

Linked Data in Scholarly Communication

AAHEP5 Information Provider Summit, Sept. 22nd 2011

Bernhard Haslhofer | Cornell University, Information Science

Overview

- **Linked Data Vision and Goals**
- Enabling Technologies (by example)
- Publishing and Consuming Linked Data
- The SciLink Project

SpringerLink - Multimedia T. X
 www.springerlink.com/content/6163735260027707/

springer.com springerprotocols.com English GO

SpringerLink

SEARCH FOR

AUTHOR OR EDITOR PUBLICATION VOLUME ISSUE GO

Advanced Search Search Tips


Cornell University Library

PAGE SPRINGERLINK BROWSE TOOLS HELP SHOPPING CART LOG IN

Related Journal

COMPUTER SCIENCE

MULTIMEDIA TOOLS AND APPLICATIONS
 DOI: 10.1007/s11042-011-0762-9 Online First

 **Linked Data and multimedia: the state of affairs**
 Bernhard Schandl, Bernhard Haslhofer, Tobias Bürger, Andreas Langegger and Wolfgang Halb

Download PDF (1.6 MB) View HTML

Abstract

Linked Data is a way of exposing and sharing d...
 semantically related resources. In the last thr...
 increasingly forming a globally connected, dist...
 a key factor for efficient management, organiz...
 multimedia and Linked Data has been rarely st...
 consequence thereof, technological deficienc...
 Data in the context of multimedia metadata, i...
 consume Linked Data. It shows that a large am

View Related Documents

- Reference Work Entry
Semantic Annotation and Retrieval: Web of Data Tom Heath
- Journal Article
A Database Perspective on Consuming Linked Data on the Web Olaf Hartig
- Book Chapter
Open, Distributed and Semantic Microblogging with SMOB Alexandre Passant
- Book Chapter
Fostering a Relationship between Linked Data and the Internet of Services John Domingue
- Book Chapter
A Web of Data: Toward the Idea of the Semantic Web Lihang Yu

Browse over

ScienceDirect - Web Semant... X
 www.sciencedirect.com/science/article/pii/S1570826811000242

Hub ScienceDirect Scopus SciTopics Applications Register Login Go to SciVal Suite

Brought to you by:
Cornell University Library

SciVerse ScienceDirect

Home Browse Search My settings My alerts Shopping cart Help

Articles All fields Author
 Images Journal/Book title Volume Issue Page Search ScienceDirect

Advanced search Search tips

Enter the Semantic Web Challenge!
 Click here for details

Back to results | < Previous 1 of 1 next

PDF (1286 K) Export citation E-mail article

Article Figures/Tables (15) References (49) Thumbnails Full-Size images

Web Semantics: Science, Services and Agents on the World Wide Web
 Volume 9, Issue 3, September 2011, Pages 266-283
 Semantic Web Dynamics

doi:10.1016/j.websem.2011.05.002 | How to Cite or Link Using DOI
 Permissions & Reprints

DSNotify – A solution for event detection and link maintenance in dynamic datasets

Niko Popitsch, Bernhard Haslhofer

University of Vienna, Research Group Multimedia Information Systems, Liebiggasse 4/3-4, A-1010 Vienna, Austria

Related Articles

- Searching and browsing Linked Data with SWSE: The Seman...
- Web Semantics: Science, Services and Agents on the Worl...
- Robust and scalable Linked Data reasoning incorporating...
- Web Semantics: Science, Services and Agents on the Worl...
- Challenges for the Multilingual Web of Data
- Web Semantics: Science, Services and Agents on the Worl...
- On-line recognition of handwritten characters—alphanu...
Pattern Recognition
- DBpedia - A crystallization point for the Web of Data
- Web Semantics: Science, Services and Agents on the Worl...

