Plan for Today's Lecture(s)

- The Semantic Web
- RDF and other Semantic Web Technologies
- Linked Data
- SemWeb and Linked Data and Libraries

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

15 October 2013
Lecture 14.1 – The Semantic Web

Why the Web Wasn't Born Semantic

- Around 1990, when the Web was being imagined by TBL, SGML (precursor of XML) was increasingly being used to define structured document models for publishing etc.
- TBL consciously chose to create HTML as a specific and simple document language rather than take on the generality, expressive power, and complexity that SGML would have meant (and he was criticized by the "experts" for doing so)
- This tradeoff enabled the Web to take off because it made it vastly easier to create web pages and software for serving and processing them
- But this made the Web work "for eyes only" rather than make it work "for machines", especially those doing business

The Vision of the Semantic Web (1)

- In a classic 2001 paper Sir Tim Berners-Lee says:
- The Web can reach its full potential only if … data can be shared and processed by automated tools as well as by people...
- The Semantic Web will bring structure to the meaningful content of Web pages...
- For the Web to scale, tomorrow's programs must be able to share and process data even when these programs have been designed totally independently.

The Vision of the Semantic Web (2)

- Services and agents can advertise their function by registering in directories
- The service is described in a way that lets other agents discover the function offered and understand how to use it, terms and conditions, invocations…
- Service discovery enables agents to delegate tasks to create the overall "value chain" in which subassemblies of information are passed from one agent to another

The Semantic Web Scenario (1)

- Lucy and Pete, 2 adult children trying to help out Mom with a medical appointment
  - "...Lucy instructed her Semantic Web agent through her handheld Web browser. The agent promptly retrieved information about Mom's prescribed treatment from the doctor's agent, looked up several lists of providers, and checked for the ones in-plan for Mom's insurance within a 20-mile radius of her home and with a rating of excellent or very good on trusted rating services"

The Semantic Web Scenario (2)

... trying to find a match between available appointment times and Pete's and Lucy's busy schedules...

…but Pete didn't like the agent's plan

… He set his own agent to redo the search with stricter preferences about location and time

Making the (Existing) Web More Semantic

- Convert existing Web page content to semantic markup
- Annotate existing Web page content with semantic metadata

Making the (New) Web More Semantic

- Create new web page content with semantic content and semantic metadata
- Create new pages or resources that are designed from the outset to be "meta-pages" that facilitate semantic processing of all the other metadata

Extracting “Semantic” Markup

- if the content is semi-structured (e.g,, a news feed that uses a "story template" for its stories, with mixed content "semantic islands" in the text, NLP and "data mining" techniques can often extract some limited semantics -- like tagging for people, place names, product names - these usually start with "named entity recognition”

Stanford Named Entity Tagger

Demo using text from Robert J. Glushko's Home Page at Berkeley

Bob Glushko is an Adjunct Full Professor at the University of California at Berkeley in the School of Information, where he has been since 2002. He has over thirty years of R&D, consulting, and entrepreneurial experience in information systems and service design, content management, electronic publishing, Internet commerce, and human factors in computing systems. He founded or co-founded four companies, including Veo Systems in 1997, which pioneered the use of XML for electronic business before its 1999 acquisition by Commerce One. Veo's innovations included the Common Business Library (CBL), the first native XML vocabulary for business-to-business transactions, and the Schema for Object-Oriented XML (SOX), the first object-oriented XML schema language. From 1999-2002 he headed Commerce One's XML architecture and technical standards activities and was named an "Engineering Fellow" in 2000.

Potential tags:

LOCATION
TIME
PERSON
ORGANIZATION
MONEY
PERCENT
DATE

Semantic Annotation

- Annotation" generally means "semantics applied to a document or information resource by a person" -- rather than by NLP
- Someone other than the author can sometimes figure out the author's intent and context if:
  - they both belong to the same narrow organization, "community of interest" or "social network"
  - there are "extraction," "summarization," and other text processing tools that can help them

Semantic Authoring

- But even when an author is creating his own semantically-encoded content or annotation ...
- How are semantic descriptors chosen?
- What do those descriptors mean?
- Can others trust what the author does?

