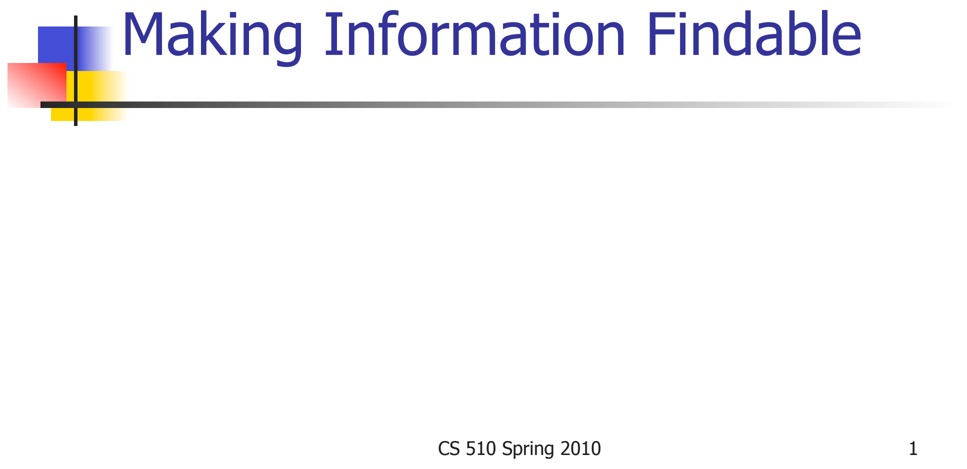
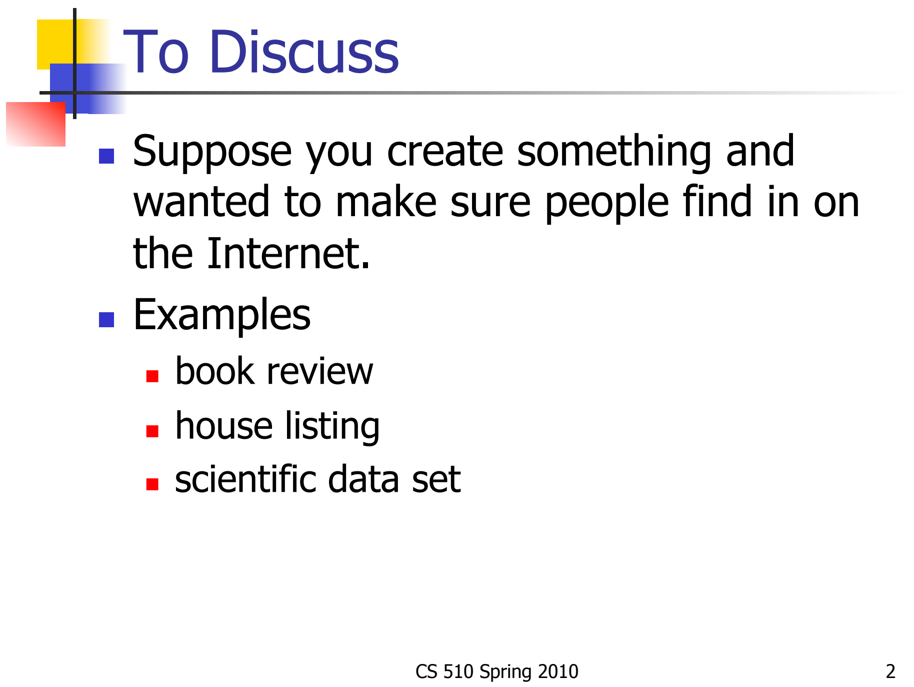


Making Information Findable



To Discuss

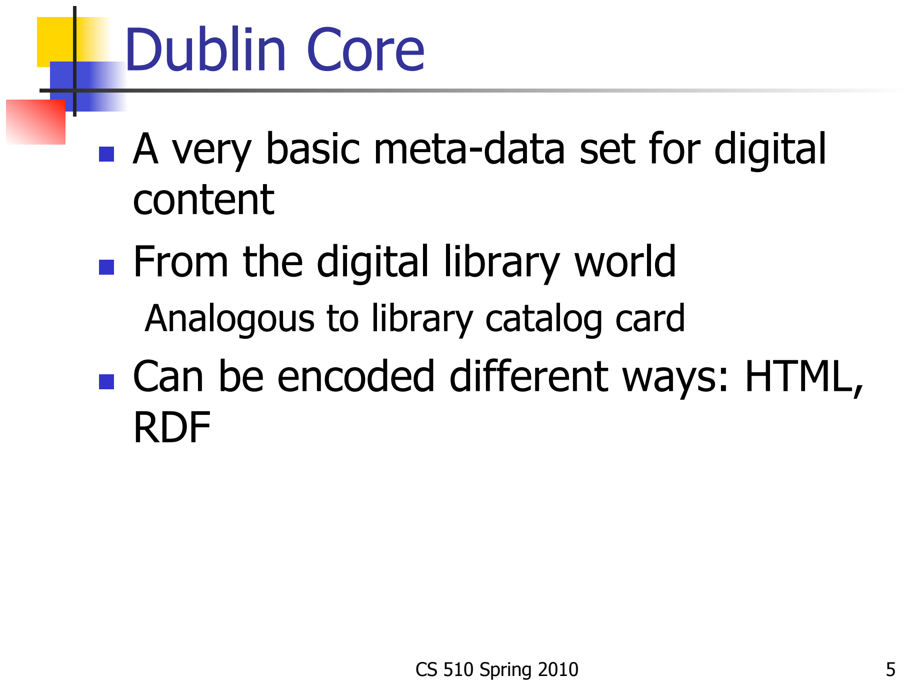
- Suppose you create something and wanted to make sure people find in on the Internet.
- Examples
  - book review
  - house listing
  - scientific data set

