From Web Services to the Semantic Web: Global Data Reuse

David Booth, Ph.D.
W3C Fellow / Hewlett-Packard

Presented at
University of Toronto
University of Waterloo
5 April 2005
(Previously schedule for 10 January 2005)

Slides: http://www.w3.org/2005/Talks/0110-dbooth-semweb/
Or: **tinyurl.com/3vcq9**

Speaker Info

- HP/Software, Technology Office
- Based in Boston
- W3C Fellow (MIT office)
- Working on W3C standards & technologies
- PhD in Computer Science from UCLA
- Many years of programming and OSs

Outline

- **Web Services**
  - SOAP, WSDL, Choreography
  - The Need for Semantics
  - The Problem of "Babelization"
- **The Semantic Web**
  - What Is the Semantic Web?
  - Google and the Semantics of Links
  - Ontologies
  - URIs as Globally Unambiguous Identifiers
  - RDF
- **Web Services and Semantic Web**

Acknowledgements

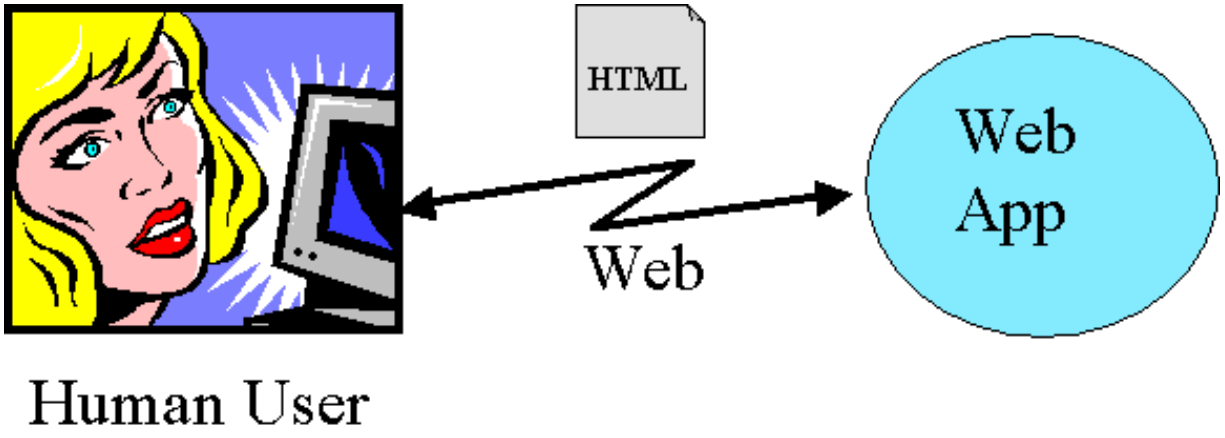
Thanks to:

- **Philippe Le Hégaret**
- **Hugo Haas**
- **Yves Lafon**
- **Tim Berners-Lee**
- **Jim Hendler**
- **Marja-Riitta Koivunen**
- **Eric Miller**
- **Eric Prud'hommeaux**
- **Ralph Swick**