Semantic Authoring: The 2001 Vision (1)

- “The clinic's web page will have more than just keywords; it will have computer-processable information about when specific doctors take appointments"
- "These semantics were encoded into the Web page when the clinic's office manager (who never took Comp Sci 101) massaged it into shape using off-the-shelf software for writing Semantic Web pages along with resources listed on the Physical Therapy Association's site”

Semantic Authoring: The 2001 Vision (2)

- What might the office manager have learned in Comp Sci 101 (or INFO 202) that is now somehow unnecessary?
- What use is the Physical Therapy Association site?
- What specifically is the "off-the-shelf" software going to do?

Semantic Authoring: Today's Reality

- Many current semantic web tools still require expertise in semantic technologies and web standards (e.g., Protégé)
- SWAT project underway to put a natural language front end to semantic authoring
  - Refining and inferring from a verbal description
- More likely to succeed are applications that aim lower, not trying to encode all the latent semantics in a document or web page

Semantic Templating with Microformats

- Microformats make the web "semantic light" by embedding content annotation into web pages
- Microformats currently exist for personal contact information, events, and a few other small chunks of structured data
- Wikis and blogs have templates to encourage the creation of more structured and semantically-annotated content (mention a movie, get schema.org/movie)
- Wikipedia has thousands of templates and "infoboxes" that encourage the creation of factual information in a uniform format

To Summarize…

- The original vision of the Semantic Web emphasized the creation of ontologies that robustly described the semantics of particular domains or contexts
- Lots of research was spawned by this vision, but the high bar of formal semantics and automated agents undoubtedly deterred “regular” people and firms from adopting it
- Semantic authoring can’t take off without tools that are simple to use as tools for designing and creating HTML pages

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

15 October 2013
Lecture 14.2 – RDF and other Semantic Web Technologies

Technologies for the Semantic Web

- XML
- RDF and RDFa
- Ontology languages

XML is a Good Start

- You can use XML to create a content-oriented vocabulary rather than the presentation-oriented one in HTML
- XML schemas allow you to specify structural, occurrence, and datatyping constraints for instances that must conform to them
- You can use XML namespaces to reuse XML constructs across a set of related document types

XML Alone is NOT Sufficient

- But the semantics associated with XML constructs are NOT explicitly represented in the instance or the schema
- Element and attribute names, container structures, etc. can suggest semantics to people, but not in a way that is “computable”
- What is the meaning of?

<quantity>5</quantity>

<price>100</price>

Resource Description Framework (RDF)

- The Resource Description Framework (RDF) is a graph-based model for making computer-processable statements about web resources and their relationships to each other
- RDF can be used to encode metadata, the usual sort of information about an information resource, like its title, author, creation date, etc.

Resource Description Framework (RDF)