I Will Be Talking About

- Meta-data: DC, GILS, RDF
- Mark-up: XML
- Putting ideas together: XMLNews, RSS, Sitemaps

Meta-Data

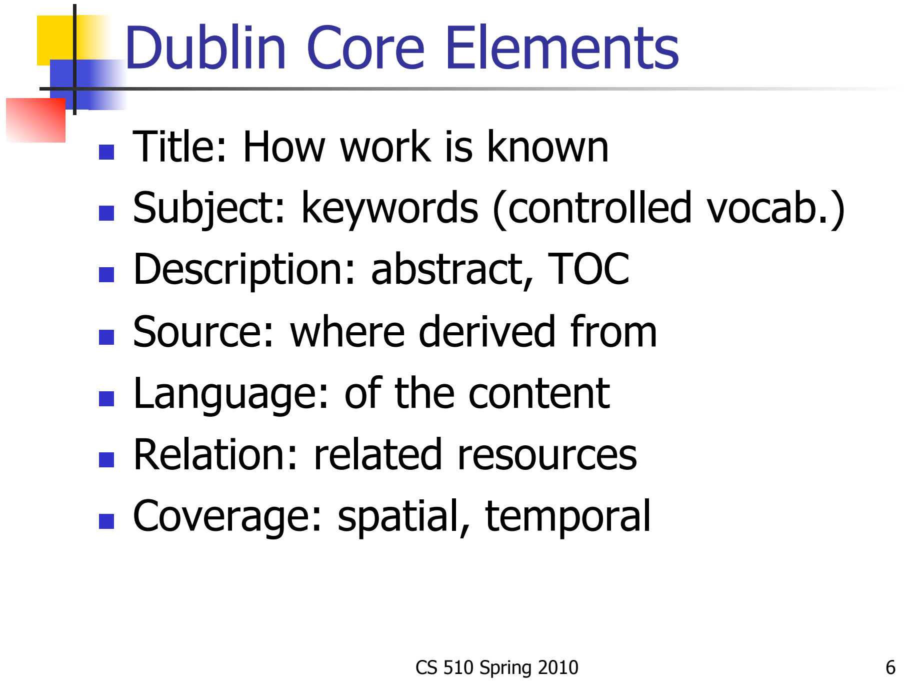
- (Structured) data about data
  - Structural (schema, format)
  - Semantic (definition of elements)
  - Characteristic
    - what it pertains to
    - subject, location
    - content rating
    - ownership
  - Descriptive (how it was created)



Dublin Core

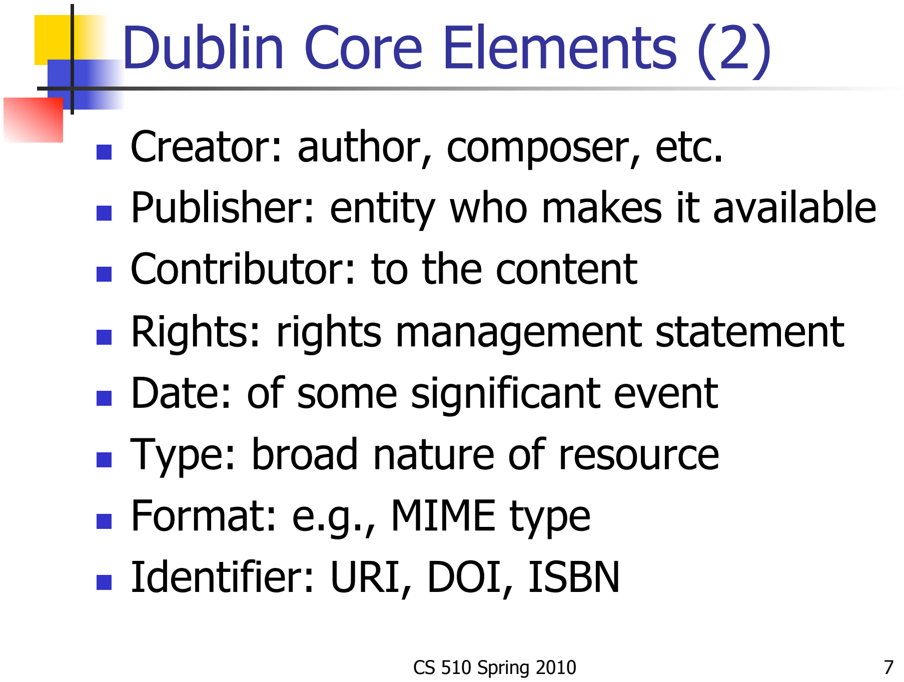
- A very basic meta-data set for digital content
- From the digital library world

