Voice Search and the Semantic Web

Jim Larson

Larson Technical Services

www.larson-tech.com

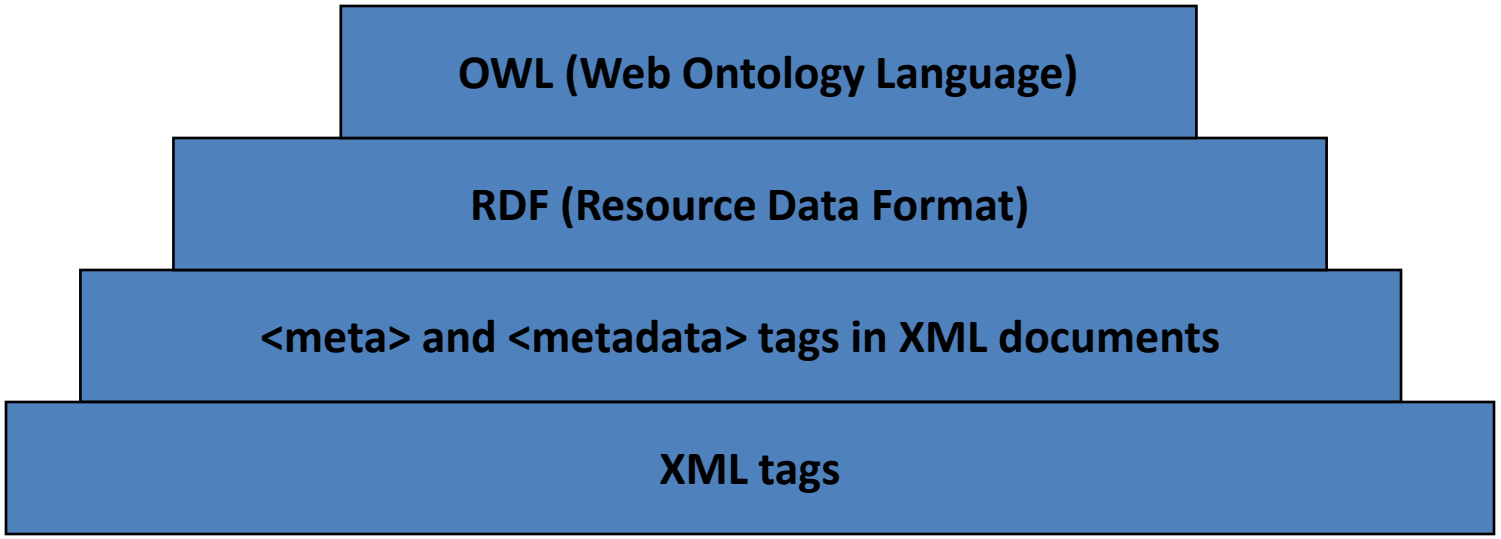
Structured and Unstructured Data

- Structured data
  - Relational databases
  - Lists (telephone numbers, calendars, etc.)
  - Header information in e-mail messages
- Unstructured data
  - Word document
  - E-mail message text
  - HTML files

