CSAIL

Session IV

*How to **publish** and **consume** linked data on the Web*

As a data consumer you need to:

- Learn one of the open source or commercial RDF toolkits
- Find and Retrieve data you need

As a data publisher you need to:

- Convert some data to RDF
- Publish it on your website like you do with HTML

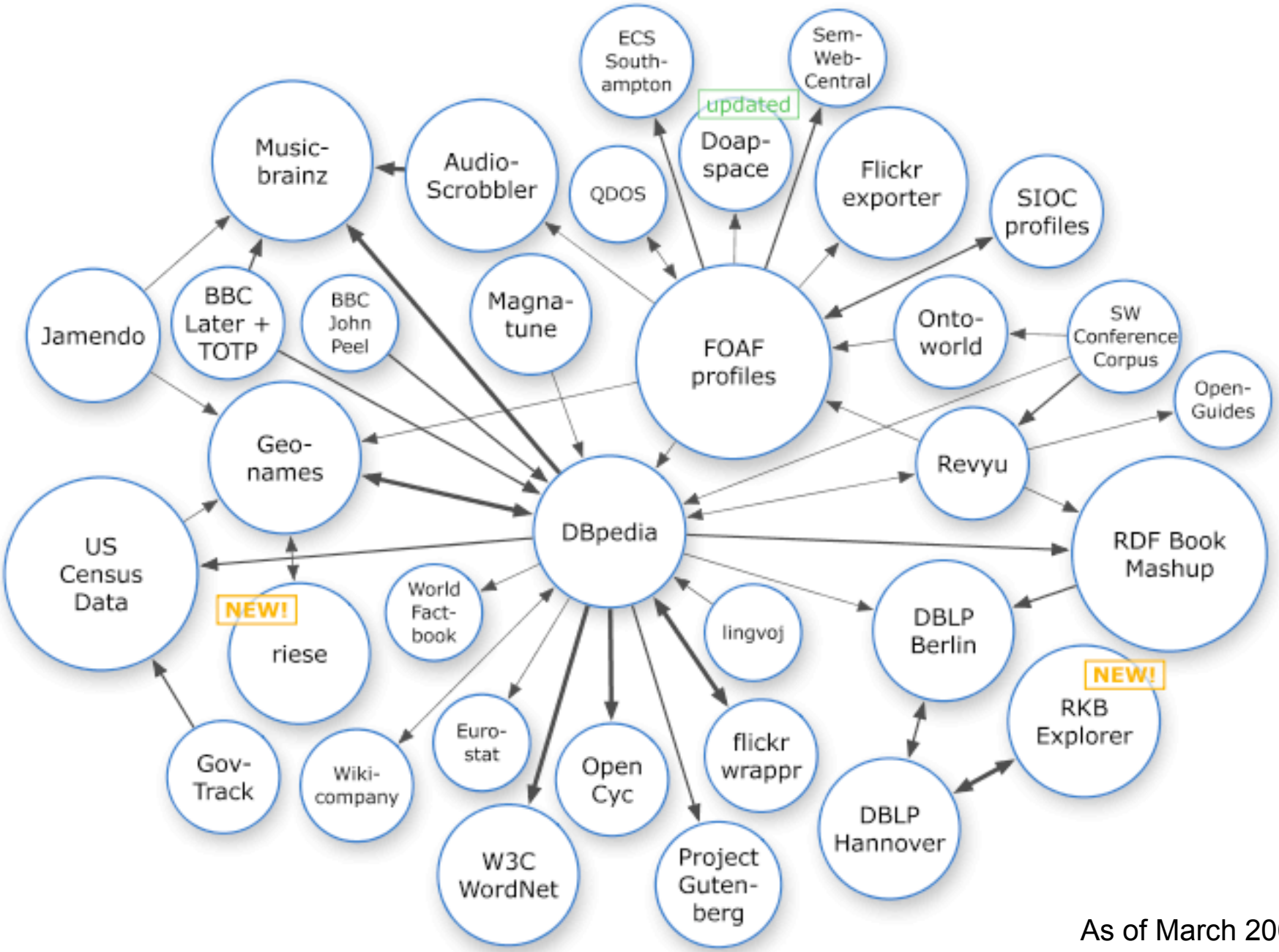
Image by Paul Miller
CC-BY

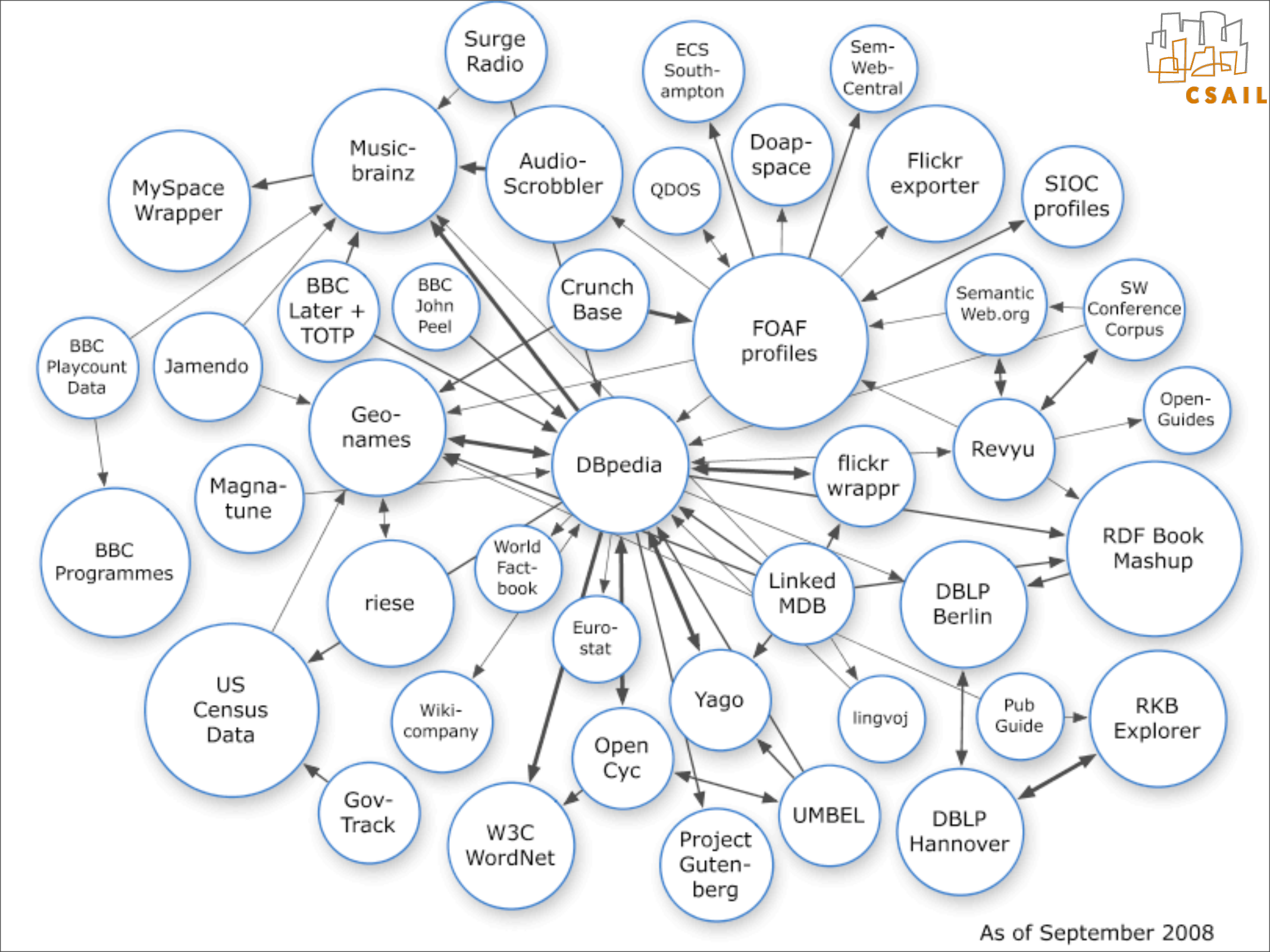
Why do I need to convert to RDF?

