



Making Information Findable

CS 510 Winter 2007

1



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

CS 510 Winter 2007

2



I Will Be Talking About

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

CS 510 Winter 2007

3



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)

CS 510 Winter 2007

4



Dublin Core

- A very basic meta-data set for digital content
- From the digital library world
- Can be encoded different ways: HTML, RDF

CS 510 Winter 2007

5



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

CS 510 Winter 2007

6



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

CS 510 Winter 2007

7



The Type Element

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

CS 510 Winter 2007

8



Qualified DC

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

CS 510 Winter 2007

9



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

CS 510 Winter 2007

10



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?)

CS 510 Winter 2007

11



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

CS 510 Winter 2007

12



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 (See DL lecture)
 - Clearinghouse doesn't necessarily have content associated with label

CS 510 Winter 2007

13



Different from Search Engine

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

CS 510 Winter 2007

14



Locators

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

CS 510 Winter 2007

15



GILS Elements

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

CS 510 Winter 2007

16



Clearinghouses

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

CS 510 Winter 2007

17



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.

CS 510 Winter 2007

18



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.

CS 510 Winter 2007

19



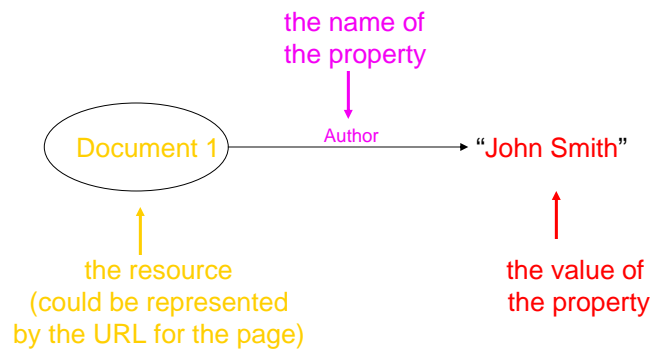
RDF Data

- **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.