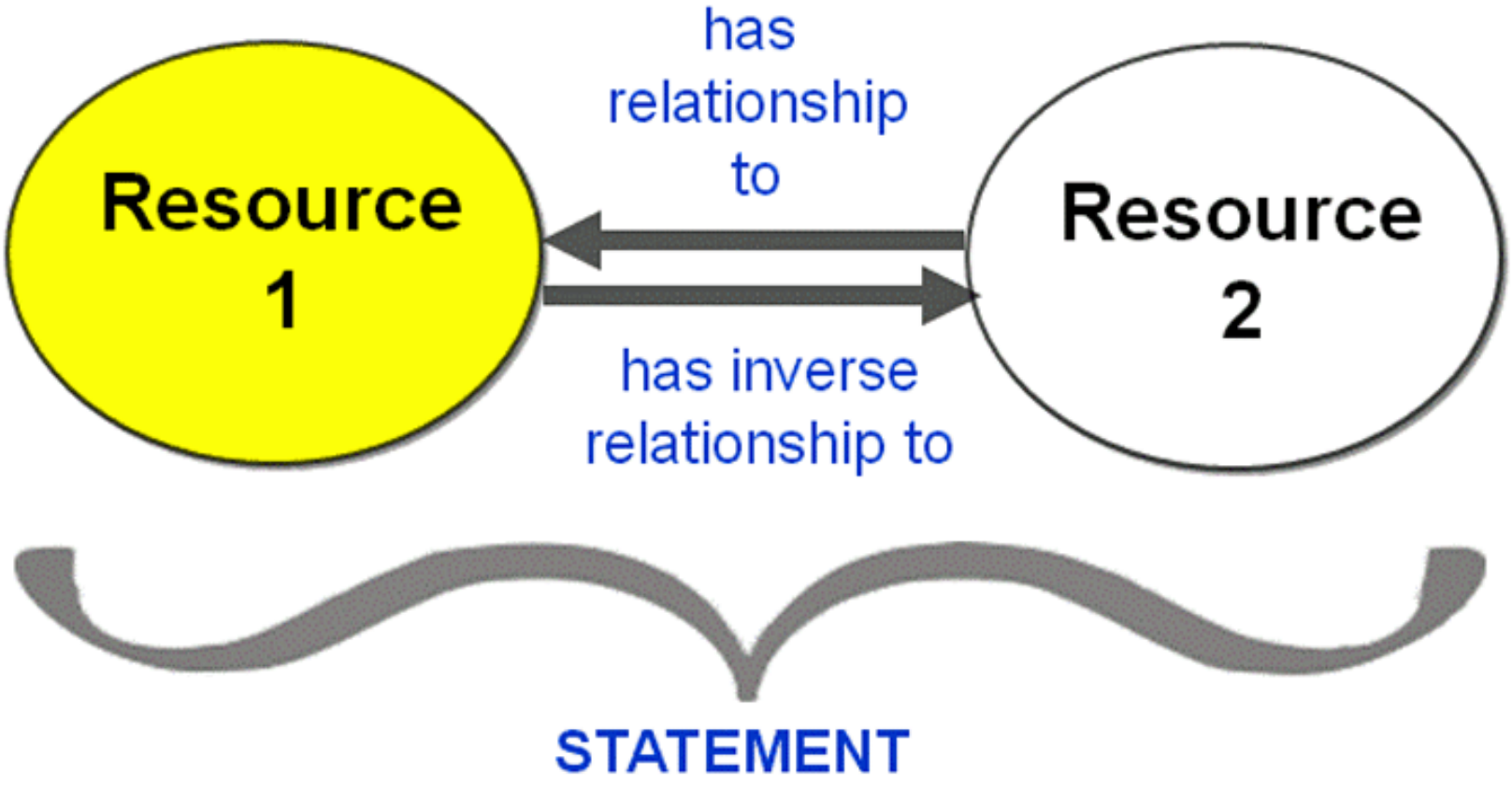
- In the context of RDF and the web, however, “resource” means something more specific: a resource is anything that has been given a Uniform Resource Identifier (URI)
- So RDF can be used to represent information about anything that can be IDENTIFIED on the Web, not just published or retrieved on it
- This broader idea about Web resource makes it a general mechanism for organizing and integrating information