  Analogous to library catalog card
- Can be encoded different ways: HTML, RDF



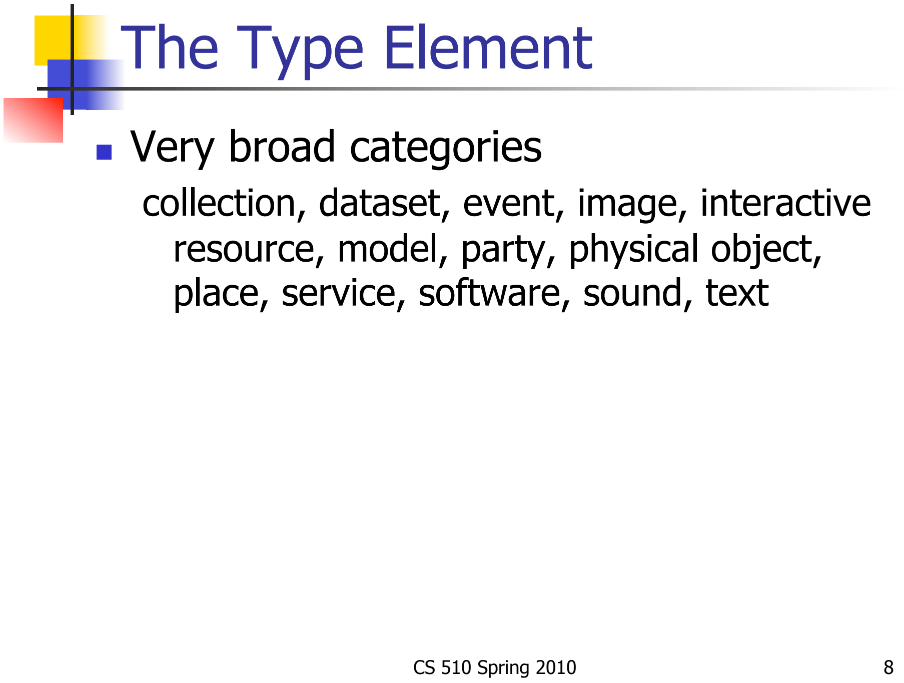
Dublin Core Elements

- Title: How work is known
- Subject: keywords (controlled vocab.)
- Description: abstract, TOC
- Source: where derived from
- Language: of the content
- Relation: related resources
- Coverage: spatial, temporal



Dublin Core Elements (2)

- Creator: author, composer, etc.
- Publisher: entity who makes it available
- Contributor: to the content
- Rights: rights management statement
- Date: of some significant event
- Type: broad nature of resource
- Format: e.g., MIME type
- Identifier: URI, DOI, ISBN



The Type Element

- Very broad categories

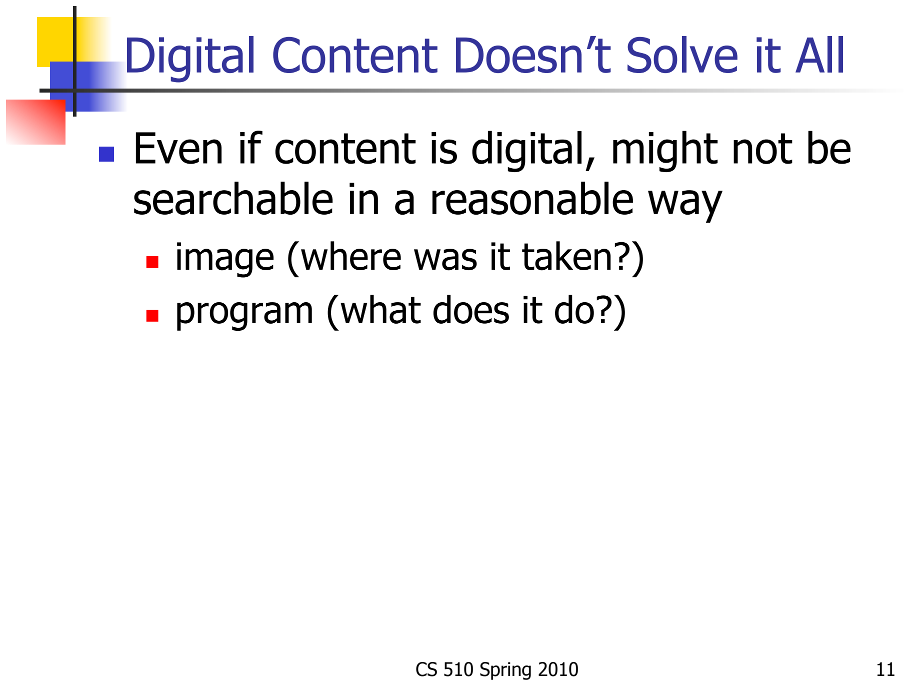
  collection, dataset, event, image, interactive resource, model, party, physical object, place, service, software, sound, text

