
  <quux>2</quux>  
  <quuux>3998SB</quuux>  
</foo>
```



Machine readable: XML

```
<order ref="2003386947">  
  <part catalogue="19J">502-224</part>  
  <quantity>2</quantity>  
  <customer>3998SB</customer>  
</order>
```



- RDF data may be freely composed with other RDF data
- Assertions can be mixed with assertions about weather, physics, business processes, weblogs and syndication, genealogy, politics, and so on, **without the need for prior coordination** between the designers of vocabularies for these domains
- The RDF position is that it is **too difficult** to agree on interchange formats
- Instead, we should use ontologies to facilitate mapping between formats



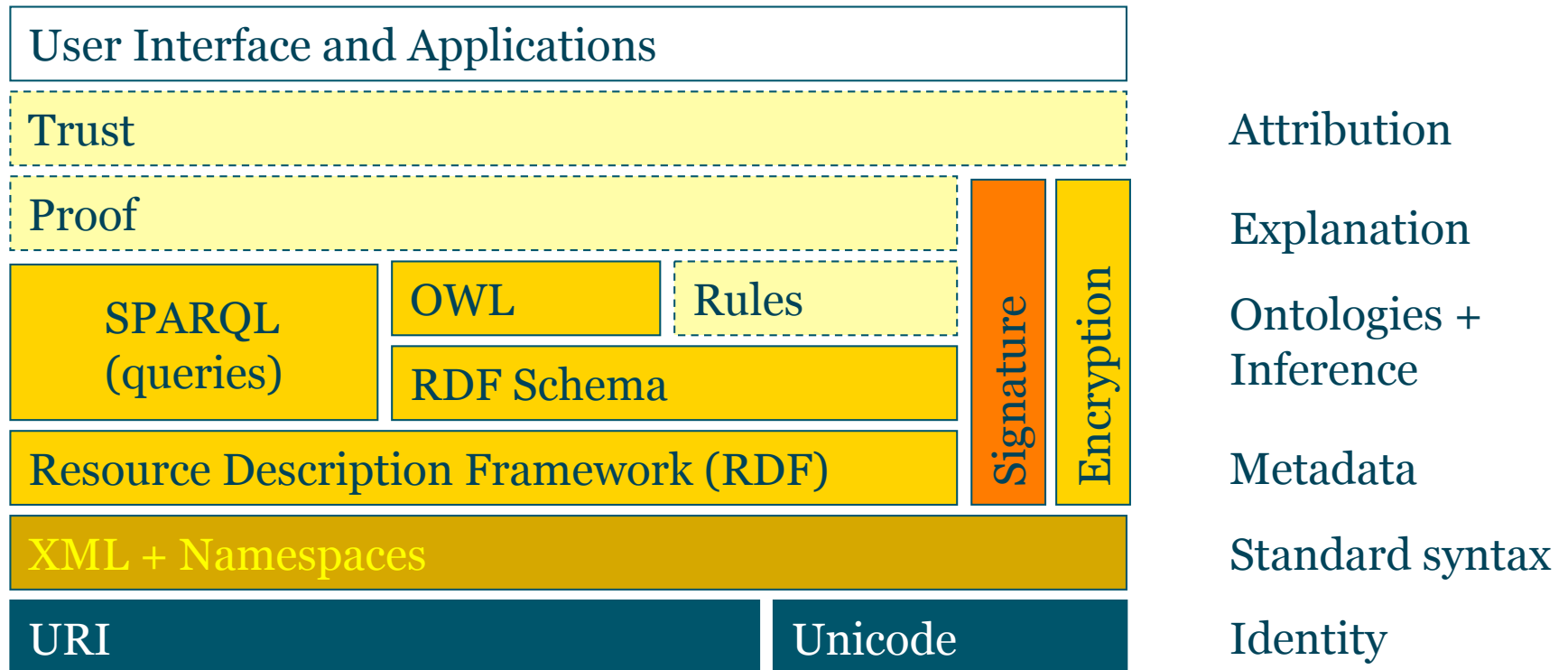
Technical Architecture

<http://www.flickr.com/photos/29225114@N08/2778223048/>

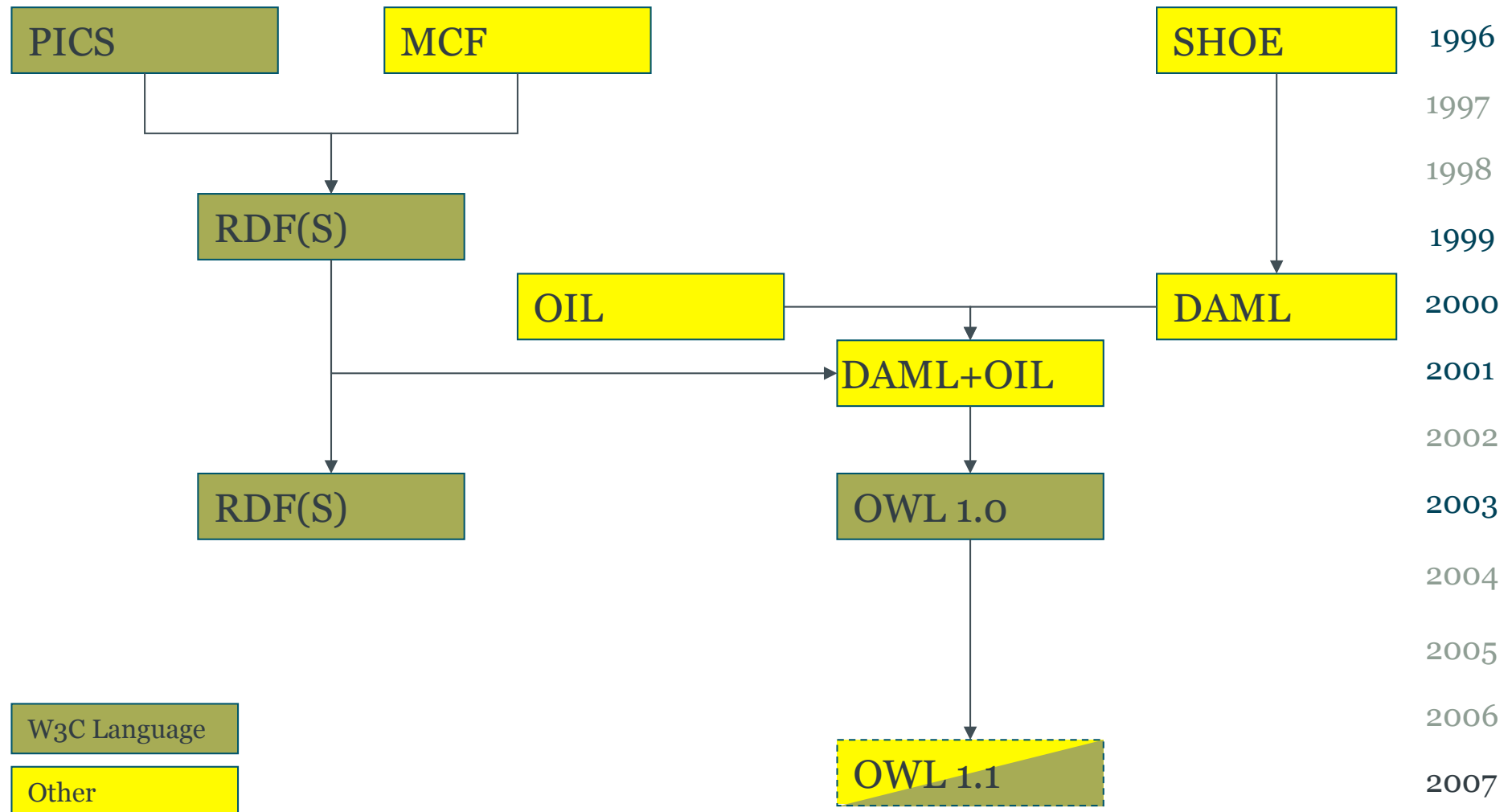
Semantic Web Principles

- Anyone can make assertions about anything
- Entities are referred to using Uniform Resource Identifiers
- Based on XML technologies
- Formal semantics

The Semantic Web layer cake

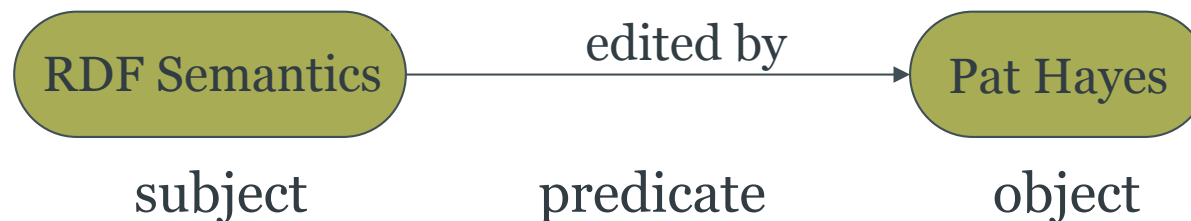


Languages of the Semantic Web



The triple

- Underlying model of triples used to describe the relations between entities in the Semantic Web