RDF Data Model -- The Conceptual View

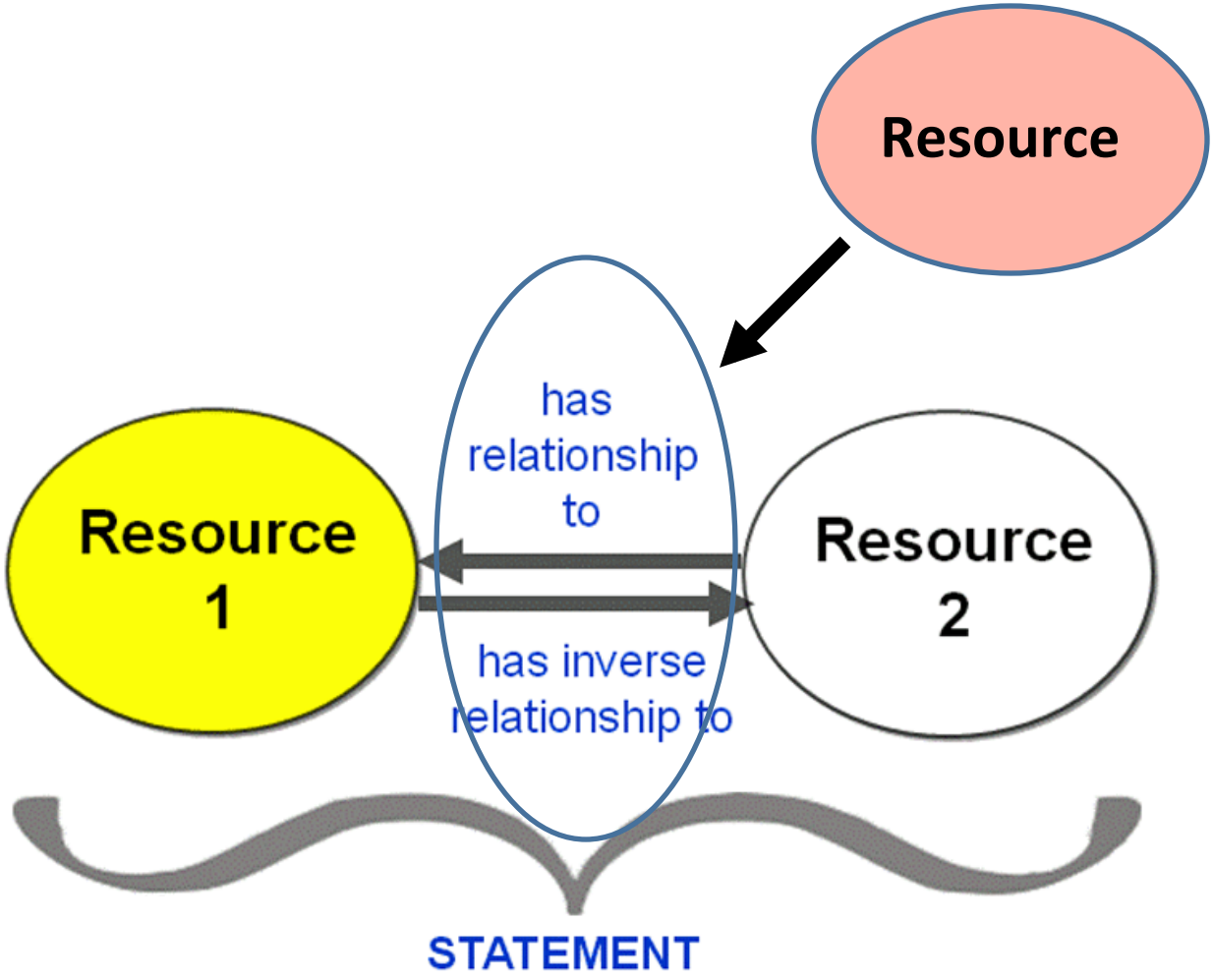
- A general way to represent information about something is in three parts:
  - The thing (or resource) being described
  - The specific property of the thing
  - The value of the property
- The data model is usually stated as Statement -> (Subject, Predicate, Object) but there are many other ways to say it

  Statement (Student, attends, University)