Qualified DC

- Can further specialize elements
  - `date.created`
  - `date.modified`
- Can give encoding scheme for value, such as a controlled vocabulary
  - `MeSH`

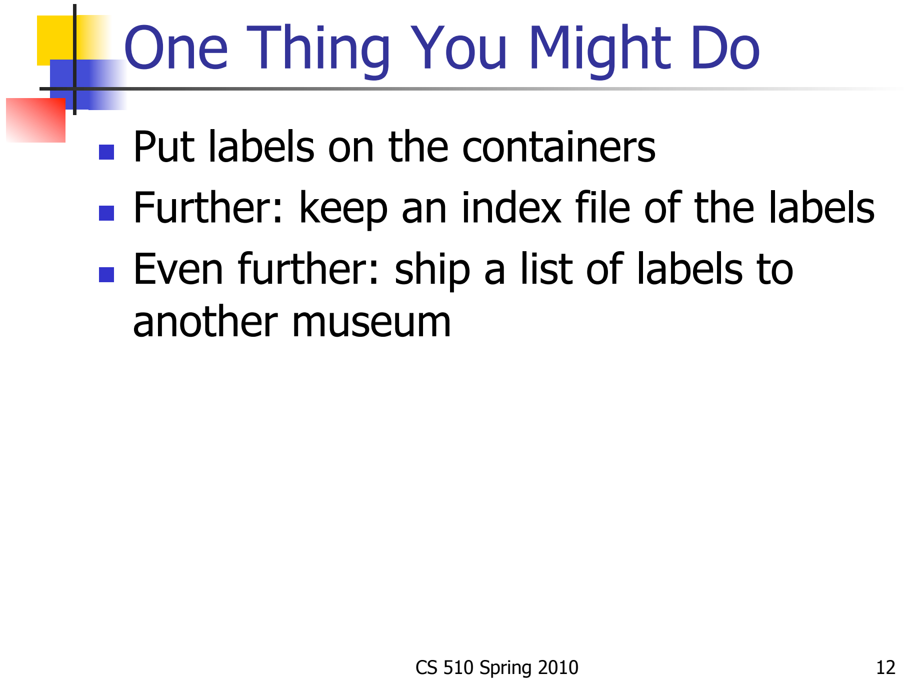
Another Point to Ponder

- Suppose you had a storeroom full of boxes (e.g., in a museum).
- How would you find stuff?
- May be hard to analyze content
  - a mineral sample
  - a piece of pottery
  - a photograph
  - a pickled worm



Digital Content Doesn’t Solve it All

- Even if content is digital, might not be searchable in a reasonable way
  - image (where was it taken?)
  - program (what does it do?)



One Thing You Might Do

- Put labels on the containers
- Further: keep an index file of the labels
- Even further: ship a list of labels to another museum

GILS: Global Information Locator Service

- Standard, structured labels for information containers
  - Could be a document
  - But could be a specimen, a collection, a person, a library, a service, a program
- Set up *clearinghouses* to collect and search these labels
  - Based on Z39.50
  - Clearinghouse doesn’t necessarily have content associated with label

Different from Search Engine

- One level removed from typical web search engine
- Search by label vs. search by content

Locators

- The container label is called a *GILS Locator Record*
- Started in global-climate-change community
- More elements, more domain-specific than DC

GILS Elements

- Examples
  - Coordinates
  - Place name, thesaurus, thesaurus keyword
  - Sources of data
  - Methodology
- About 70 standard elements, plus 100 more from Z39.50