- This is the basis of the RDF data model
- (subject, predicate, object)
 - e.g. “RDF Semantics”, “edited by”, “Pat Hayes”

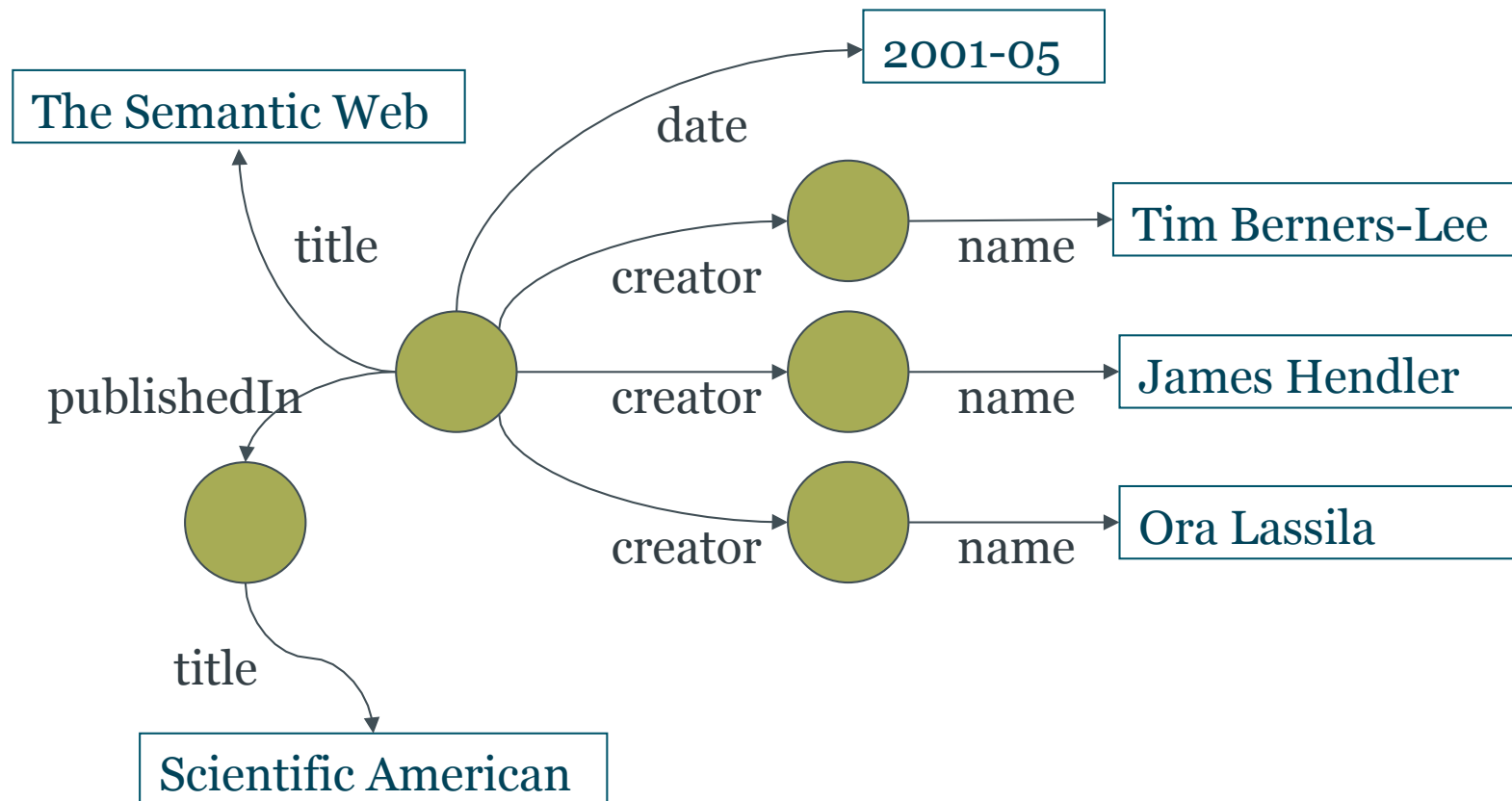


Example

- Take a citation:
 - Tim Berners-Lee, James Hendler and Ora Lassila. The Semantic Web. Scientific American, May 2001
- We can identify a number of distinct statements in this citation:
 - There is an article titled “The Semantic Web”
 - One of its authors is a person named “Tim Berners-Lee” (etc)
 - It appeared in a publication titled “Scientific American”