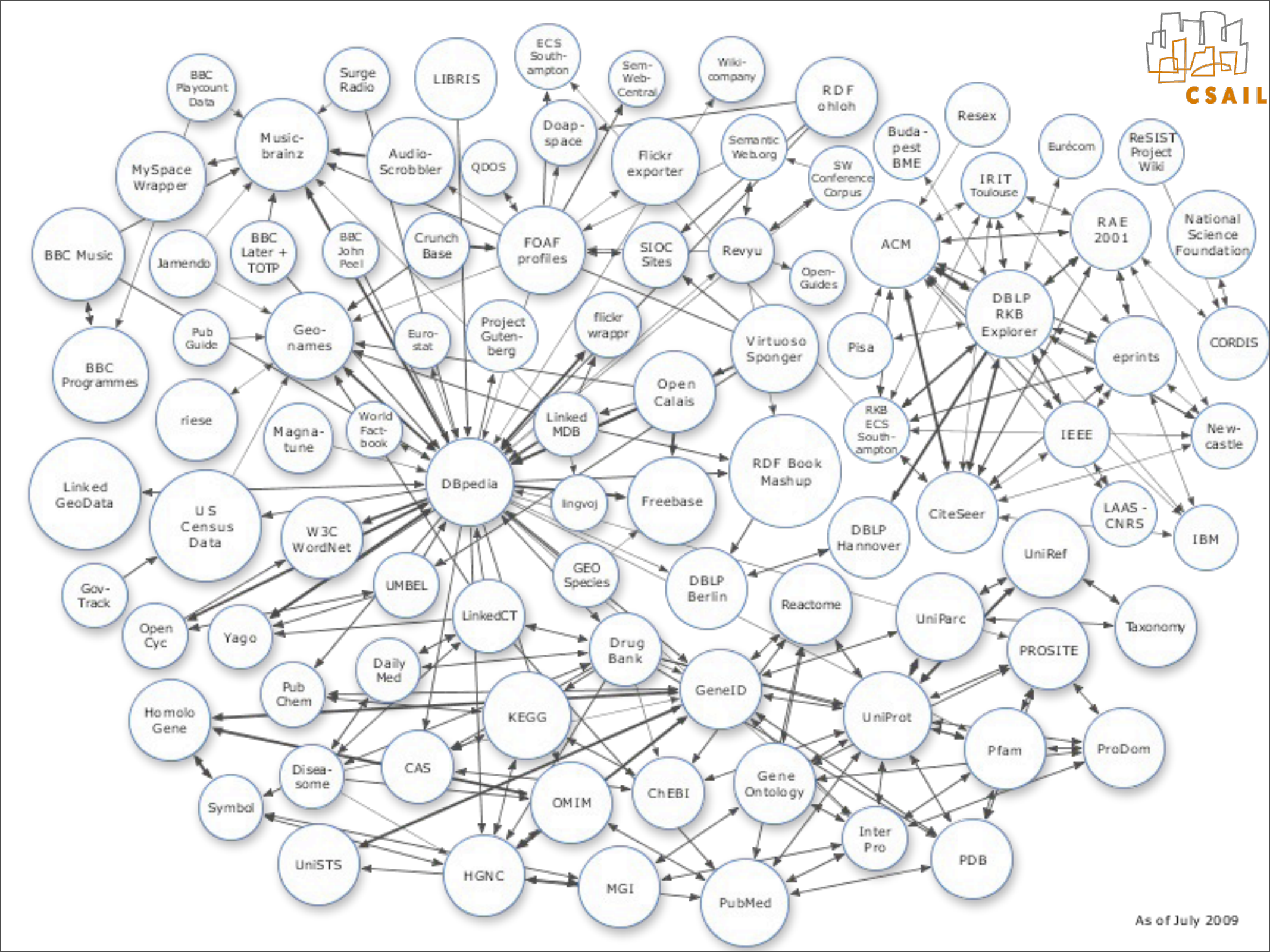
- After all, aren't there so many Web 2.0 APIs such as eBay, Amazon, Yahoo and Google Base?
- Linked data:
  - provides a single standardized mechanism instead of having to rely on diverse interfaces and results formats.
  - more easily crawled by search engines.
  - can be accessed by generic data browsers.
  - enables links between data from different data sources.