Clearinghouses

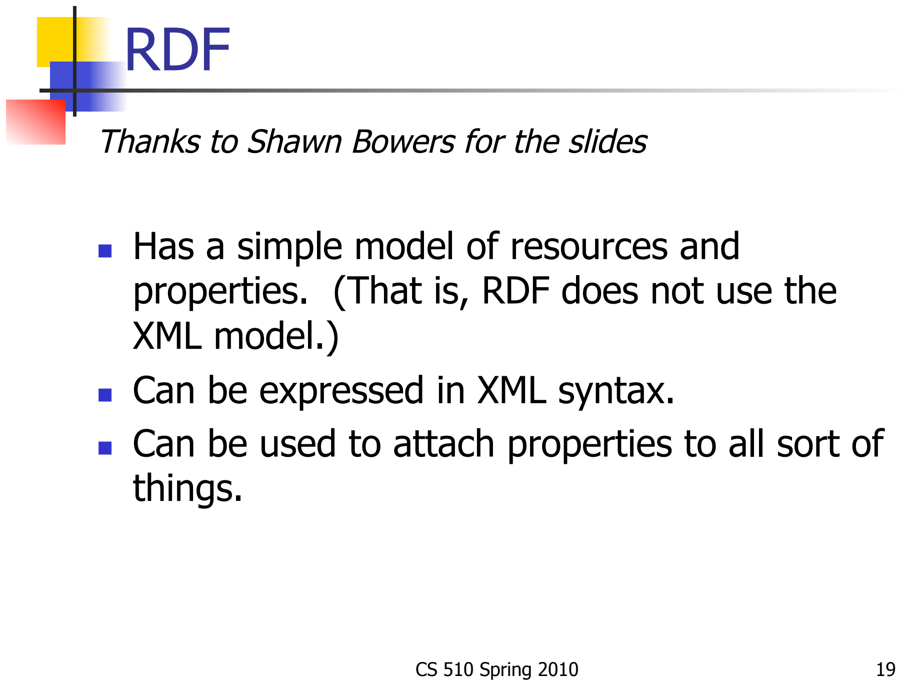
- A number of governments have set up clearinghouses
- Search one clearinghouse, rather than servers (or catalogs) of dozens of agencies.

Resource Description Framework (RDF)

- To add metadata about web pages and other resources
- Example:
  Add “rating” information to web pages…

  then allow parents to filter what their children see based on the rating.

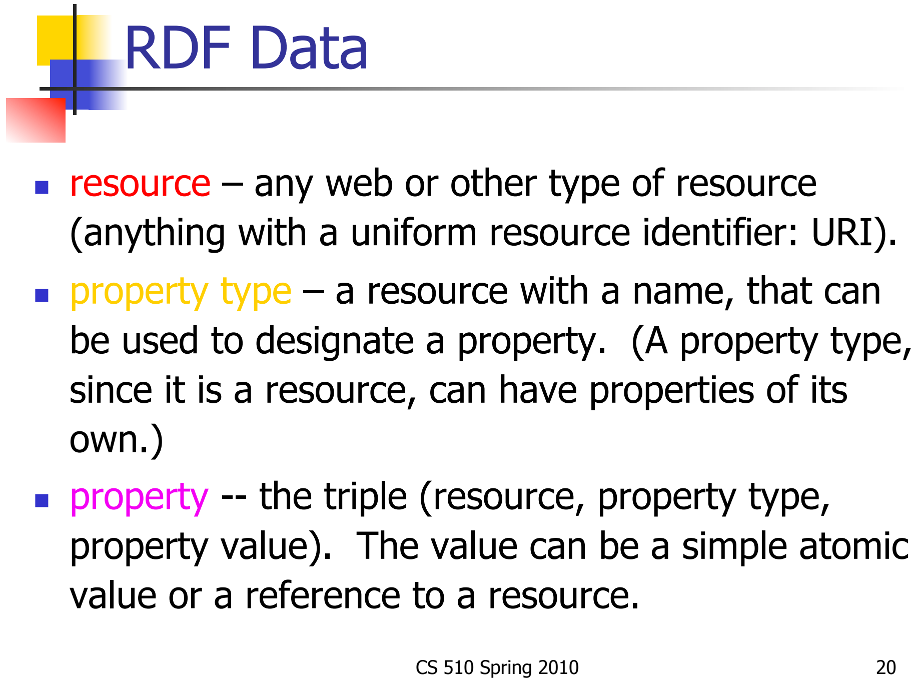
  Original work was part of the Platform for Internet Content Selection (PICS) effort.



RDF

Thanks to Shawn Bowers for the slides

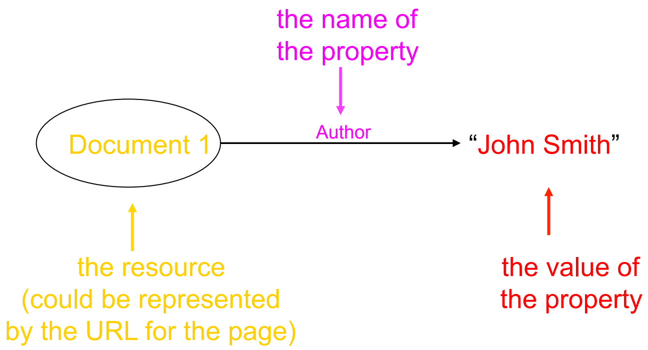
- Has a simple model of resources and properties. (That is, RDF does not use the XML model.)
- Can be expressed in XML syntax.
- Can be used to attach properties to all sort of things.



RDF Data

- resource – any web or other type of resource (anything with a uniform resource identifier: URI).
- property type – a resource with a name, that can be used to designate a property. (A property type, since it is a resource, can have properties of its own.)
- property -- the triple (resource, property type, property value). The value can be a simple atomic value or a reference to a resource.