W3C

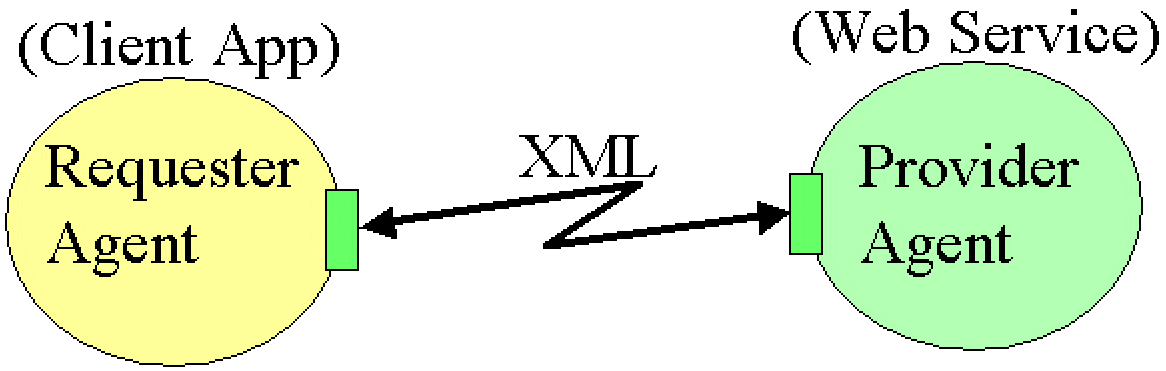
- International consortium
- Mission: "Lead the Web to its full potential"
- Defines Web standards:
  - XML, HTML
  - Digital signatures
  - Web Accessibility Initiative (WAI)
    - See http://www.w3.org/WAI/
  - Web Services
    - XMLP Working Group (for SOAP 1.2)
    - Web Services Description Working Group (for WSDL 1.2)
    - Choreography Working Group
    - Considering work on WS Addressing
    - Workshop planned on WS Constraints and Capabilities
  - Semantic Web
    - RDF
    - Web Ontology Language ("OWL")
  - . . . (and many others)

Traditional Web Application



- Human to Machine interaction

Web Service



- For machine-to-machine interaction
- Client application interacts with Web Service application
- Client and Service exchange XML

Existing protocols were not XML-oriented.

How should messages be packaged?

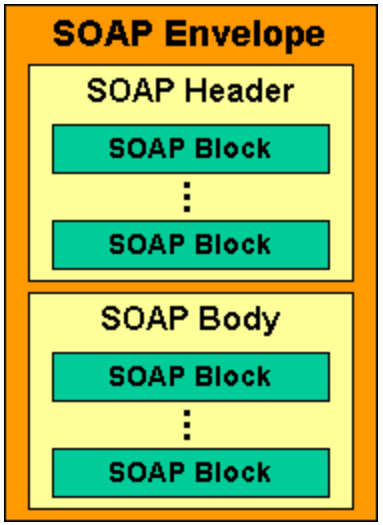
SOAP

- Framework for representing XML messages
- Layered on top of transmission protocols (**HTTP, etc**.)
- SOAP 1.1 produced by individual companies
- SOAP 1.2 has been standardized at W3C

SOAP Structure

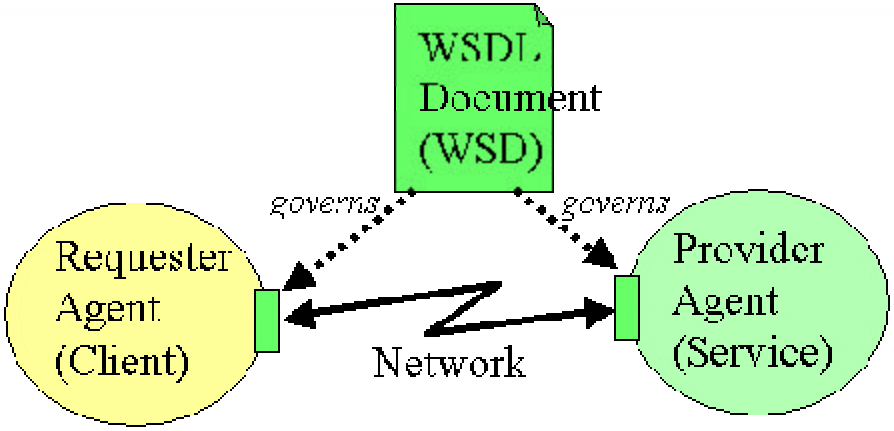
SOAP message is in an "envelope":

```
<Envelope>
  <Header>...</Header>*
  <Body>...</Body>
</Envelope>
```



How can service and client agree on message (body) signatures and transport details?