Linking Open Data Project

- Community project with W3C support started in early 2007
- Idea: take existing open data sets and make them available on the web in RDF
- Interlink them with other data sets
- Setup query endpoints
- Altogether billions of triples, millions of links...


ECS South-ampton
Sem-Web-Central
updated
Doap-space
Music-brainz
Audio-Scrobbler
QDOS
Flickr exporter
SIOC profiles
Jamendo
BBC Later + TOTP
BBC John Peel
Magna-tune
FOAF profiles
Onto-world
SW Conference Corpus
Open-Guides
Geo-names
Revyu
US Census Data
DBpedia
RDF Book Mashup
NEW!
riese
World Fact-book
lingvoj
DBLP Berlin
NEW!
RKB Explorer
Gov-Track
Wiki-company
Euro-stat
Open Cyc
flickr wrappr
DBLP Hannover
W3C WordNet
Project Guten-berg
As of March 2008


Surge Radio
ECS South-ampton
Sem-Web-Central
Music-brainz
Audio-Scrobbler
Doap-space
Flickr exporter
MySpace Wrapper
QDOS
SIOC profiles
BBC Later + TOTP
BBC John Peel
Crunch Base
FOAF profiles
Semantic Web.org
SW Conference Corpus
BBC Playcount Data
Jamendo
Geo-names
DBpedia
Revyu
Open-Guides
Magna-tune
flickr wrappr
RDF Book Mashup
BBC Programmes
World Fact-book
riese
Linked MDB
DBLP Berlin
Euro-stat
US Census Data
Yago
Pub Guide
RKB Explorer
Wiki-company
lingvoj
Open Cyc
Gov-Track
W3C WordNet
UMBEL
DBLP Hannover
Project Gutenberg
CSAIL
As of September 2008


CSAIL
BBC Playcount Data
Surge Radio
LIBRIS
ECS South-ampton
Sem-Web-Central
Wiki-company
RDF ohloh
Resex
Buda-pest BME
Eurécom
ReSIST Project Wiki
Music-brainz
MySpace Wrapper
Audio-Scrobbler
QDOS
Doap-space
Flickr exporter
Semantic Web.org
SW Conference Corpus
IRIT Toulouse
RAE 2001
National Science Foundation
BBC Music
Jamendo
BBC Later + TOTP
BBC John Peel
Crunch Base
FOAF profiles
SIOC Sites
Revyu
Open-Guides
ACM
DBLP RKB Explorer
CORDIS
Pub Guide
Geo-names
Euro-stat
Project Guten-berg
flickr wrappr
Virtuoso Sponger
Pisa
eprints
BBC Programmes
riese
Open Calais
RKB ECS South-ampton
IEEE
New-castle
Magna-tune
World Fact-book
Linked MDB
Linked GeoData
US Census Data
DBpedia
lingvoj
Freebase
RDF Book Mashup
CiteSeer
LAAS - CNRS
IBM
W3C WordNet
DBLP Hannover
UniRef
Gov-Track
UMBEL
GEO Species
DBLP Berlin
Reactome
UniParc
Taxonomy
Open Cyc
Yago
LinkedCT
Drug Bank
PROSITE
Daily Med
Pub Chem
GeneID
UniProt
Homolo Gene
KEGG
Pfam
ProDom
Disea-some
CAS
Gene Ontology
Symbol
OMIM
ChEBI
Inter Pro
UniSTS
HGNC
MGI
PubMed
PDB
As of July 2009

Steps to publishing data

Step 1: Understand the principles

Step 2: Understand your data

- What are the key things present in your data? For e.g. is it people, places, events, books, films, musicians?
- What vocabularies can be used to describe those?