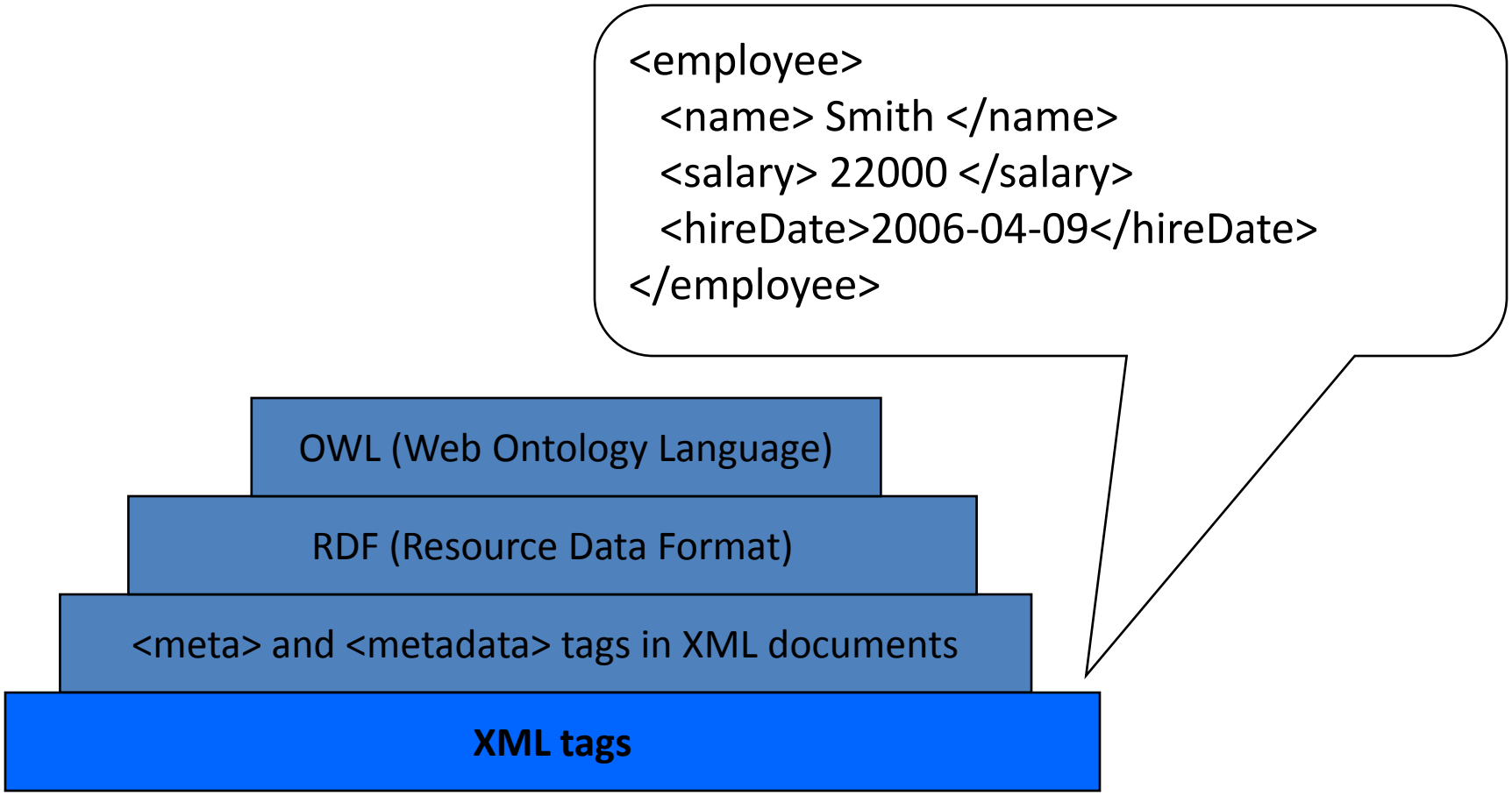
Two approaches for searching unstructured data

- Brute Force
  - Scan entire file for words matching a pattern
  - Time consuming, semantic misses
- Create metadata ahead of time that facilitates searching
  - Examples of “added structure” from real life
    - Table of contents at the front of a book
    - Index at the back of a book
    - Directory of the day’s events posted in a hotel lobby
    - Road signs and billboards