Statements About Resources

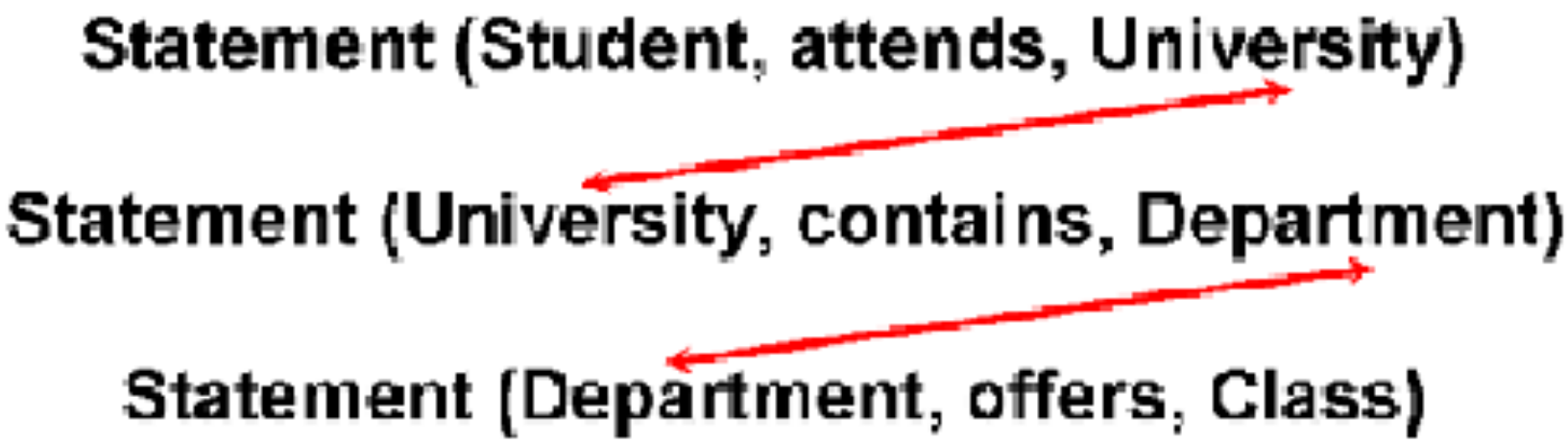


The Relationship is also a “Resource” with a Unique Identifier



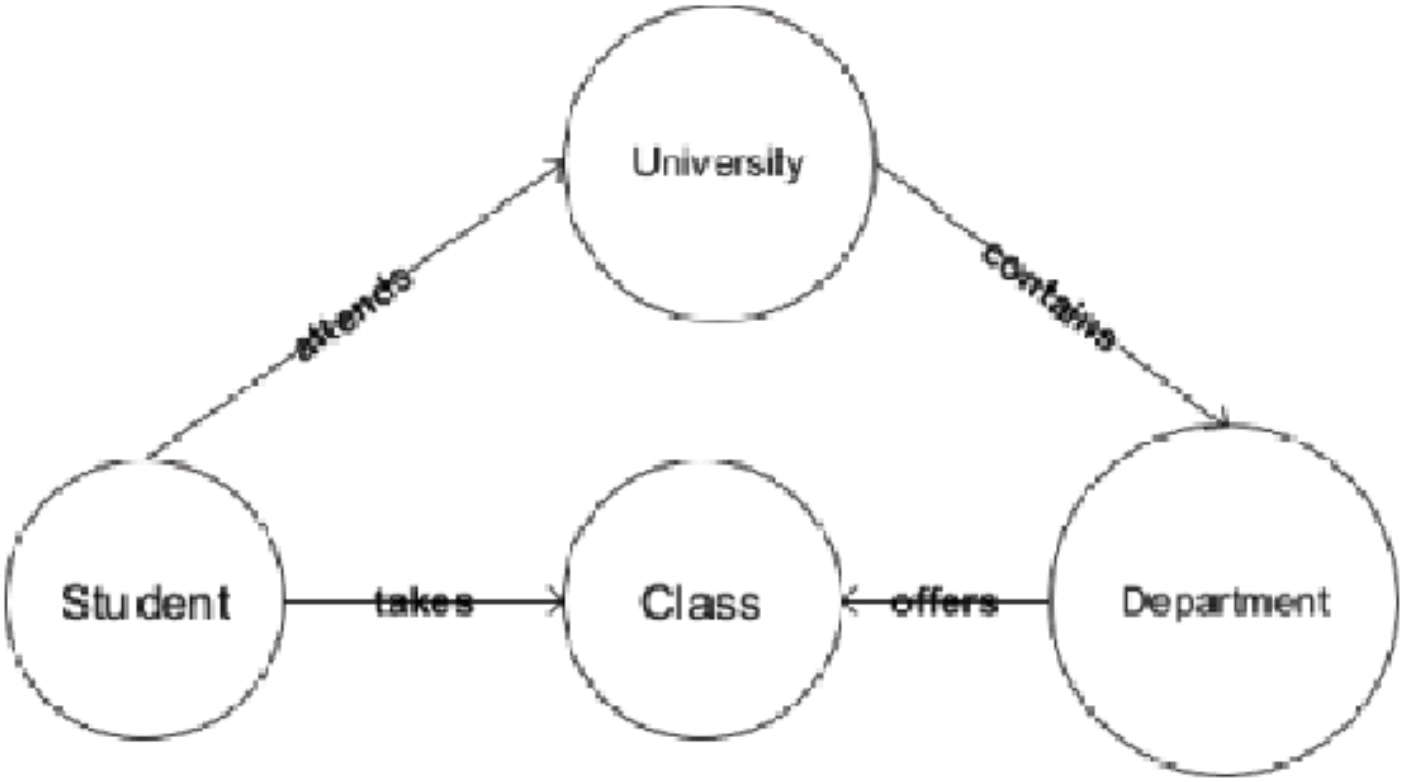
Statements as Interconnecting Building Blocks