View more related articles

My Applications

Add

Browser: [1106.5178] The Open Annc x
 URL: arxiv.org/abs/1106.5178

Cornell University Library
 We gratefully acknowledge supporting institutions

arXiv.org > cs > arXiv:1106.5178

Computer Science > Digital Libraries

The Open Annotation Collaboration (OAC) Model

Bernhard Haslhofer, Rainer Simon, Robert Sanderson, Herbert van de Sompel
 (Submitted on 25 Jun 2011)

Annotations allow users to associate additional information with existing resources. Using proprietary and closed systems on the Web, users are already able to annotate multimedia resources such as images, audio and video. So far, this information is lost from the Web of Data. We and the Linked Data Annotation Collaboration interoperability specifications use cases and early learned and discuss

Download:
 • PDF
 • Other formats

Current browse context:
 cs.DL
 < prev | next >
 new | recent | 1106

Change to browse by:

Comments: 5 pages, 3
 Subjects: Digital Lib
 ACM classes: H.3.7
 Cite as: arXiv:1106

Submission history
 From: Bernhard Haslhofer
 [v1] Sat, 25 Jun 2011 22:
 Which authors of this pap

feedback

Browser: My Library | Mendeley
 URL: www.mendeley.com/library/

WELCOME back Bernhard Haslhofer
 My Account Upgrade
 Invite colleagues / Support

MENDELEY

Dashboard My Library Papers Groups People

Add Document Delete Documents Remove from Folder Create Folder Create Remove Web Account

My Library
 All Documents
 Recently Added
 Favorites
 Needs Review
 My Publications
 Unsorted
 (Linked) Research Data
 Databases
 Human Computation
 Information Retrieval
 Linked Data
 Multimedia
 Social Computing
 Social Search
 Web Science
 Web Stuff

Groups
 Annotations
 Linked Open Data in Science

Browser: data.europeana.eu - The Eur x
 URL: eprints.cs.univie.ac.at/2919/

inf Fakultät für Informatik
 universität wien

Home About Browse by Year Browse by Subject Browse by Research Area Browse by Author
 Login

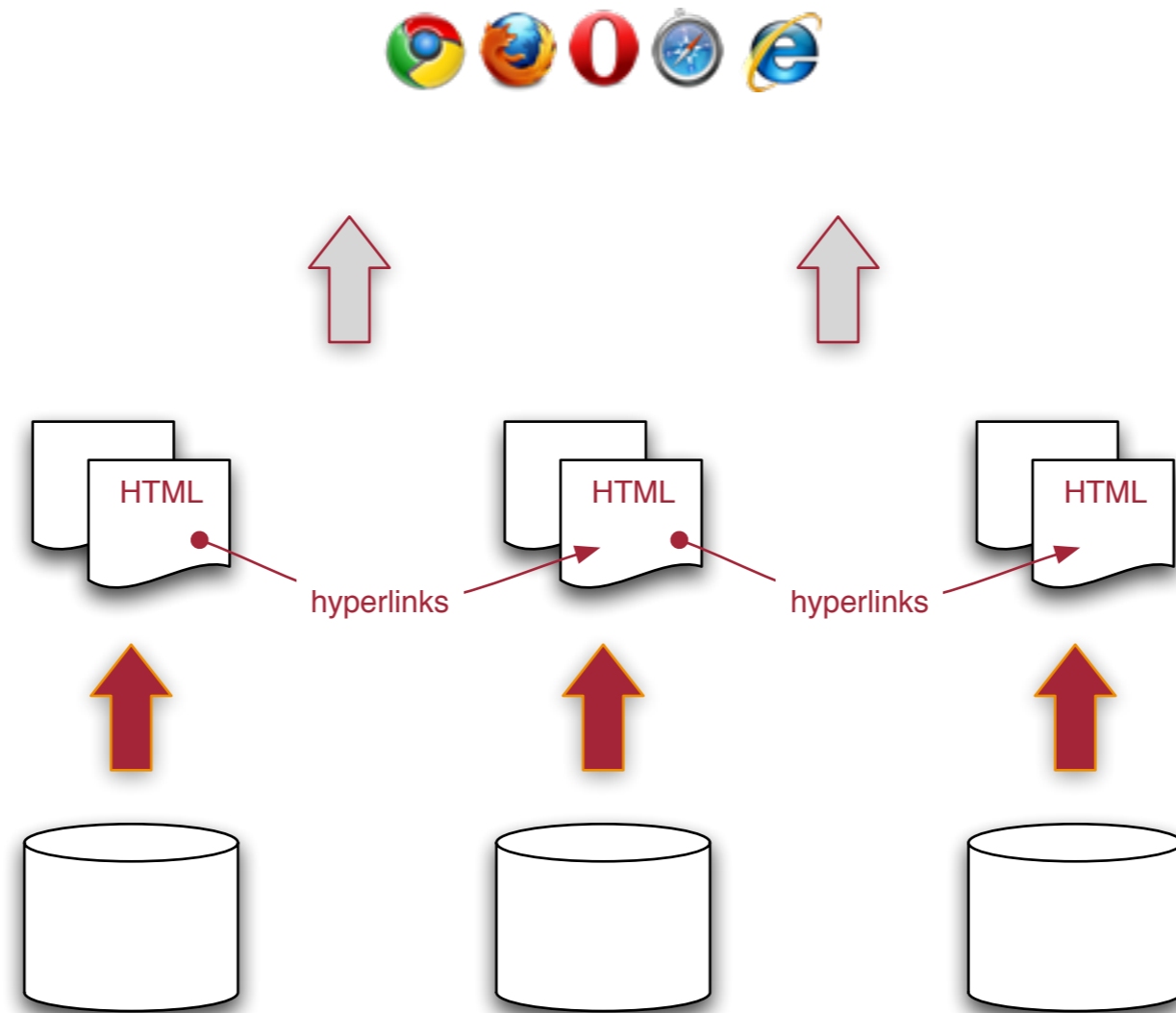
data.europeana.eu - The Europeana Linked Open Data Pilot