W3C XML-based Metadata



XML Tag

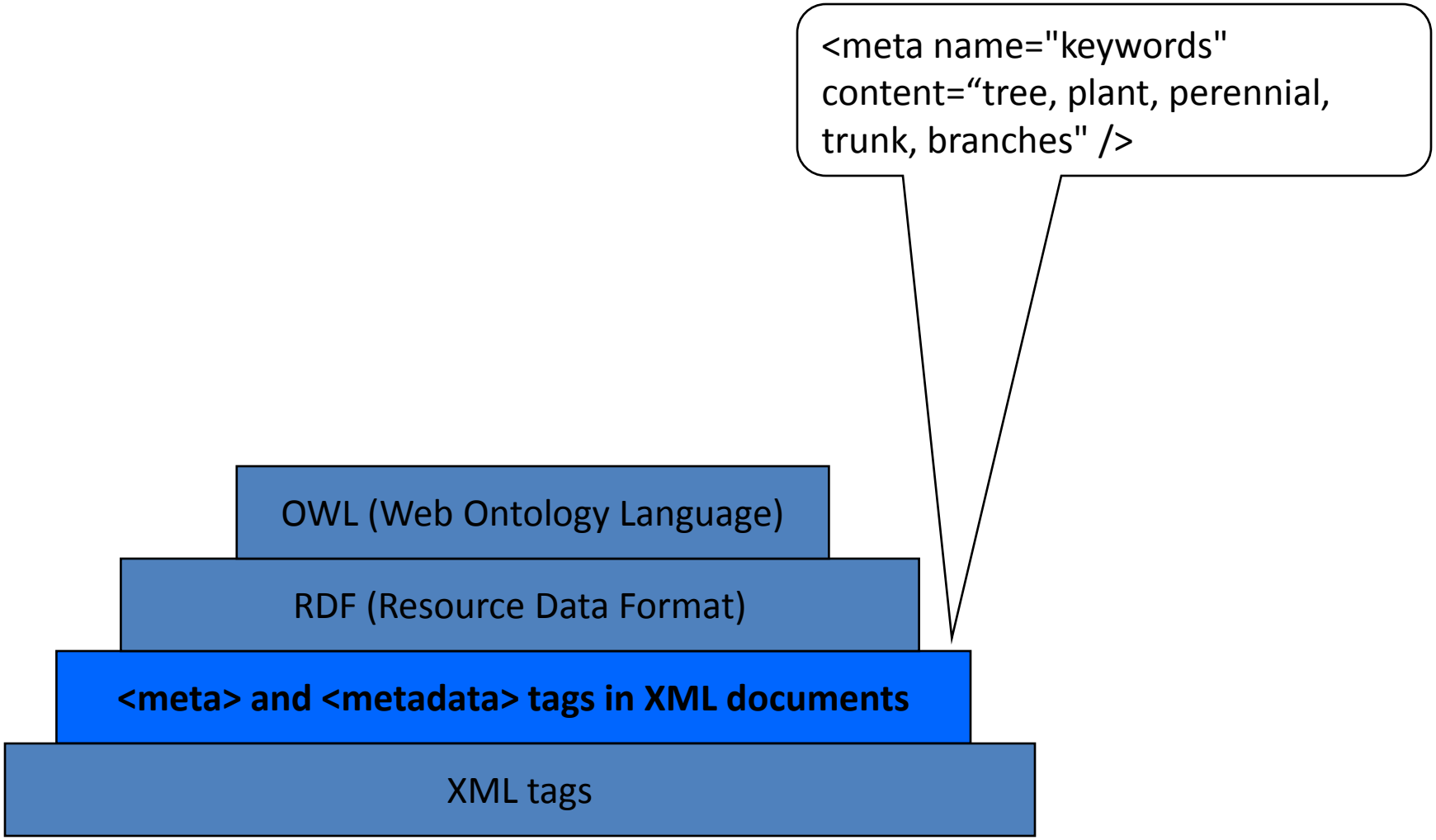


Searching Using XML Tags

- Query:

  Display employees with name = “Smith”
- Search each document looking for the <name> tag with value “Smith”
- To accelerate search, build an index on “name”

<meta> Tag



Searching using <meta> tags

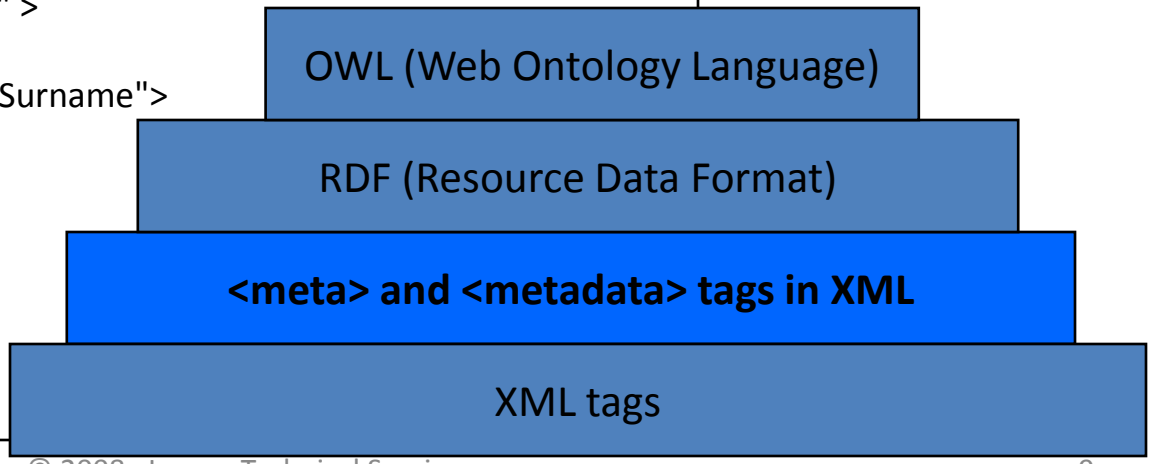
- Query:

  Display employees with keyword = “tree”
- Instead of searching every word in every document, search only the <meta> tags
- To accelerate search, build an index on meta tag attributes