 - It was published in May 2001

Example

- We can represent these statements graphically:



Example

- There are two types of node in this graph:
 - **Literals**, which have a value but no identity (a string, a number, a date)

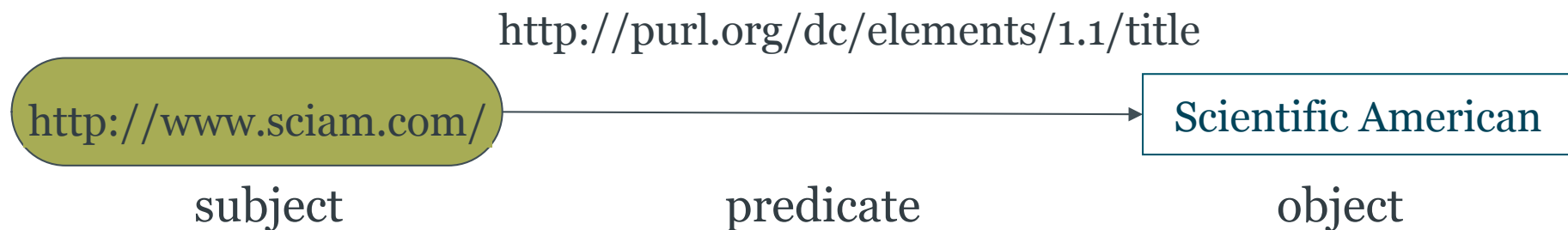
Scientific American

- **Resources**, which represent objects with identity (a web page, a person, a journal)



Example

- Resources are identified by URIs
- Property labels are also identified by URIs, and are drawn from a vocabulary or ontology



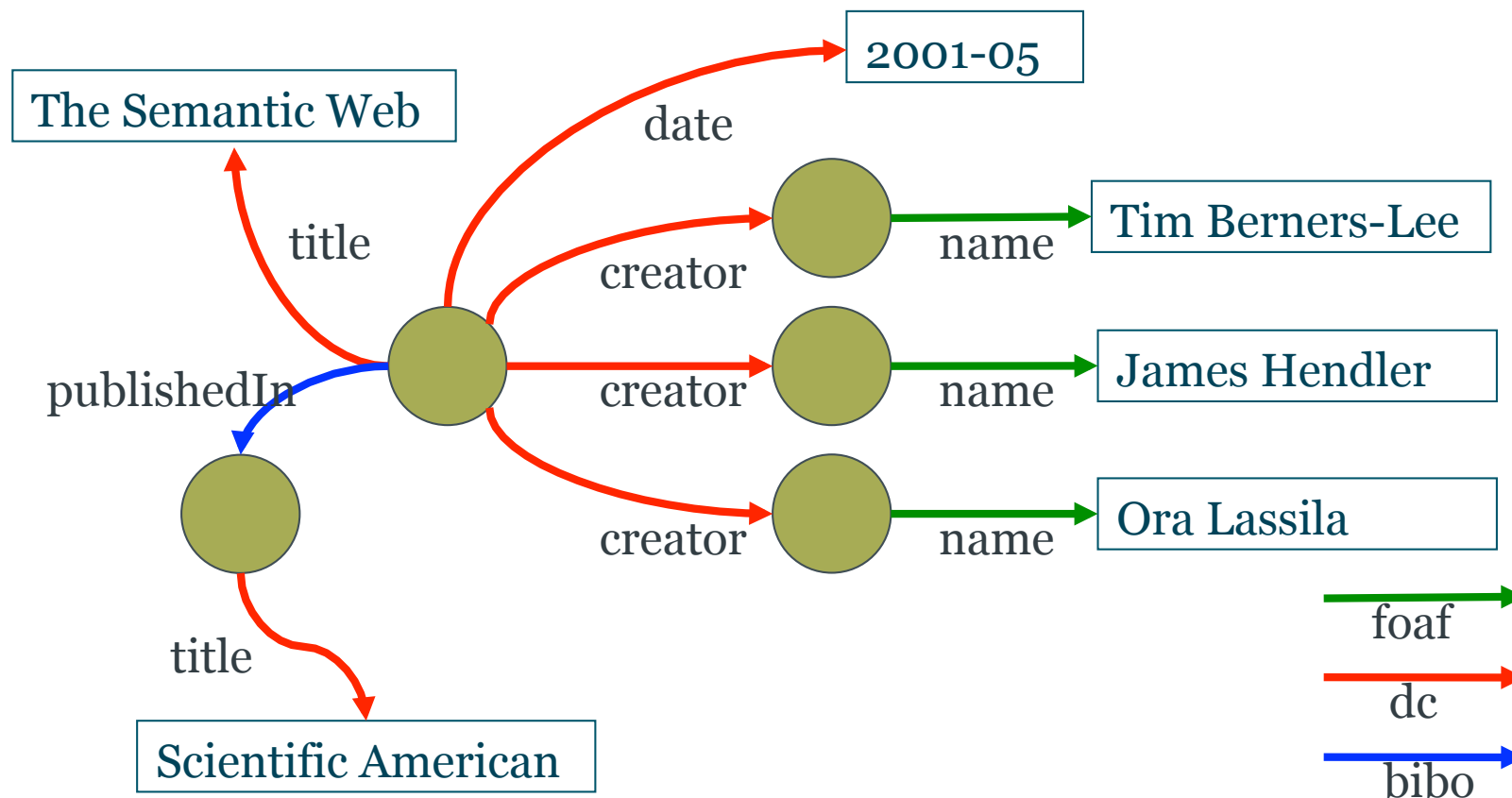
Blank nodes (bNodes)