RDF Example



the name of the property

Document 1

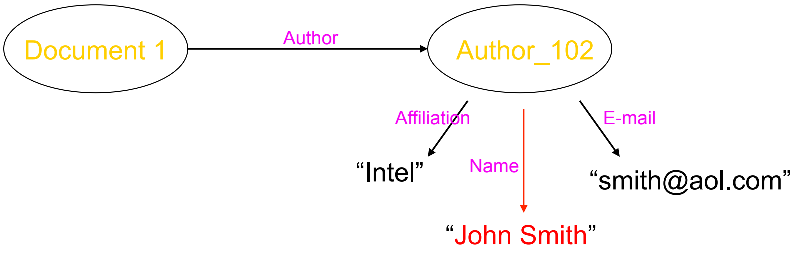
Author

"John Smith"

the resource (could be represented by the URL for the page)

the value of the property

Slightly More Complex RDF Example



Document 1

Author

Author_102

Affiliation

Name

E-mail

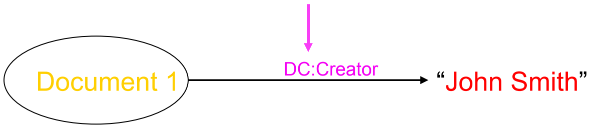
"Intel"

"smith@aol.com"

"John Smith"

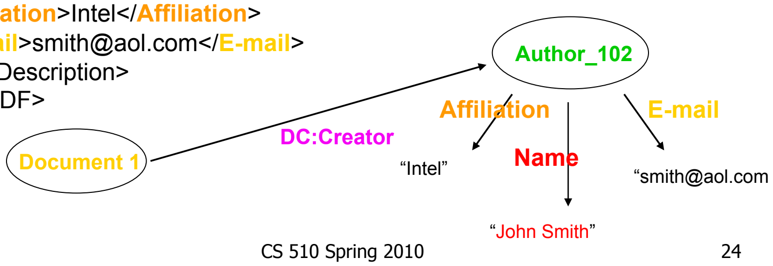
Namespace

The name of the property can
come from a predefined vocabulary
(for property names).
DC stands for the Dublin Core.



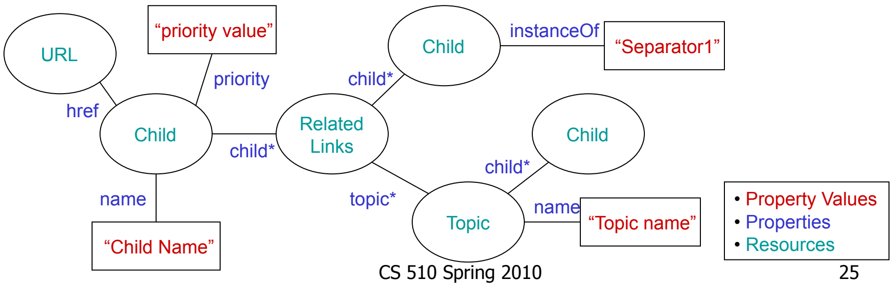
RDF represented in XML syntax

```
<RDF:RDF        xmlns:RDF = “http://www.w3.org/TR/RDF/”
                xmlns:DC = “http://purl.org/dc/elements/1.0/”>

  <RDF:Description about = “….uri of Document 1…”>
    <DC:Creator resource =“Author_102”/>
  </RDF:Description>

  <RDF:Description ID=“Author_102”>
    <Name>John Smith</Name>
    <Affiliation>Intel</Affiliation>
    <E-mail>smith@aol.com</E-mail>
  </RDF:Description>
</RDF:RDF>
```



Alexa 'What's Related' Vocabulary

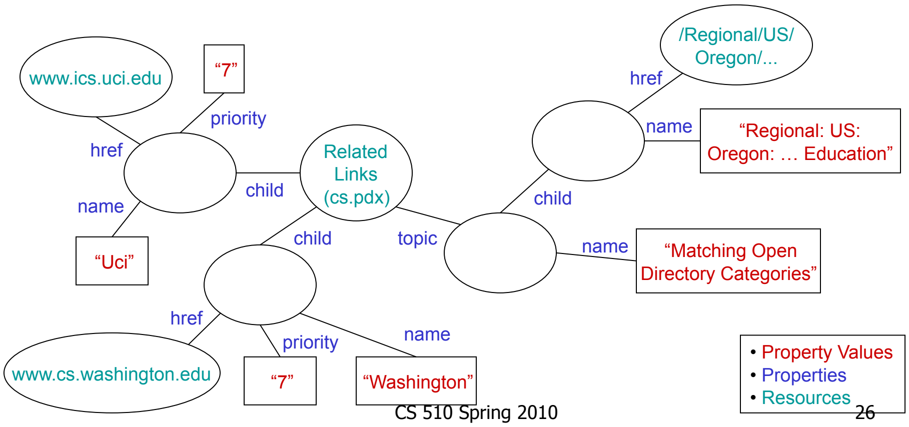
- "What's Related" -- similar to Open Directory Vocabulary
  - The data is created by Alexa Internet, which collects data about the web (see http://www.alexa.com)



CS 510 Spring 2010 25

Alexa 'What's Related' Vocabulary (cont.)