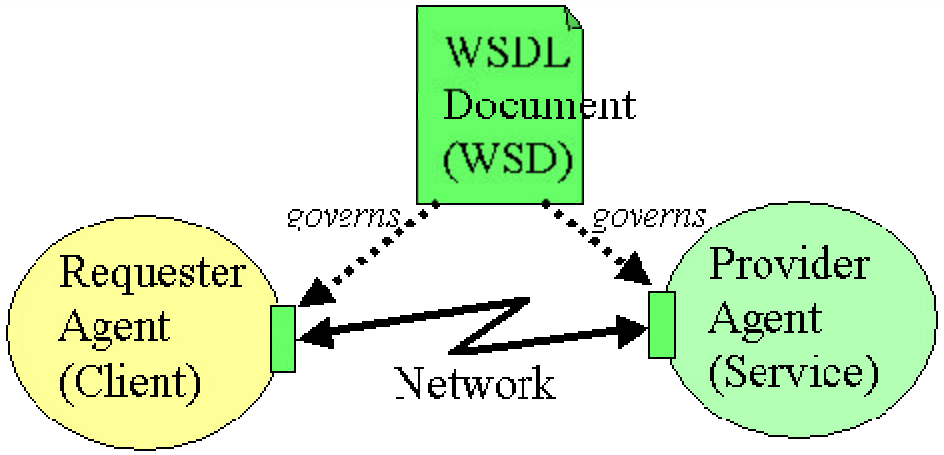
Web Service Description



- **Machine-processable document**
- **Written in Web Service Description Language (WSDL)**
- **Specifies syntax and mechanics of message exchange**
  - Message formats, Data types, Protocols, etc.

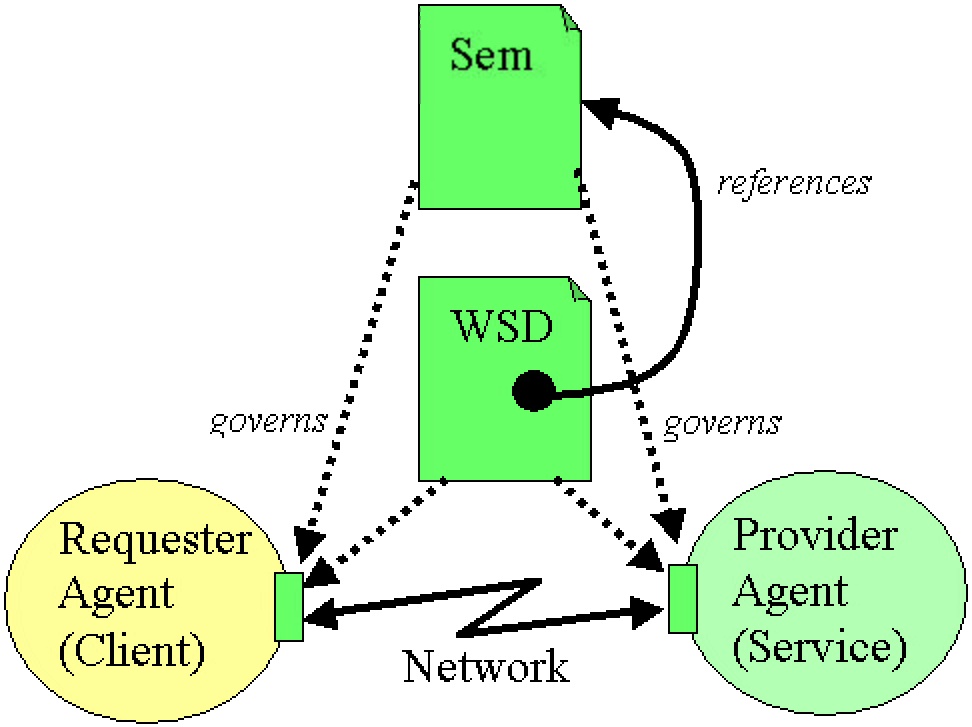
But what will the messages mean?