Content
 → Abstract
 → Authors

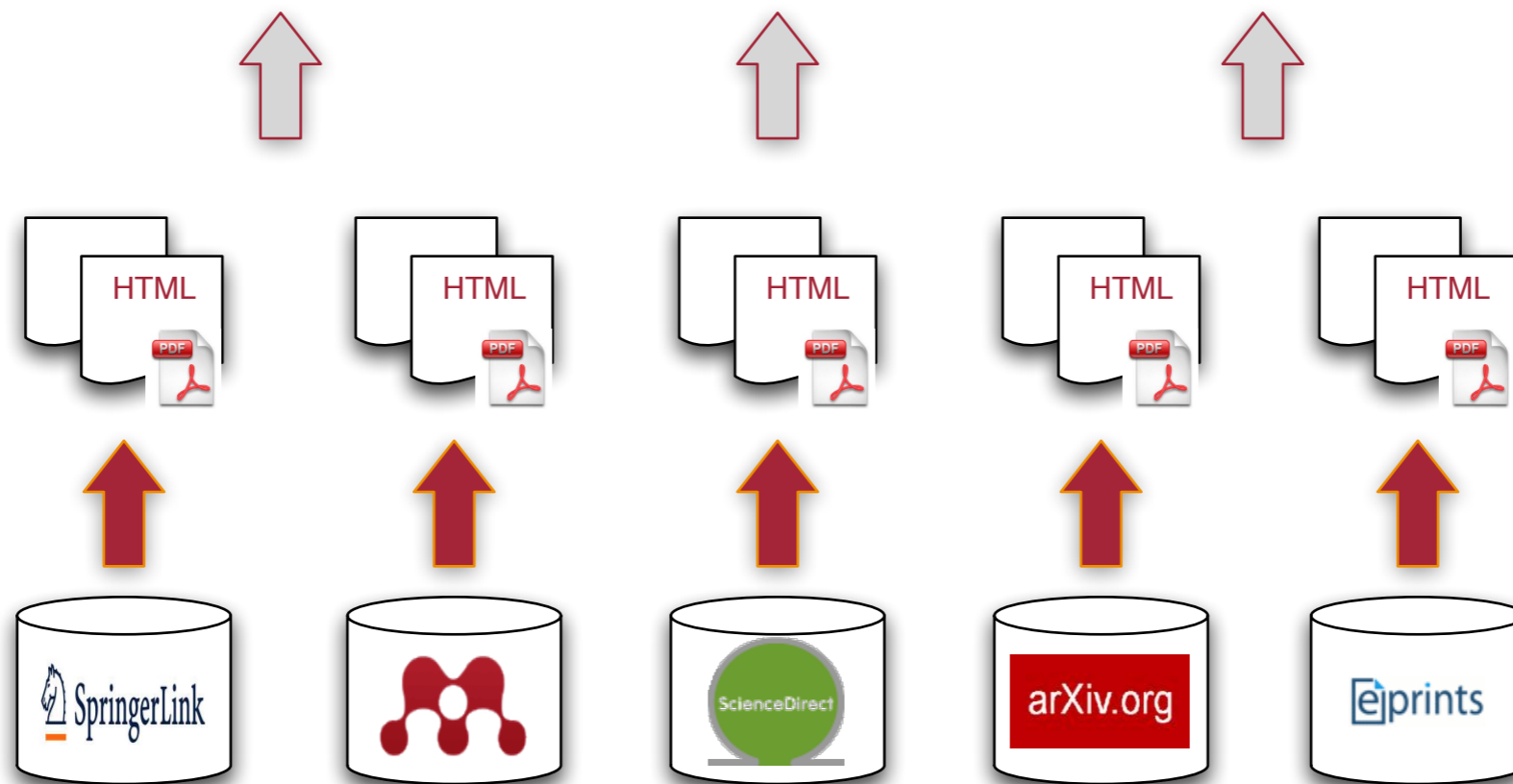
Shortfacts

Category	Conference or Workshop Item (Paper)
Event Title	DCMI International Conference on Dublin Core and Metadata Applications
Research Areas	Multimedia Information Systems
Event Location	The Hague, The Netherlands
Event Type	Conference
Event Dates	September 21-23, 2011
Date	25 July 2011

Web Architecture