- Example of instantiated schema



CS 510 Spring 2010 26

What’s Related Example RDF/XML

```
<RDF:RDF>
<RelatedLinks>
  <child href=“http://www.ics.uci.edu/” name=“Uci” priority=“7”/>
  <child href=“http://www.cs.washington.edu/” name=“Washington”
   priority=“7”/>
  <-- etc. -->
  <child href=“http://editorial.alexa.com/netscape_editor” name=“Suggest
   related …/”>
  <child instanceOf=“Separator1”/>
  <Topic name=“Matching Open Directory categories”>
    <child href=“http://…/Regional/US/Oregon/Localities/P/Portland/
   Education”
     name=“Regional: US: Oregon: Localities: P: Portland: Education”/>
    <child href=“http://info.netscape.com/…” name=“Submit a site to the
   Open Directory”/>
    <!-- etc. -->
  </Topic>
  <child instanceOf=“Separator1”/>
  <Topic name=“Site info for www.cs.pdx.edu”>
    <child href=“http://info.netscape.com/…” name=“Owner: Computer
   Science Department”/>
    <child href=“http://info.netscape.com/…” name=“Date Established: 08-
   Nov-89”/>
    <!-- etc. -->
  </Topic>
  <!-- etc. -->
  </RelatedLinks>
</RDF>
```

More Expressive Documents

- Explicitly capture more of the semantics and structure of documents
  - more precise searching
  - easier information extraction
- For example, a real estate ad

```
<Residential-Listing>
  <General>
     <Price units=“usdollars”>629000</Price>
     <Structure>
          <Num-beds>5</Num-beds>
          <Num-baths>3.5</Num-baths>
          <Building-area units=“sq-feet”>
               4506</Building-area>
     </Structure>
     <Land-Area units=“acres”>.26</Land-Area>
  </General>
  <Features> … </Features>
  <Contact> … </Contact>
</Residential-Listing>
```

What’s RELML for?

- Trying to get around non-interoperable Multiple Listing Services (MLS) for each local board of Realtors.
- Gives more semantics, not great layout
  - Can use XSL to covert it to HTML
- Could use common format to exchange listings between areas
  - e.g., commercial property in NW

RSS

What does RSS stand for?

- Real Simple Syndication
- RDF Site Summary
- Rich Site Summary

Earlier syndication approaches

- Information Content and Exchange (ICE) Protocol

  Supports catalogs, subscriptions, push/pull
- Channel Definition Format (CDF, MS)
- BackWeb (SW downloads, automatic pull)
- Pointcast (desktop news, push-based initially)

Content vs. Links

- RSS aimed at syndication of links, rather than direct content

  Not so much publishing as an alert service for addition or changes on a web site

- Site provides a feed (channel), which is a list of links

  With meta-data per channel and per links

Basic component

Item or *Entity*: Link + descriptive meta-data

Could put content in description

```
<item>
  <title>Homework 6</title>
  <link>http://www.cs.pdx.edu/
     ~maier/cs510iri/hw6.pdf</link>
  <description>Homework 6 covers
     clustering; it is due 20 February.
  </description>
</item>
```

Pull technology

RSS, in all its versions, is a pull technology

- Site makes feed available at a known address
- Clients use a tool (called an *aggregator)* to poll multiple feeds.

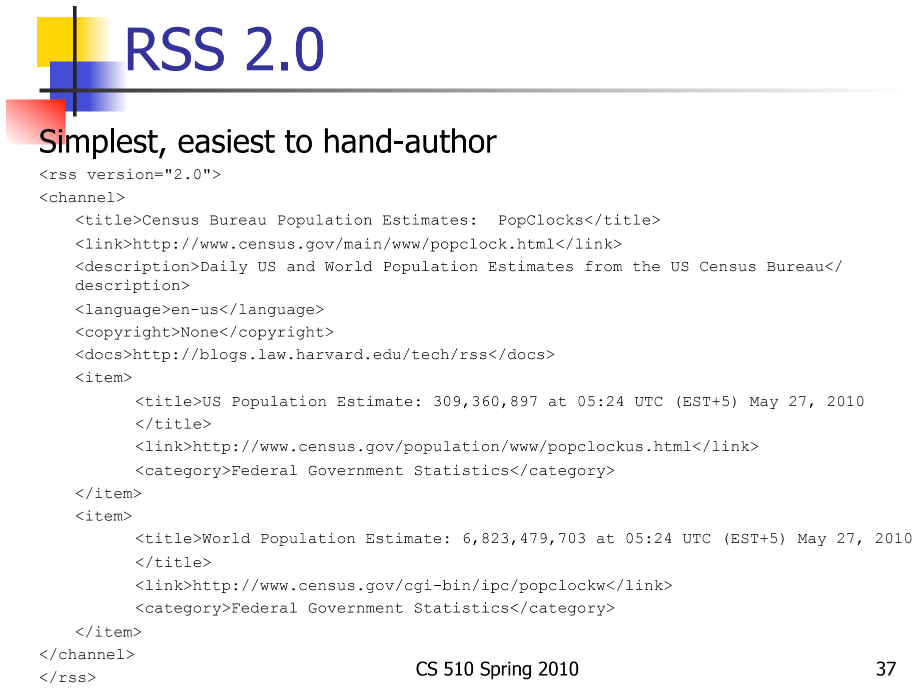
What is in an item?