The Need for Semantics



- **WSDL only defines syntactic-level interface**
- **Client and Service must also agree on <u>semantics</u>**
  - "Semantics" = "meaning"
  - Can be oral or written (preferably)
  - Can be human-oriented (e.g., English) or machine-processable (e.g., RDF)

Referencing Semantics in WSDL

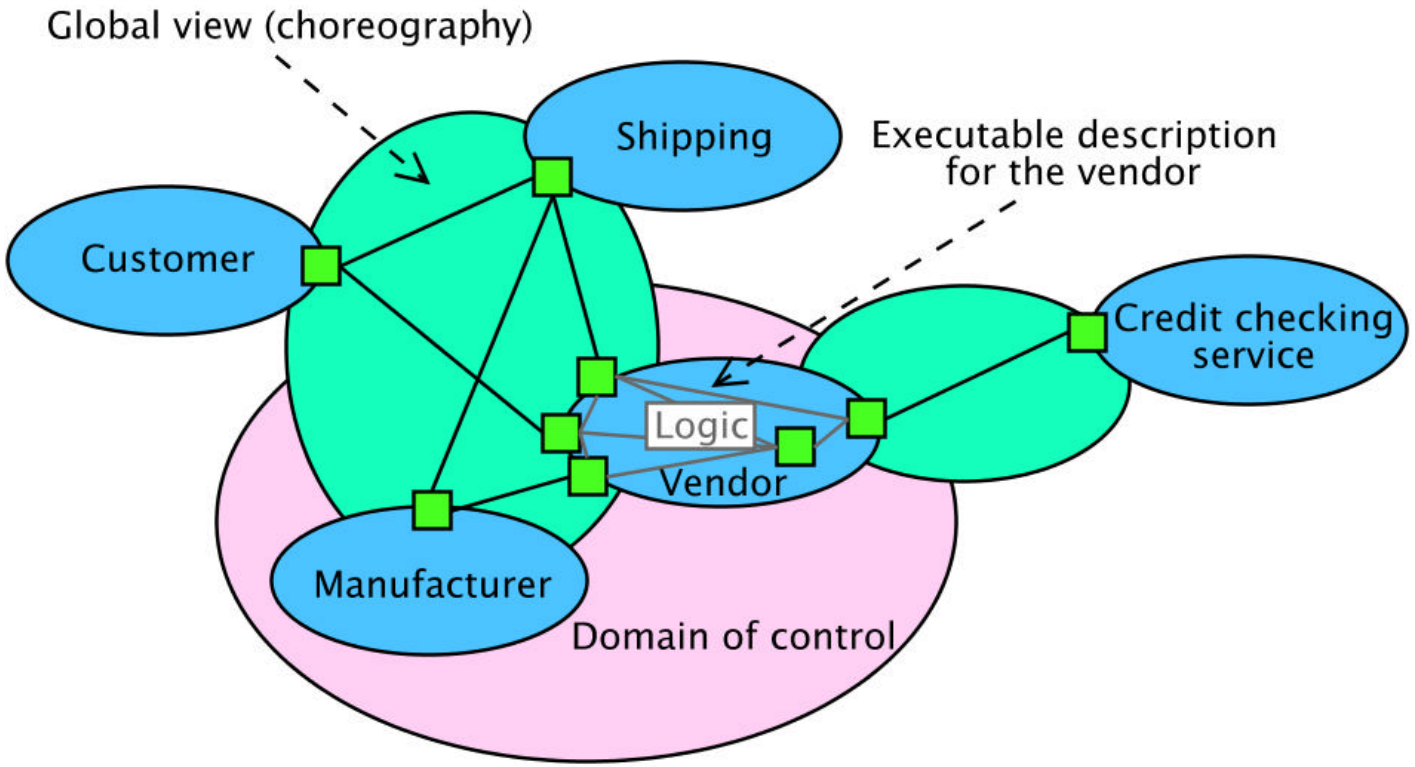


- **Suggestion: WSD should point to document describing semantics**

Web Service Choreography

- WSD describes:
  - Single Web service
  - Simple interactions (e.g., Request-Response)
- "Choreography" can describe:
  - More complex interactions
  - Compositions of services
- Machine processable