- Sometimes we have resources which we do not wish to identify with a URI
- These are **blank nodes** or **anonymous resources**



Mixing Vocabularies

- The triple-based graph model makes it possible to mix terms from different vocabularies in the same graph
- Simplifies the task of information integration





Linked Data

<http://www.flickr.com/photos/reedsturtevant/4288406152/>

Linked Data Principles

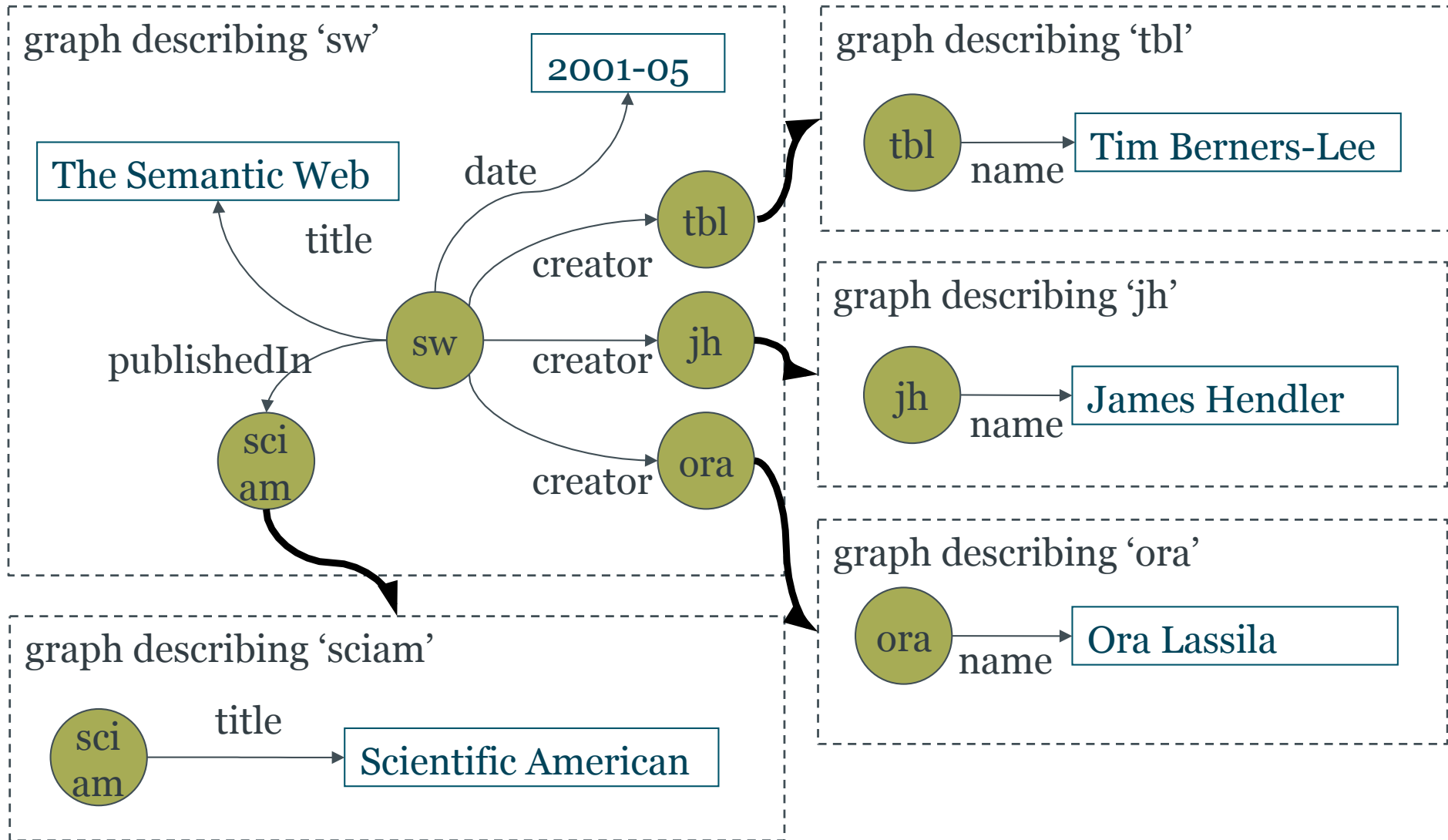
Set of publishing practices for SW data:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information
4. Include links to other URIs. so that they can discover more things

Effectively, putting the hypertext back into the Semantic Web

Simplifies integration between datasets while maintaining loose coupling

Example





In a nutshell

<http://www.flickr.com/photos/arielarielariel/322301228/>

“It’s the integration, stupid!”

(Tim Berners-Lee, speaking at ISWC2003)

YORKS.W.P.
COWGILL
761869

NEWBY HEAD 4¹/₂
HAWES 11¹/₂
INGLETON 14¹/₄

DENT
SED...

Further Reading

- <http://www.w3.org/standards/semanticweb/>
- <http://www.w3.org/standards/techs/rdf>
- <http://www.w3.org/standards/techs/owl>
- <http://www.w3.org/TR/swbp-vocab-pub/>
- <http://linkeddata.org/>