Step 3: Apply common patterns for URIs

Step 4: Setup infrastructure

Step 5: Link to other data sets

Step 1 : Linked Data Principles

- Use URIs as names for things
- Use HTTP URIs, so people can look up those names
- When someone looks up a URI, provide useful RDF information
- Include RDF statements that link to other URIs, so they can discover related things

Step 2 : Ontologies to describe data

- Many ontologies: FOAF, DC, SIOC, DOAP, SKOS, CC, etc
- Lookup namespaces for prefixes at http://prefix.cc (but don't rely on that entirely!)
- Common practice is to mix terms from different vocabularies (for e.g. use of rdfs:label and foaf:depiction)
- Reuse as much as you can!

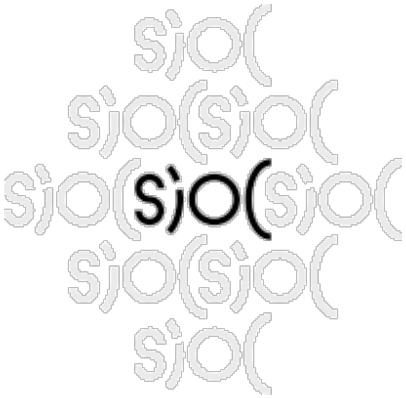
Step 2 : Ontologies to describe data
FOAF

- “Friend Of A Friend”
- Vocabulary for describing people
- See http://www.foaf-project.org/

Step 2 : Ontologies to describe data DC

- “Dublin Core”
- Defines general metadata attributes.
- These include information about information resources (dc:title, dc:creator), digital document license (dc:license)
- See http://dublincore.org/documents/dces/

Step 2 : Ontologies to describe data
SIOC



- “Semantically Interlinked Online Communities”
- Vocabulary for representing online communities
- Adds a social aspect to the Semantic Web
- See http://sioc-project.org

Step 2 : Ontologies to describe data

DOAP

- “Description Of A Project”
- Vocabulary for describing projects
- Describes concepts (such as doap:Project, doap:Repository), and properties (such as doap:mailing-list, doap:developer) related to software project management
- See http://trac.usefulinc.com/doap

Step 2 : Ontologies to describe data

CC

- “Creative Commons”
- Vocabulary for describing license terms
- “Rights Expression Language” describing terms such as cc:attributionName, cc:attributionURL, etc...
- See http://www.w3.org/Submission/ccREL

Step 2 : Ontologies to describe data
SKOS

- “Simple Knowledge Organization System”
- Vocabulary for representing taxonomies and loosely structured knowledge (for e.g. thesauri, classification schemes)
- Practical use is interlinking common categories with SIOC; to share tags and topics across different content
- See http://www.w3.org/2004/02/skos

Exercise 1

- Think of ways you can extend your FOAF profile.

Exercise 1 Solution 1

- You can give a license to your FOAF document

```
<> dc:license <http://creativecommons.org/licenses/by-nc/3.0/>;
```

Exercise 1 Solution 2

- You can say you are a developer of a particular project

```
<URI_of_some_project> doap:developer <#i>;
```

Step 3 : What are common patterns for URIs?

http://dbpedia.org/resource/MIT	Thing
http://dbpedia.org/data/MIT	RDF data
http://dbpedia.org/page/MIT	HTML page