Web Service Choreography



- **"Orchestration": Single party (conductor) directly all activities**
- **"Choreography": Parties follow document. No single party directs.**

Problem: "Babelization"

What will happen if Web services become popular?

- Each WSD defines a "language" for interacting
- Proliferation of "languages" / terms
- "Babelization"

  <bar:publication_author>
  <bar:Author>
  <foo:DocumentCreator>
  <foo:Pub-Creator>
  <foo:PlaneTicket>
  <foo:NonrefundablePlaneTicket>
  . . . .
- Meaning may be same, similar or different

Broader Problem: Difficult to Reuse Data

- Data developed independently
- Administered separately
- "Silos"
- Hard to reuse data across administrative boundaries
  - Format, meanings don't always match

Root Difficulty

- Different terms, same meaning
  - Rule can say <foo:DocumentCreator> is the same as <bar:Author>
- Same term, inconsistent meanings
  - XML namespaces can prevent this (if used)
- Different terms, related meaning (but not identical)
  - How are meanings related?
  - <foo:DocumentCreator> == <bar:Author>?
  - If I understand <foo:PlaneTicket>, what can I infer about <foo:NonrefundablePlaneTicket>?

Underlying Needs

Need machine processable:

- Common vocabularies
- Unambiguous names
- Common data model for expressing information

Same needs as for the Semantic Web!

What Is the Semantic Web?

- Enhancement of the current Web, in which
- Meaning is machine-processable

Goal: A More Useful Web

How?

- Make it easier to *Find, Share,* and *Combine* information
  - I.e. global data reuse
  - Data integration
- Allow machines to automate more (Mostly tedious operations)

"The bane of my existence is doing things that I know the computer could do for me."

-- Dan Connolly, The XML Revolution

The Problem of *Finding* Information

Scenario: Find information on specific person, Mary Cook

Problem:

- Many extraneous hits

The Problem of *Sharing* Information

Scenario: Web page shows list of meeting attendees
Want to add to my address book

Problem:

- Address book doesn't understand Web page
- Must copy-and-paste manually

The Problem of *Combining* Information

Scenario:

- Web site A has product reviews
- Web site B has product prices
- Want to *combine* reviews and prices

Problem: Must do it manually

The Lack of Machine-Processable Semantics

Essential Problem:

- Computer doesn't understand meaning ("semantics") of Web pages
- Meaning conveyed by:
  - Human language (e.g., English, German, etc.)
  - Graphics, multimedia
  - Page layout
- Okay for human understanding
- Difficult for machine processing

Analogy: What We Say to Dogs

(Gary Larson cartoon -- local link only)

"Stay out of the garbage! Understand, Ginger?
Stay out of the garbage!"