<metadata> Tags with “Dublin Core” Annotations

```
<metadata>
  <rdf:RDF
     xmlns:rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:rdfs = "http://www.w3.org/TR/1999/PR-rdf-schema-19990303#"
     xmlns:dc = "http://purl.org/metadata/dublin_core#">
 <rdf:Description about="http://www.example.com/meta.vxml"
     dc:Title="Directory Enquiry Service"
     dc:Description="Directory Enquiry Service for London in VoiceXML"
     dc:Publisher="W3C"
     dc:Language="en"
     dc:Date="2002-02-12"
     dc:Rights="Copyright 2002 John Smith"
     dc:Format="application/voicexml+xml" >
     <dc:Creator>
       <rdf:Seq ID="CreatorsAlphabeticalBySurname">
         <rdf:li>Jackie Crystal</rdf:li>
         <rdf:li>William Lee</rdf:li>
       </rdf:Seq>
     </dc:Creator>
  </rdf:Description>
 </rdf:RDF>
</metadata>
```

- Used to search and query



Searching using <metadata> tags

- Query:

  Display documents with publisher = “W3C” and creater = “William Lee”

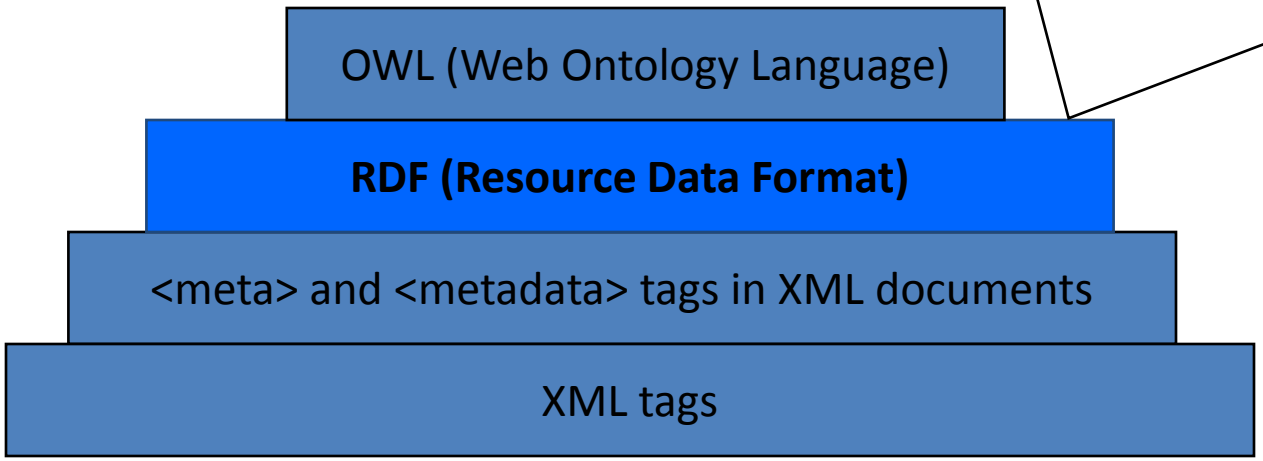
- Use standardized names of Dublin Core
- Instead of searching every word in every document, search only the <metadata> tags
- To accelerate search, build an index on standard names within the metadata tag

What Is RDF?

- Resource Description Framework
- Describes web resources
  - Resource is anything that can have a URI
  - A Property is a Resource that has a name, such as "author" or "homepage"
  - A Property value is the value of a Property, such as "Jan Egil Refsnes" or "http://www.w3schools.com" (note that a property value can be another resource)