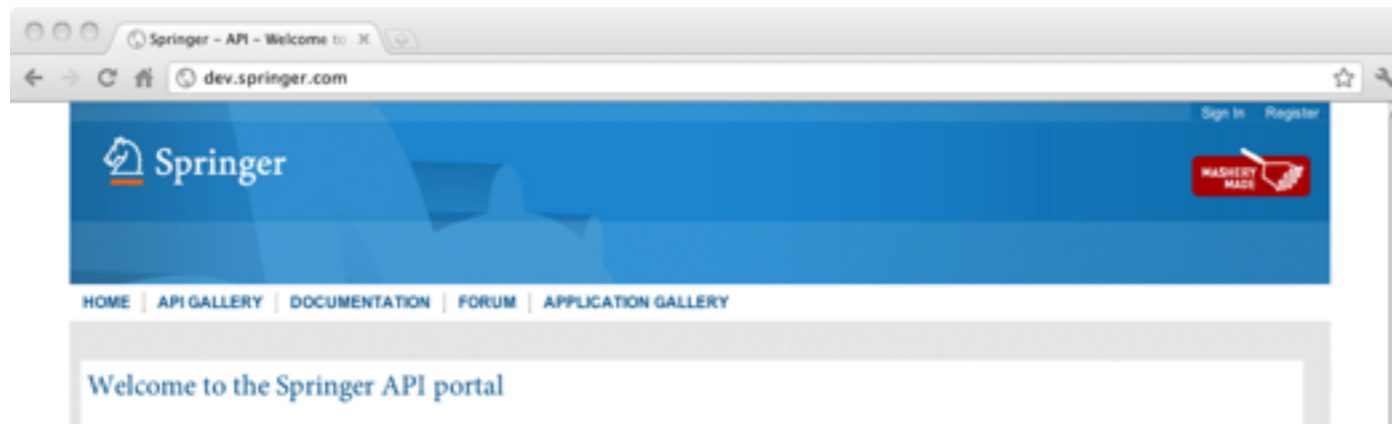
Web Architecture



Web Architecture

- A set of simple standards
 - Uniform document encoding (HTML)
 - Uniform global (!) addressing (URI)
 - Uniform transportation (HTTP)
- Hyperlinks connecting documents
- Works pretty well for accessing and exchanging **documents**

But sometimes we need to access to the
underlying (meta)data.