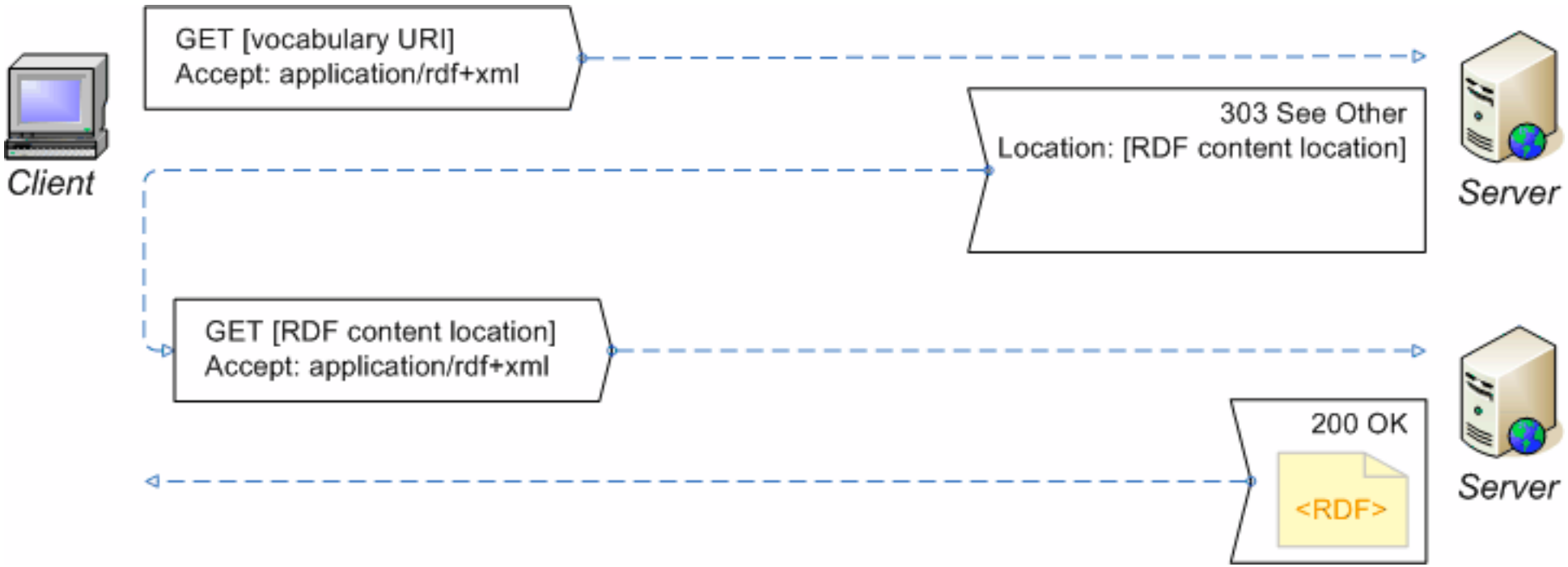
http://example.com/thing	Thing
http://example.com/thing/rdf	RDF data
http://example.com/thing/html	HTML page

http://example.com/thing	Thing
http://example.com/thing.rdf	RDF data
http://example.com/thing.html	HTML page


Ceci n'est pas une pipe.

You are not your homepage!

Step 4 : Setup infrastructure



Step 5 : Link to other data sets

Popular predicates for linking:

- owl:sameAs
- rdfs:seeAlso
- rdfs:subClassOf
- foaf:homepage
- foaf:topic
- foaf:based_near
- foaf:maker
- foaf:depiction
- foaf:page
- foaf:primaryTopic

Other ways of linking and extracting data

RDFa

- “Resource Description Framework in attributes”
- Uses RDF with XHTML
- By adding some meta information, the same resource can be reused for data integration, better mashups, etc
- See http://www.w3.org/TR/xhtml-rdfa-primer

RDFa Example

```
<div about="http://uri.to.newsitem">
  <span property="dc:date">March 23, 2004</span>
  <span property="dc:title">Rollers hit casino for £1.3m</span>
  By <span property="dc:creator">Steve Bird</span>. See
  <a href="http://www.a.b.c/d.avi" rel="dcmtype:MovingImage">
  also video footage</a>…
</div>
```

yields, through and RDFa processor

```
<http://uri.to.newsitem>
  dc:date                "March 23, 2004";
  dc:title               "Rollers hit casino for £1.3m;
  dc:creator             "Steve Bird";
  dcmtype:MovingImage <http://www.a.b.c/d.avi>.
```

GRDDL

- “Gleaning Resource Descriptions from Dialects of Languages”
- Uses intelligent “scrapers” or “wrappers” (XSLT script) to extract structure from web pages or XML files
- See: http://www.w3.org/2004/01/rdxh/spec