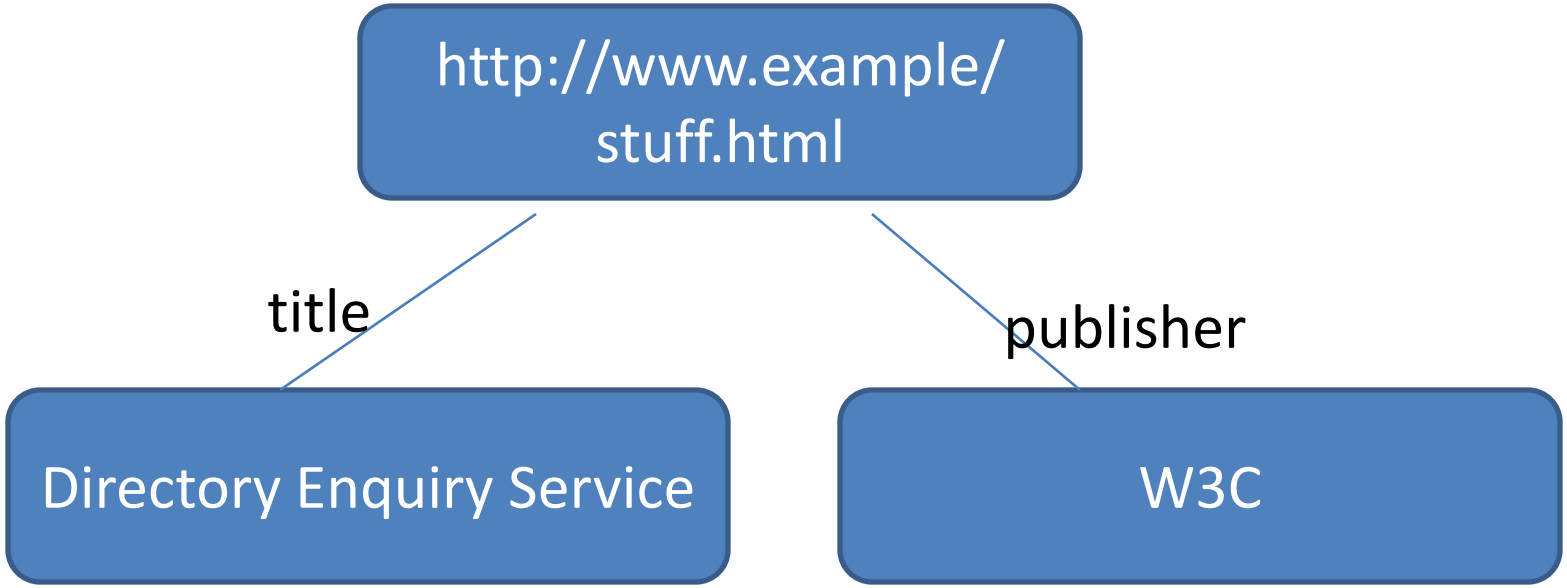
RDF (Resource Description Framework)

```
<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://http://www.example/stuff.html">
              <dc:title> Directory Enquiry Service</dc:title>
              <dc:publisher> W3C </dc:publisher>
        </rdf:Description>
</rdf:RDF>
```

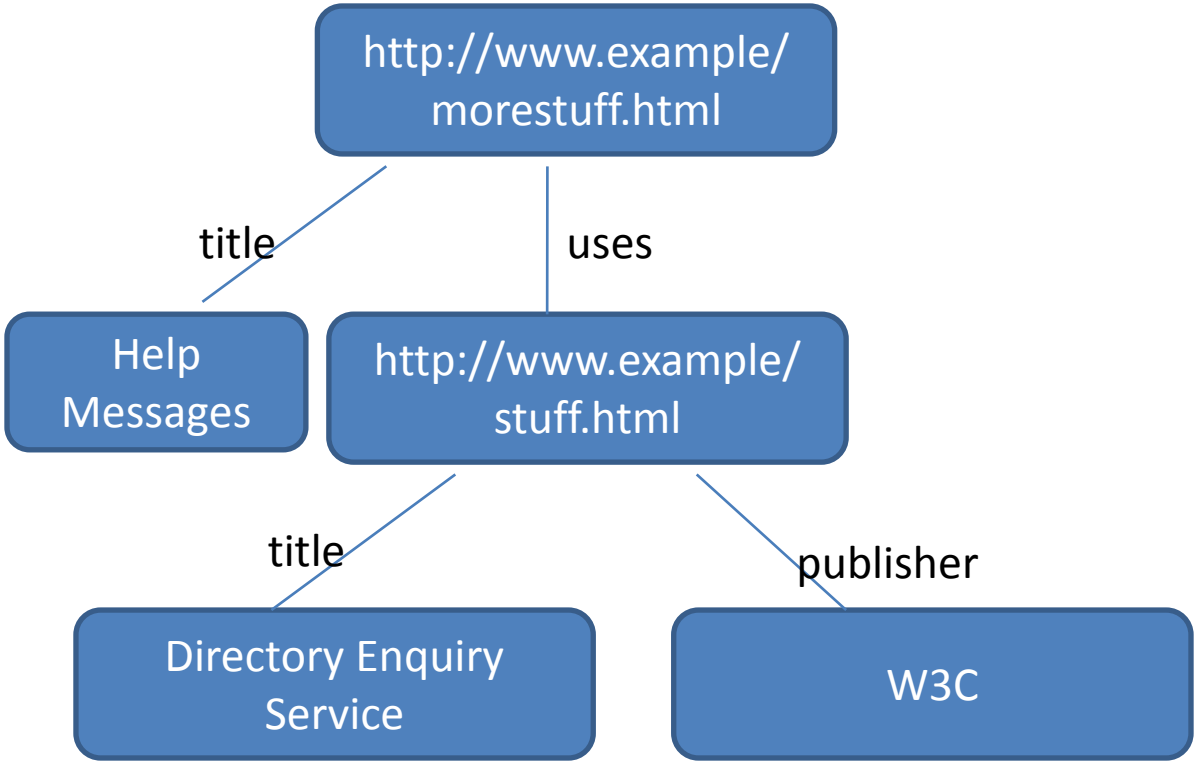


RDF is conceptually a graph

```
<rdf:Description rdf:about="http://http://www.example/stuff.html">
        <dc:title> Directory Enquiry Service</dc:title>
        <dc:publisher> W3C </dc:publisher>
</rdf:Description>
```