- Because each resource is uniquely identified, statements that involve the same resource can be interconnected



Interconnected Statements Create a Graph

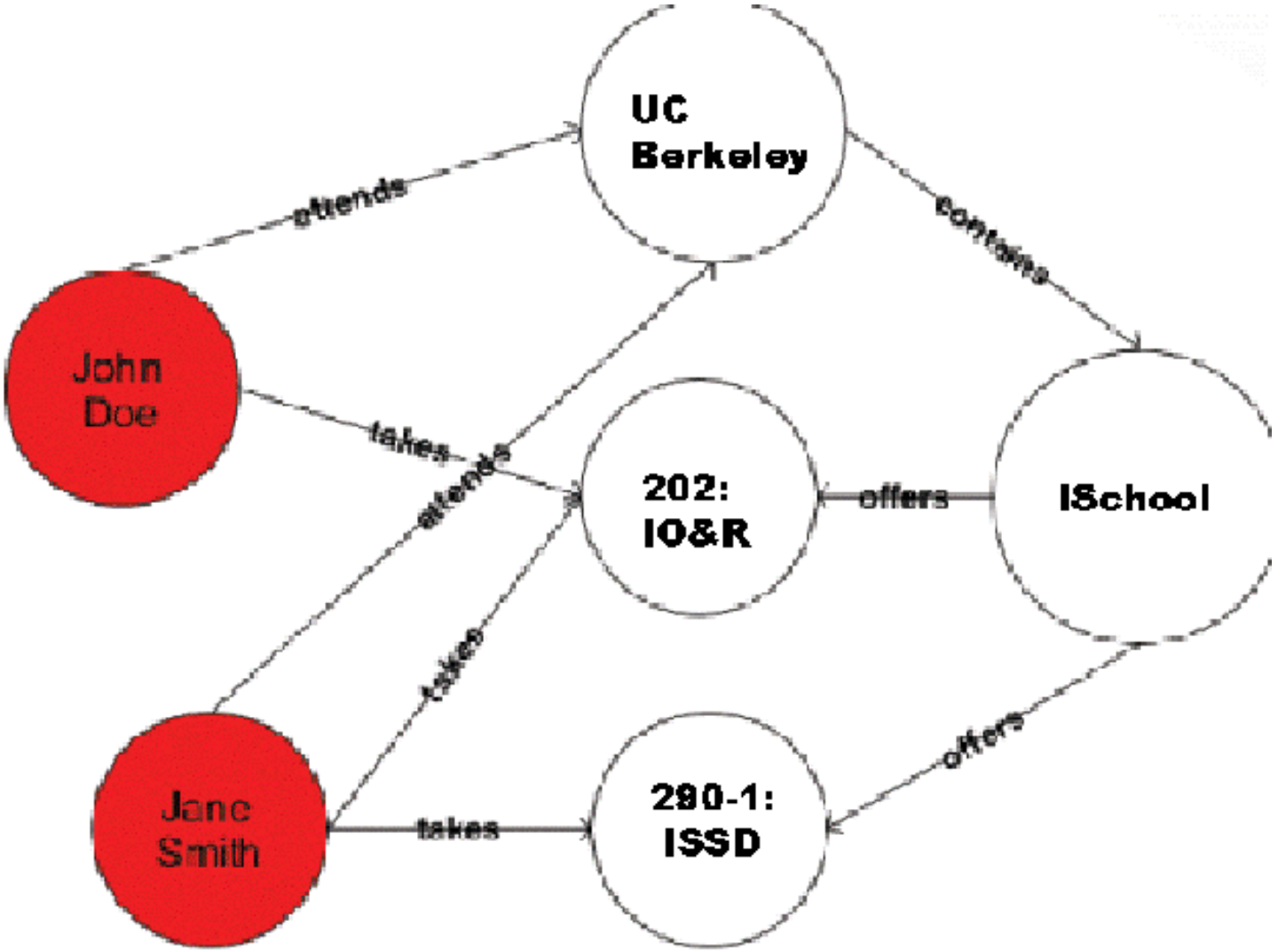
- These are statements about resource types that are part of the conceptual model of a domain