What Dogs Understand

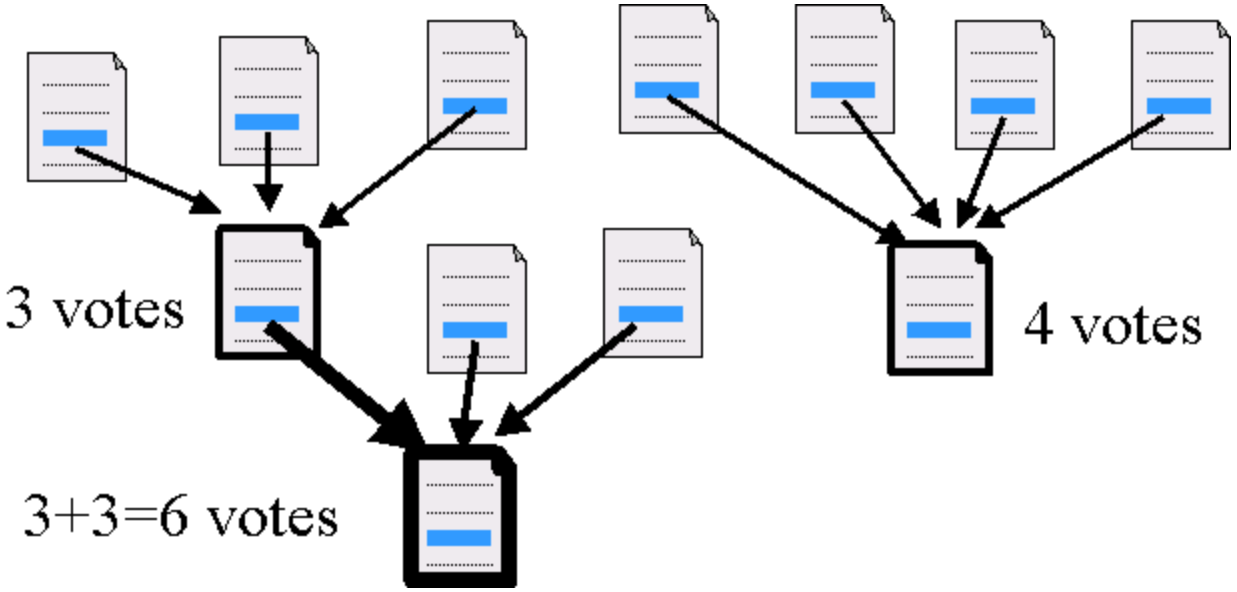
"Blah blah blah blah **GINGER** blah blah blah . . . ."

What Computers Understand

"Blah blah blah blah **<A HREF=...>** blah blah blah . ."

- Need the computer to "understand" more
  - Not human concept of "understanding"
  - Just useful machine processing

How Google Works



- Links into page determine importance
- "Importance" is cumulative (see article)

Exploiting Machine Processable Semantics

Google:

- Links are machine processable
- Links have (Minimal) semantics
  - "This refers to that"
- Amazing results from minimal semantics

What if Web pages had *more* semantics?

(See Article by Bijan Parsia)

Ways to Enable Machine Processing

Two approaches:

- Smarter machines
- Smarter data

Approach 1: Smarter machines

- Teach computers to understand the meaning of Web data
  - Natural language processing
  - Image recognition
  - Etc.
- The Artificial Intelligence (AI) approach
  - WARNING: Hard problems!
  - *Not* the Semantic Web approach

Approach 2: Smarter Data

- Make data easier for machines to understand
  - Express meaning in a machine-processable format
- The Semantic Web approach

How?

Underlying Needs

Need machine processable:

- Common vocabularies (Ontologies)
- Unambiguous names (URIs)
- Common data model for expressing information (RDF)
  - (Especially meta-data)

Ontologies

- Ontology: Formal description of concepts and their relationships
- Example:
  - Definition of "*plane ticket*" and "*non-refundable plane ticket*"
  - A "*non-refundable plane ticket*" is-a-kind-of "*plane ticket*"
- Common, machine-processable "vocabulary"

Example Ontologies

- Dublin Core
  - Defines ~14 basic concepts for documents and publishing:
    - "title", "creator", "subject", "publisher"
- OWL-S
  - Ontology for Web Services
- And many others

How can concepts be unambiguously identified?

URIs as Globally Unambiguous Identifiers

- URI has two different uses:
  1. Unambiguous name for something
  2. Location of a document
- Name is still useful even without accessing a location!
- URIs can be used to identify concepts
  - Especially useful for ontologies & metadata
  - Also useful for other data
- Document at URI can describe the concept

Examples of URIs as Identifiers

- http://example.org/staffids#85740
  - Identifies a particular person (e.g. "John Smith")
- http://purl.org/dc/elements/1.1/creator
  - Defines Dublin Core concept of a document's "creator"

Advice: Use URIs as unambiguous identifiers!