Using RDF

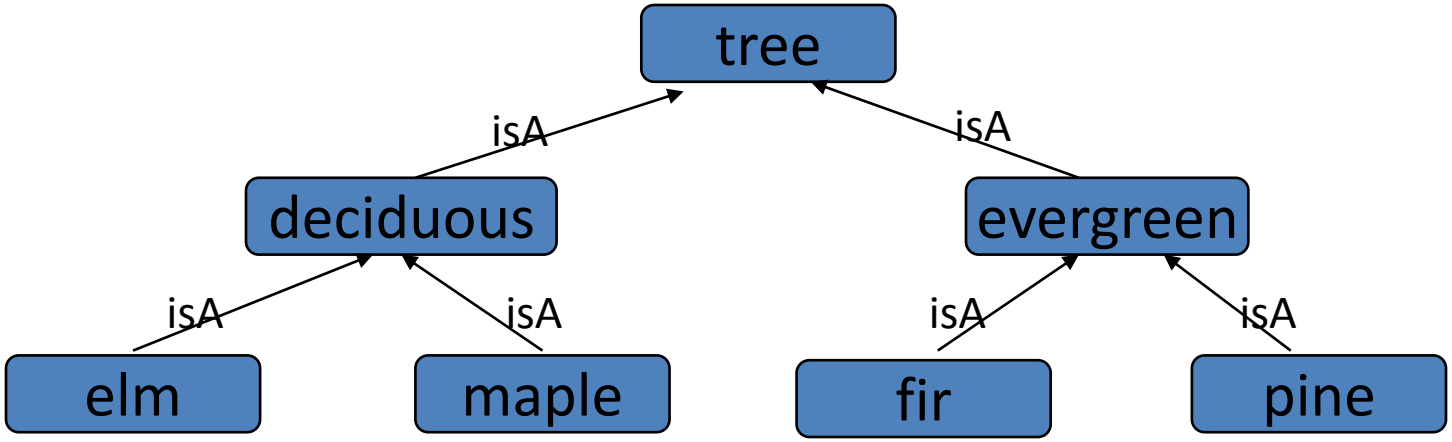


- Query:

  Display “Help Messages” document for “Directory Enquiry Service”
- Search RDF to locate URLs of relevant documents

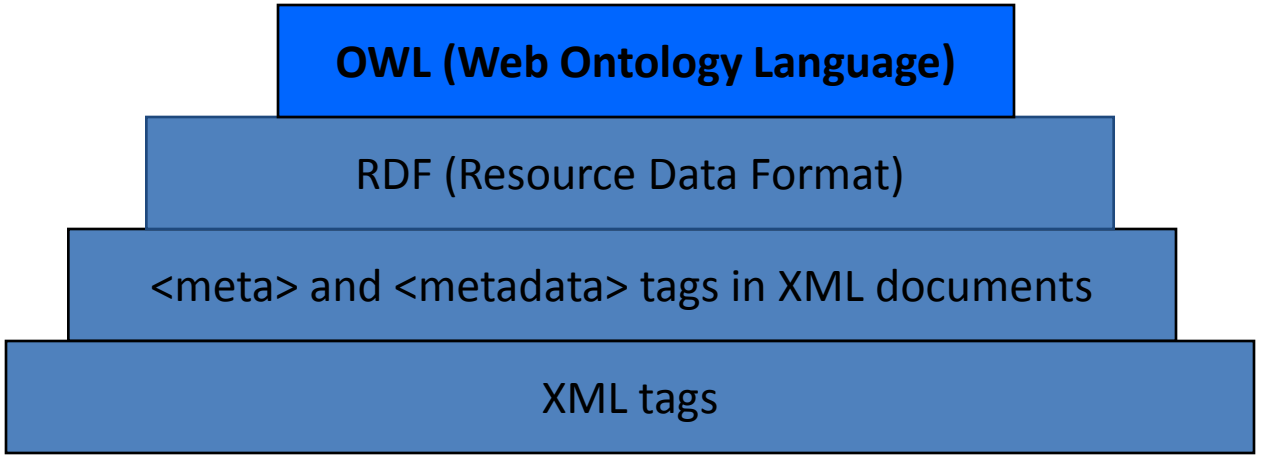
What Is Ontology?