Exercise 2

Identify the embedded RDF in this web page:

http://dig.csail.mit.edu/2010/LinkedData/testdata/rdfa.html

Exercise 2 Solution

Human readable form

content

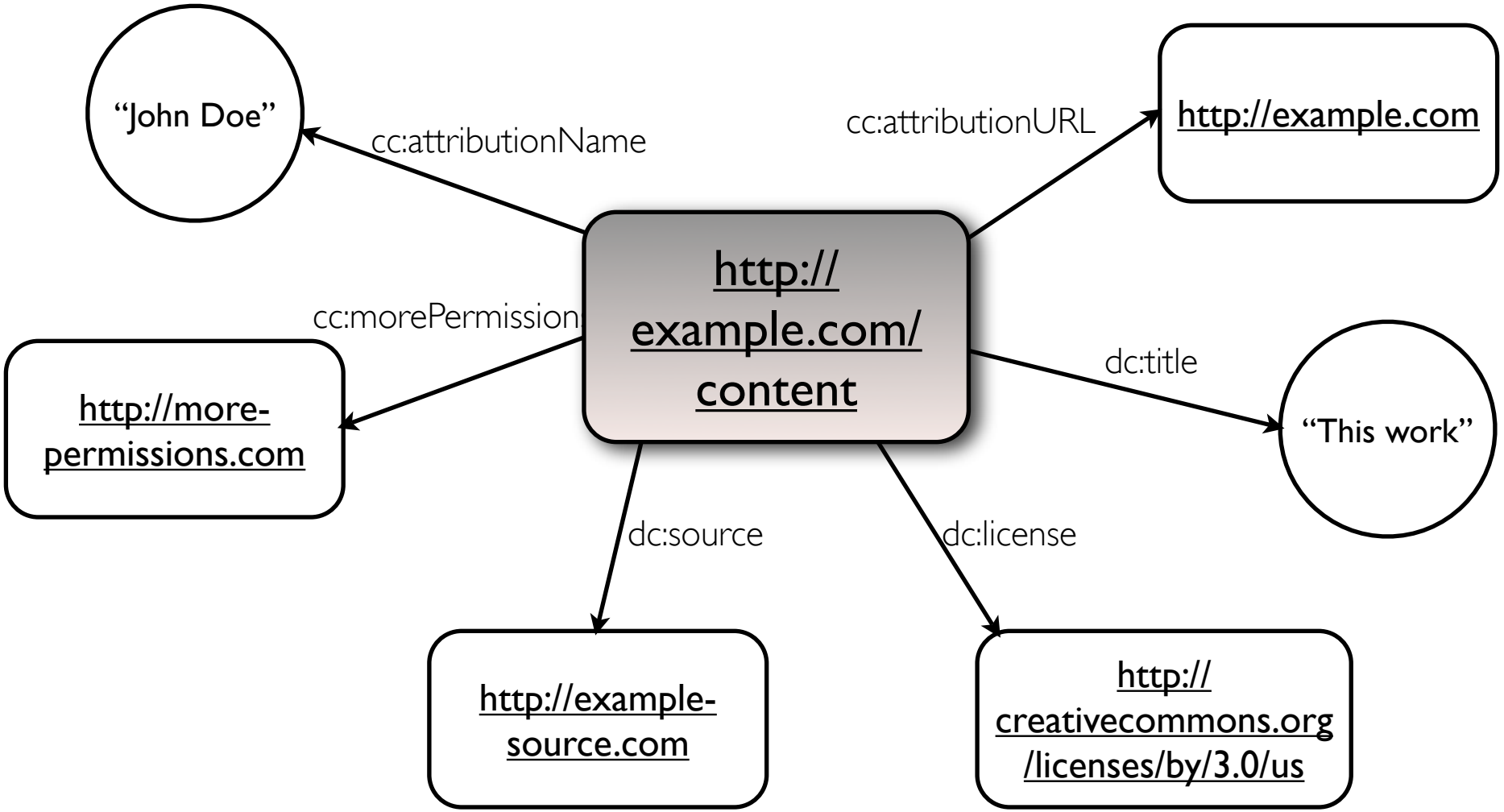
Exercise 2 Solution

Under the hood (HTML + RDFa)