Some Statements About Instances

- John Doe attends UC Berkeley
- John Doe takes 202
- Jane Smith attends UC Berkeley
- Jane Smith takes 202
- Jane Smith takes 290-1

Graphical Representation



Can We Make These Inferences?

- John Doe takes classes at the ISchool
- 202 is offered by UC Berkeley
- Jane Smith is a classmate of John Doe

RDF Syntax – Simplified Implementation View

- RDF is a conceptual model that must be “serialized” into some specific data syntax
- <Description> describes a resource
- Attributes or elements contained in <Description> are properties of the resource
- Their content is the value of the property

RDF Description Example
(from Heath & Bizer p. 18)

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:foaf="http://xmlns.com/foaf/0.1/">

  <rdf:Description rdf:about="http://biglynx.co.uk/people/dave-smith">
    <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
    <foaf:name>Dave Smith</foaf:name>
    ...
```

- Two RDF triples in the description
  - This URI is a person
  - The person is named Dave Smith

RDFa

- RDFa is a second serialization syntax for RDF
- RDFa uses attributes (hence the "a") to embed RDF statements into XHTML elements
- So instead of using RDF in standalone documents, or using the RDF namespace to embed statements in other XML documents, RDFa lets you embed these statements into XHTML instances
- Very useful in situations where data publishers can’t change publishing technology from HTML to XML but could change templates

RDFa {and,or,vs} Microformats

- This makes RDFa like microformats in that you can add some new attributes to an HTML document
- But unlike microformats that are pre-defined for specific types of data, RDFa can be used to say anything about anything
- But the greater expressiveness of RDFa comes at a cost – harder to use

RDFa Description Example
(from Heath & Bizer p. 19)

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN"
    "http://www.w3.org/MarkUp/DTD/xhtml-rdfa-1.dtd">
<html xmlns="http://www.w3.org/1999/xhtml"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/">

<head>
    <meta http-equiv="Content-Type" content="application/xhtml+xml;
        charset=UTF-8"/>
    <title>Profile Page for Dave Smith</title>
</head>

<body>
  <div about="http://biglynx.co.uk/people#dave-smith" typeof="foaf:Person">
    <span property="foaf:name">Dave Smith</span>
  </div>
</body>

</html>
```

The Need for Ontologies (1)

- Suppose we have RDF statements:
  - (Bob Glushko, teaches, INFO202)
  - (Information System & Service Design, is-taught-by, Dr. Robert J. Glushko)
- No RDF processor can link these into a graph and make the inference that both classes are taught by the same person unless…

The Need for Ontologies (2)

- An ontology could define:
  - "Bob Glushko" and "Dr. Robert J. Glushko" to be the same person
  - "teaches" and "is-taught-by" to be inverse relationships

Ontology assertions on the Web are expressed using the OWL language

RDF Schema Features:

- *Class (Thing, Nothing)*
- *rdfs:subClassOf*
- *rdf:Property*
- *rdfs:subPropertyOf*
- *rdfs:domain*
- *rdfs:range*
- *Individual*

(In)Equality:

- *equivalentClass*
- *equivalentProperty*
- *sameAs*
- *differentFrom*
- *AllDifferent*
- *distinctMembers*

Property Characteristics:

- *ObjectProperty*
- *DatatypeProperty*
- *inverseOf*
- *TransitiveProperty*
- *SymmetricProperty*
- *FunctionalProperty*
- *InverseFunctionalProperty*

Identity, Authority, Truthiness

- The assertion that "Bob Glushko" and "Dr. Robert J. Glushko" is the same person enables statements about either to be interconnected
- But if anyone can make assertions, how do we know whether they are correct? What authorities should we trust?
  - (Barack Obama, has birthplace, Hawaii)
  - (Barack Obama, has birthplace, Kenya)
- “Defining a new URI somewhere on the Web” isn’t the same as "defining a new concept”

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

15 October 2013
Lecture 14.3 – Linked Data

Linked Data

- "Linked Data" is a reframing of basic Web principles reframed for the Semantic Web
  - instead of links as relationships in hypertext documents written in HTML, links are between arbitrary resources described by RDF
- Use URIs to identify everything
- Use HTTP URIs so that they can be de-referenced
- When someone looks up a URI, provide useful information, including additional links

From AI to BI (Erik Wilde)

- Erik Wilde nicely characterized the shift from the "Semantic Web" to "Linked Data" as "From AI to BI"
- Semantic Web is mainly about data and ontologies; its starting point was Artifiicial Intelligence(AI) efforts like CYC
- Linked Data changes the focus to real-world entities, linking data and making data discoverable; this is the Business Intelligence (BI) design pattern (ETL)

ETL – Extract, Transform, Load

- You EXTRACT some data from your repository
- You can TRANSFORM its native data model into RDF by using custom-defined mappings into a data model that is standard (enough) to be used by the "outside world"
- The RDF "triple store" can then be LOADED into the "Linked Data Cloud"
- This might be a one time conversion of your data into RDF
- A better way would be to create a linked data service that responds to requests from web agents, and then transforming those requests into queries that can be handled by a "live" triple store

Libraries & Semantic Web / Linked Data : Opportunity

- Libraries have a very long history in trying to systematize and catalog the world's knowledge
- Library data tends to be high quality when created and it is managed and maintained by people who keep it that way
- This should make library data a key trusted resource about people, publications, and organizations that could be used to improve the quality of non-library data
- The Web is increasingly the first source that people search, and the library needs to interconnect with it or it will be ignored

Authoritative Library Information

- Subject headings
- Authoritative names
- Classification

- If each of the terms in these vocabularies is identified by a URI, it would be easy to detect or prevent conflicts

Libraries & Semantic Web / Linked Data

- "Libraries could collaboratively develop a large shared knowledgebase that could act as a library "linking hub". The linking hub would expose a network of tightly linked information from publishers, aggregators, book and journal vendors, subject authorities, name authorities, and other libraries"

Libraries & Semantic Web / Linked Data : Obstacles

- However, libraries have had a deeply engrained model of how it is supposed to be done, by a few heavyweight and fairly centralized organizations, and many are reluctant to interconnecting their resources to the open and uncontrolled web
- In particular, the conceptual foundations for library data assume it will be created by professional people for people
- "The struggle to accommodate technological change with data created using the old rules is clearly not optimal, and hinders the ability of libraries to create innovative services“

Libraries & Semantic Web / Linked Data : Progress

- It hasn't been quick, but the library world seems to be adapting to the decentralized and heterogeneous world view of the Semantic Web and Linked Data
- A Bibliographic Framework for the Digital Age, Library of Congress "Manifesto" (10/31/11)
- Library of Congress Linked Data Service

Readings for Next Lecture

- TDO 6
- Glushko, Robert J., Paul P. Maglio, Teenie Matlock, and Lawrence W. Barsalou. “Categorization in the wild.”