CS 510 Winter 2007

20

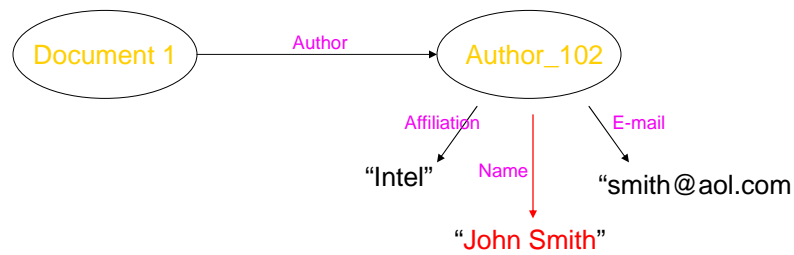
RDF Example



CS 510 Winter 2007

21

Significantly More Complex RDF Example

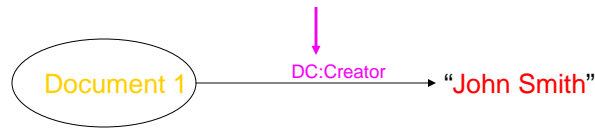


CS 510 Winter 2007

22

RDF Example

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

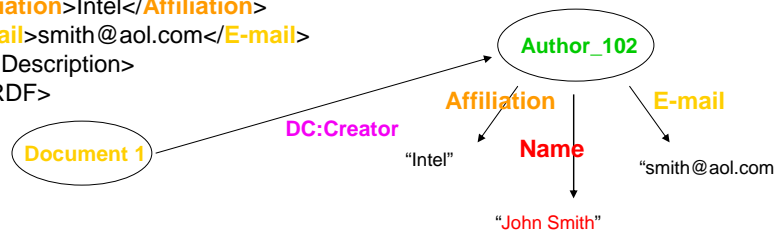


RDF represented in XML syntax

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

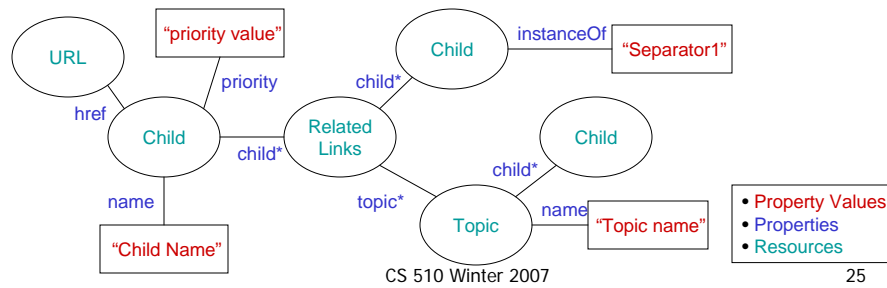
```
<RDF:Description about = "...uri of Document 1...">
  <DC:Creator resource = "Author_102"/>
</RDF:Description>
```

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



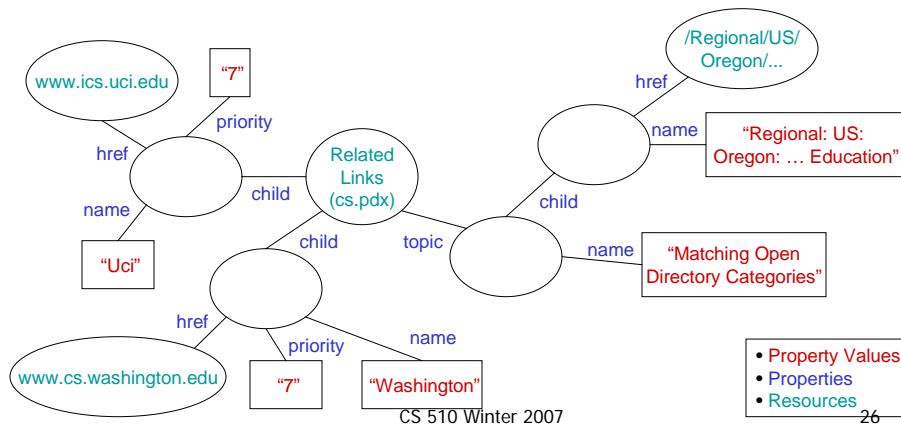
Netscape 'What's Related' Vocabulary

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



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

- Example of instantiated schema





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/B/Beaverton/Education"
      name="Regional: US: Oregon: Localities: B: Beaverton: 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.cse.ogi.edu">
    <child href="http://info.netscape.com/..." name="Owner: Computer
      Science AND Engineering Department"/>
    <child href="http://info.netscape.com/..." name="Date Established: 08-
      Nov-89"/>
    <!-- etc. -->
  </Topic>
  <!-- etc. -->
</RelatedLinks>
</RDF>

```

CS 510 Winter 2007

27




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

CS 510 Winter 2007

28



```

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

CS 510 Winter 2007 29



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 convert it to HTML
- Could use common format to exchange listings between areas
 - e.g., commercial property in NW



XMLNews: Putting Technologies Together

- For news dissemination services
 - MediaXpress
 - iSyndicate
- Mainly aimed at web-site maintainers
 - Want news stories
 - easy to determine if interesting to audience
 - easy to format to match look of site

CS 510 Winter 2007

31



Business Model

- Gather news from different providers
 - Reuters, UPI, Sports Network
- Mark up stories in XML, add meta-data
- Push out to web servers via satellite, Internet delivery service
- Receive at web site
- Tools to parse, format, select content

CS 510 Winter 2007

32



XMLNews-Story

- XML markup for news stories
- Based on NITF: News Industry Text Format
- Head: headline, byline, dateline
- Body: labeled text
 - quote, country, chron, person, org, money
- Tables (HTML-style)
- Media objects: image, audio, video

NITF seems to have switched to XML, too.

CS 510 Winter 2007

33



XMLNews-Meta

- RDF properties on news stories; extends Dublin core
 - more precise searching
 - adds info not in story, e.g., stock symbol

NewsML seems to be subsuming it

CS 510 Winter 2007

34



XMLNews-Meta Elements

- DC: title, creator
- Other descriptive: classification, fixture name
- milestones: release time, expire time
- provenance: supplier, distributor
- subject matter: company name, industry code, event name
- version links