- Ontology is a description of things and their relationships.

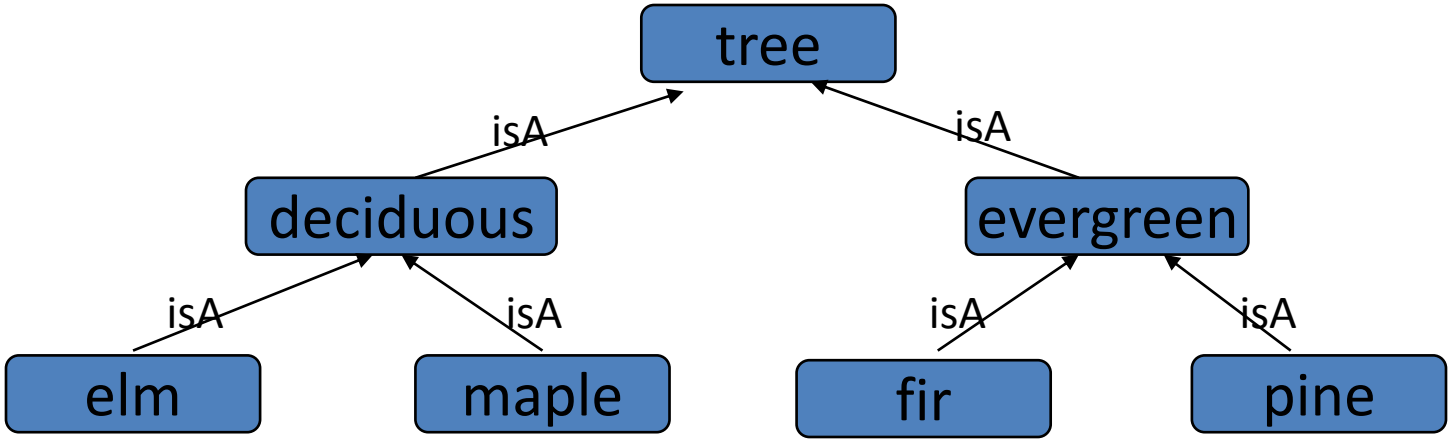


What is OWL?

- OWL is Web Ontology Language
- Stronger language with greater machine interpretability than RDF.
- Larger vocabulary and stronger syntax than RDF.

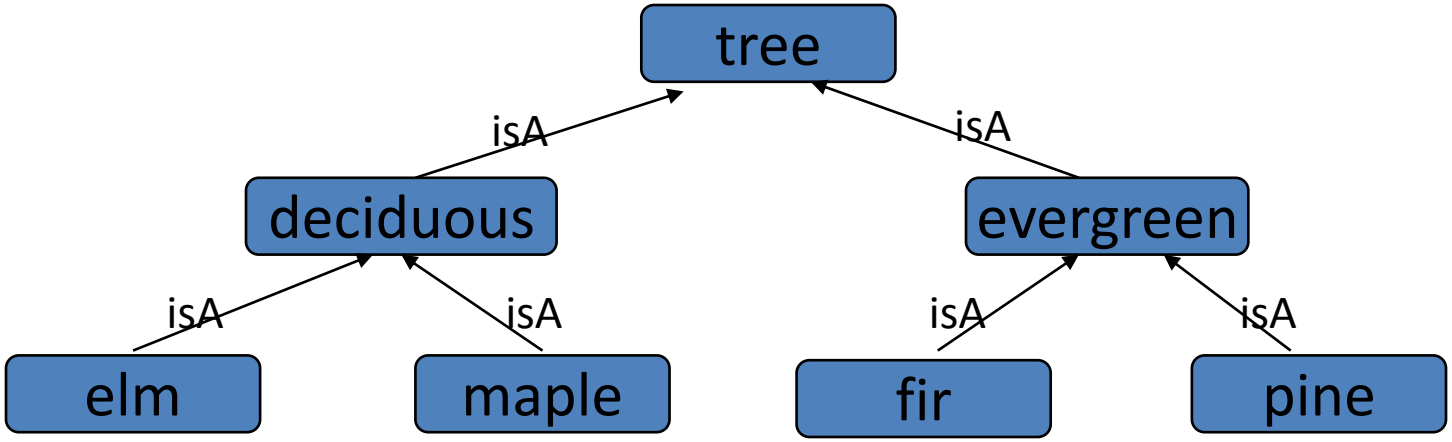


OWL and query modification



- If *plant_type = "evergreen"*
  - Returns too few responses, replace *"evergreen"* by *"tree"*
  - Returns too many responses, replace *"evergreen"* by *"pine"* or *"fir"*

OWL and reasoning



Display documents containing “evergreen”

=>

Display documents containing “fir”

UNION

Display documents containing “pine”

So what?