Usually at least

- Link
- Title
- Description

Other common elements

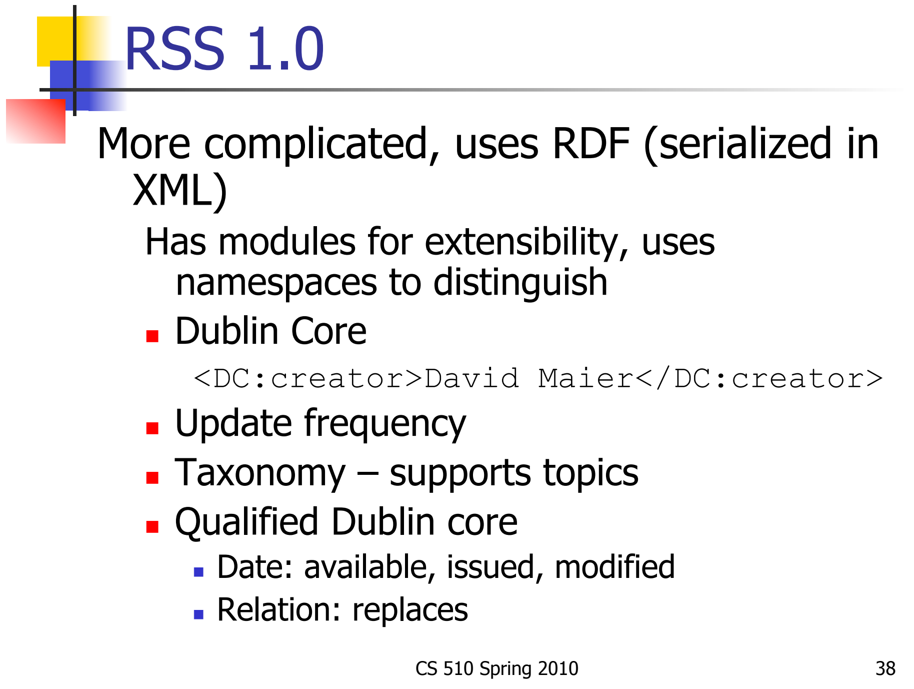
- Unique identifier
- Date
- Author



RSS 2.0

Simplest, easiest to hand-author

```
<rss version="2.0">
<channel>
   <title>Census Bureau Population Estimates:  PopClocks</title>
   <link>http://www.census.gov/main/www/popclock.html</link>
   <description>Daily US and World Population Estimates from the US Census Bureau</
   description>
   <language>en-us</language>
   <copyright>None</copyright>
   <docs>http://blogs.law.harvard.edu/tech/rss</docs>
   <item>
        <title>US Population Estimate: 309,360,897 at 05:24 UTC (EST+5) May 27, 2010
        </title>
        <link>http://www.census.gov/population/www/popclockus.html</link>
        <category>Federal Government Statistics</category>
   </item>
   <item>
        <title>World Population Estimate: 6,823,479,703 at 05:24 UTC (EST+5) May 27, 2010
        </title>
        <link>http://www.census.gov/cgi-bin/ipc/popclockw</link>
        <category>Federal Government Statistics</category>
   </item>
</channel>
</rss>
```



RSS 1.0

More complicated, uses RDF (serialized in XML)

Has modules for extensibility, uses namespaces to distinguish

- Dublin Core

  `<DC:creator>David Maier</DC:creator>`

- Update frequency
- Taxonomy – supports topics
- Qualified Dublin core
  - Date: available, issued, modified
  - Relation: replaces

How are feeds produced?

1. By hand
2. Web-site scraper (local or 3rd party)
3. By-product of content-management system
4. Auto-generated along with site
5. Create feed, construct site from it

Sitemaps

- Introduced by Google
  Now supported by Bing, Yahoo, Ask
- A means to let search engines know about key pages at a site
  - importance, changes
  - can expose pages behind forms
- Represented in XML
  - Can be at `domain/sitemap.xml`
  - `robots.txt` can give location
  - Or can submit to search engine

Sitemap Components

```
<url>
  <loc>http://www.pdx.edu/about-us</loc>
  <lastmod>2010-04-05T23:59:06+00:00
    </lastmod>
  <changefreq>monthly</changefreq>
  <priority>0.8</priority>
</url>
```

Sitemap Index

- Limits to a single sitemap file
  - 50K URLs, 10MB
  - Can only index pages in its directory and sub-directories
  - Same protocol (`http:`, `ftp:`)
- Sitemap Index can point to multiple sitemaps