Technical Issue: httpRange-14

Does http://www.example.org/DansCar identify

- **a document?**
- **or the thing described by that document?**

Convention: Use # when identifying the thing:

- **http://www.example.org/DansCar#**
- **http://www.example.org/DansThings#DansCar**

Controversy about whether # is needed

- **See TAG issue httpRange-14**

My suggestion: Use #

Standardizing Ontologies

- Option 1: Standardize on one big ontology
  - Not realistic or practical
- Option 2: Allow arbitrary, conflicting ontologies
  - Not good either
- **Option 3: Allow multiple ontologies, but use URIs to avoid accidental conflicts**
  - Can be merged later

W3C Web Ontology Working Group

- Defining a common language ("OWL") for writing ontologies
- W3C WebOnt Working Group
  - Part of W3C Semantic Web activity
- Based on RDF

What Is RDF?

- "Resource Description Framework"
  - (But think: "**R**elational **D**ata **F**ormat")
  - (Or: "**R**eusable **D**ata **F**ormat")
- W3C Recommendation
  - Part of Semantic Web activity

- Language for making statements about things
- Often used for metadata
  - E.g., Author, Title, Subject
- Has XML syntax
- Simple, universal data representation

RDF Triples

- Info expressed as triples:

  (Subject, Verb, Object)

  Or:

  (Subject, Property, Value)

- Subject, Verb and Object can all be URIs
  - Globally unambiguous

Example RDF Triple

(Not RDF/XML syntax)

http://www.example.org/foo.html *(Subject)*

http://purl.org/dc/elements/1.1/creator *(Verb)*

http://example.org/staffids#85740 *(Object)*

Meaning:

"Web page foo.html was created by John Smith"

RDF for Semantic Web vs. RDF for Data Integration

Two views of RDF:

- Basis for Semantic Web vision
- Techology for solving data integration problems

Web Services and Semantic Web Communities

- Mostly separate communities
- Web Services:
  - More "business" reputation
  - Focus on immediate products
- Semantic Web:
  - More "academic" reputation
  - Focus on long term solution
- Some cross-over
- My opinion: Lots of potential for mutual benefit

RDF and Web Services

- OWL-S (Ontology for Web services):
  - http://www.daml.org/services/owl-s/1.0/
- WSDL 2.0 will have a mapping to RDF. Some work so far:
  - http://www.w3.org/2002/02/21-WSDL-RDF-mapping/
  - http://www-106.ibm.com/developerworks/library/ws-rdf/?dwzone=ws
- "Web Service Composer" (Demo app):
  - http://www.mindswap.org/~evren/composer/
- SWAD-Europe "Semantic Web Services":
  - http://www.itd.clrc.ac.uk/Activity/ACTIVITY=SWAD-Europe;SECTION=1999;
- "Semantic Web Services Initiative":
  - http://www.nextwebgeneration.org/swsi/
- "Semantic Web Enabled Web Services":
  - http://swws.semanticweb.org/
- "WSDL and the Semantic Web" (Bijan Parsia):
  - http://www.mindswap.org/~bparsia/talks/may2003-wsd-wg/Overview-3.html
- More links:
  - http://www.w3.org/2001/11/11-semweb-webservices

Outline

- **Web Services**
  - SOAP, WSDL, Choreography
  - The Need for Semantics
  - The Problem of "Babelization"
- **The Semantic Web**
  - What Is the Semantic Web?
  - Google and the Semantics of Links
  - Ontologies
  - URIs as Globally Unambiguous Identifiers
  - RDF
- **Web Services and Semantic Web**

END

- W3C Mission: *Lead the Web to its full potential*