- Meta data accelerates searching web sites for specific information
- Meta data helps users construct query
  - User navigate through OWL terms to select query terms
  - Suggest query modifications to user
- Retrieve documents that don’t contain words used in the original query

Constructing meta data

- Reuse existing RDF definitions and OWL onthologies
- Extract attribute value pairs from tags in XML, HTML, SVG, and other XML languages
- Extract attribute value pairs from <meta> and <metadata>
- (Research) Automatic extraction of meta data from text

Semantic Web Tools

http://esw.w3.org/topic/SemanticWebTools

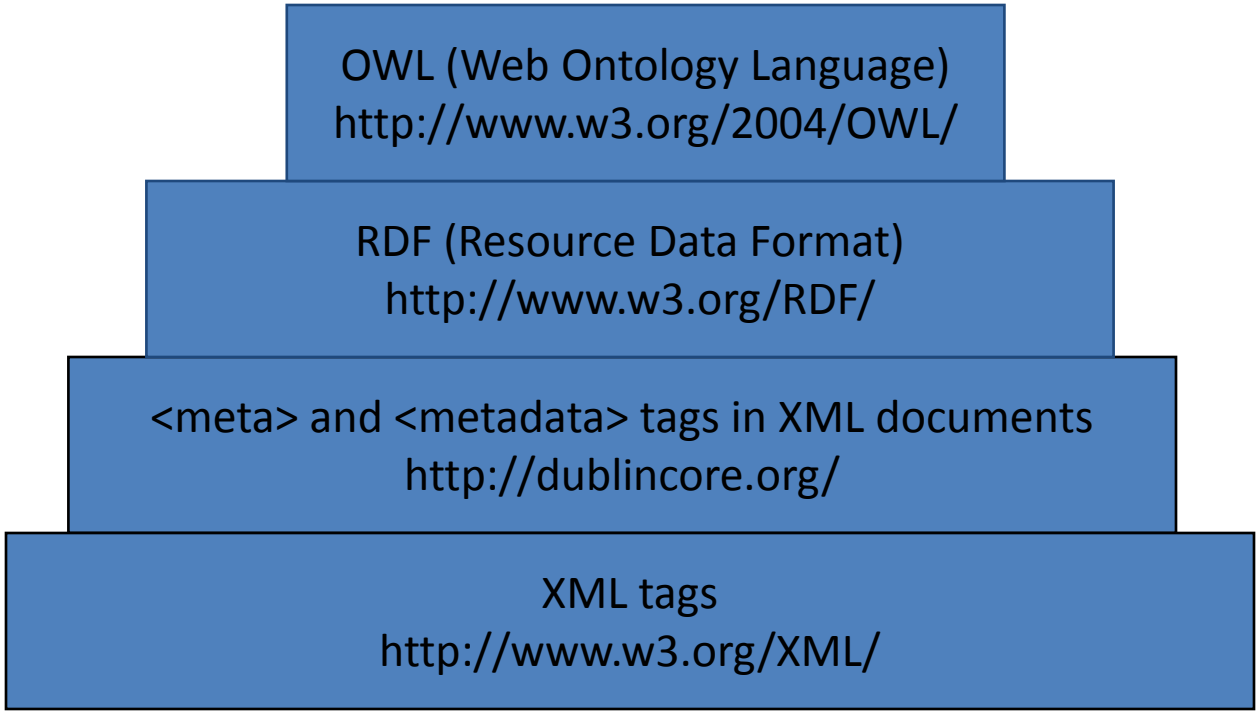
- General Development Environments, Editors, Content Management Systems, ...
- Reasoners and related tools
- RDF Generators
- On-line Validators
- Semantic Web Browsers
- Search Engines
- Tools for Converting Classifications and Other KOS into RDF-S or OWL Ontologies
- Tagging and Semantic Web Vocabularies

OWL tools

- SchemaWeb
  - directory of RDF schemas and OWL ontologies.
- DAML Ontology Library
  - Organizes hundreds of ontologies in a variety of different ways (keyword, organization, submission date, etc.)
- Swoogle
  - is a search engine for Semantic Web documents, including OWL ontologies,

Summary

- Use metadata to reflect the content of your web pages
- Make metadata part of your searching strategy
- Review detailed language specifications



Questions