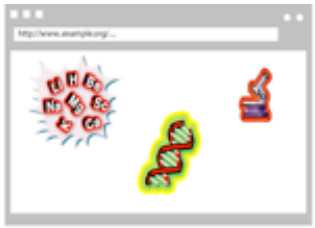
```
<div xmlns="http://www.w3.org/1999/xhtml " xmlns:dc="http://purl.org/dc/terms/ " xmlns:cc="http://creativecommons.org/ns#" about=
▼ <a rel="license" href="http://creativecommons.org/licenses/by/3.0/us/ ">
    <img alt="Creative Commons License" style="border-width:0" src="http://i.creativecommons.org/l/by/3.0/us/88x31.png ">
  </a>
  <br>
  <span href="http://purl.org/dc/dcmitype/Text " property="dc:title" rel="dc:type">This work </span>
  " by "
  <a href="http://www.example.com " property="cc:attributionName" rel="cc:attributionURL">John Doe </a>
  " is licensed under a "
  <a rel="license" href="http://creativecommons.org/licenses/by/3.0/us/ "> Creative Commons Attribution 3.0 United States License </a>
  ". "
  <br>
  " Based on a work at "
  <a href="http://www.examples-source.com " rel="dc:source">http://www.examples-source.com </a>
  ". "
  <br>
  " Permissions beyond the scope of this license may be available at "
  <a href="http://more-permssions.com " rel="cc:morePermissions">http://morepermssions.com </a>
</div>
```

Exercise 2 Solution

Things conveyed in the underlying RDF

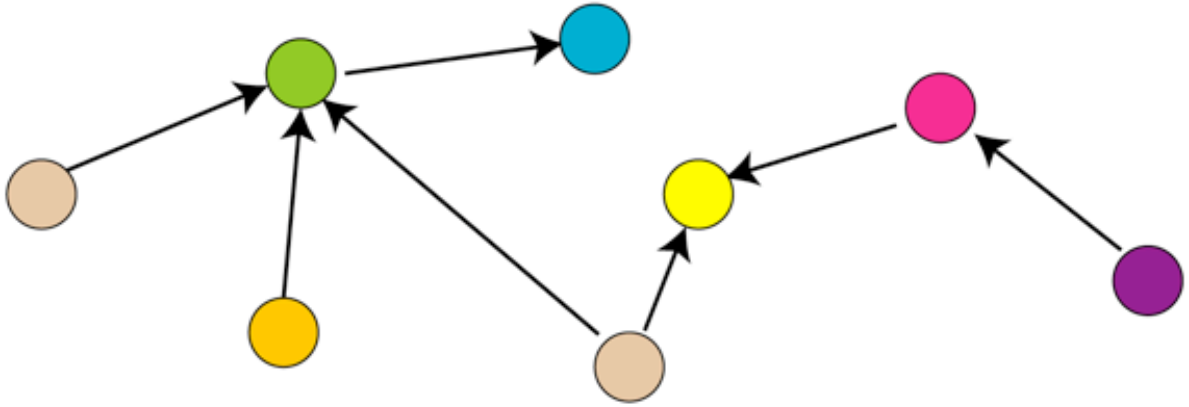


What we have learnt...



Applications

SPARQL,
OWL inferences,
etc.



Data represented in RDF, possibly with extra knowledge (RDFS, OWL, SKOS, Rules, …)

SQL <=> RDF,
GRDDL, RDFa
etc.

Data in various formats

References

- How to publish Linked Data: http://linkeddata.org/docs/how-to-publish
- List of well-known vocabularies: http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/CommonVocabularies
- Ivan Herman's Semantic Web Tutorial: http://www.w3.org/2009/Talks/0829-Nanjing-IH
- SIOC tutorial: http://www.johnbreslin.com/blog/2008/04/28/slides-from-the-sioc-tutorial-at-www2008

Some other important resources

- Semantic Web Interest Group
  - A forum of developers with publicly archived mailing lists: http://www.w3.org/2001/sw/interest
  - Constant IRC presence on irc.freenode.net#swig
- PlanetRDF aggregates a number of Semantic Web blogs: http://planetrdf.com
- Cambridge Semantic Web Gathering meets every month on the 2nd Tuesday at 6pm at MIT: http://www.meetup.com/The-Cambridge-Semantic-Web-Meetup-Group