Hub | ScienceDirect | Scopus | SciTopics | Applications Login | Go to SciVal Suite

Developers Home | My projects | Resources | Community | Blog | My settings | Applications Help

SciVerse Content APIs

The SciVerse Content APIs provide direct access to Elsevier content using RESTful URL requests. Currently, they can only be used by applications running within the SciVerse application framework.

Note: The terms "application" and "gadget" are used interchangeably throughout this documentation. SciVerse applications are our proprietary implementation of the [Gadgets API](#), but access to the Content APIs isn't limited to SciVerse Gadgets.

Working with the SciVerse Content APIs

1. Getting Started
2. Content Basics
3. Building A Request
4. Request Examples
5. Content API Access

1. Getting Started

The SciVerse Application Framework provides a customized environment for building self-contained JavaScript applications (often called 'Gadgets' or 'Widgets'), that are basically extensions to the already existing functionality on SciVerse web sites. For more information about the Framework, [start here](#).

Applications running within the framework are not limited to using data from the SciVerse Content APIs. Any external data source that can be targeted from an application running in a browser can be integrated into the application. The SciVerse Application Framework has [functionality](#) to help a gadget search and retrieve SciVerse data through back-end API calls. The APIs that are called by these methods are the SciVerse Content APIs. Currently, the Content APIs [cannot be accessed from outside the framework](#).

Getting started

- [Learn](#)
- [Start your application](#)
- [Community](#)

Resources

Documentation & Tools

- [Download our SDK](#)
- [Content APIs](#)
- [Framework API](#)
- [Application framework](#)
- [Integration points](#)
- [XML definition file](#)
- [User interface guidelines](#)
- [How to...](#)
- [Troubleshooting](#)
- [Example applications](#)
- [Developer links](#)
- [FAQ](#)

SciVerse Developers

- [What is SciVerse?](#)
- [Request access](#)
- [Policies](#)
- [Application approval](#)

[Blog](#)

arXiv.org help - arXiv API

arxiv.org/help/api/index

Cornell University Library

We gratefully acknowledge supporting institutions

arXiv.org > help

Search or Article-id (Help | Advanced search)

Help Table of Contents | Search arXiv Help

arXiv API

This is the home site of the arXiv API. The goal with an easy-to-use programmatic interface. To list and contact other developers and maintainers.

For more information about the arXiv API, please see:

- API News
- About the arXiv API
- Quickstart
- Using the arXiv API
- arXiv API documentation
- Community

API News

See the [arXiv API Blog](#) for the latest news updates.

About the arXiv API

The Cornell University Library e-print arXiv, has mathematics and computer science communities ongoing research. The open-access arXiv e-print communication. Manuscripts are often submitted or published elsewhere, and in other

Mendeley Developers Portal

dev.mendeley.com

Return to Mendeley.com Home Documentation Mendeley data challenge My applications Register application Support Sign out

1. Getting Started
Learn how the API works with our detailed documentation.

2. Register an app
Register your application online, and start building an app.

3. Get Involved
Join the community discussions, ask questions, and help others.

Mendeley has recently released its Open API! [Read more](#)

Mendeley/PLOS API Binary Battle

Build an app, make science more open, win \$10,001 + \$1,000 Amazon Web Services Credits!

Grand Prize \$10,001 + \$1,000 Amazon Web Services Credits

Runner up prizes [more!](#)

★ Submit your app by September 30th, 2011

We've built the world's largest crowdsourced research database, available under a Creative Commons license. We want to see a world in which science is mashed up... with anything. Science can benefit everyone, but it needs the support of the tech community to make it happen.

http://eprints.cs.univie.ac.at

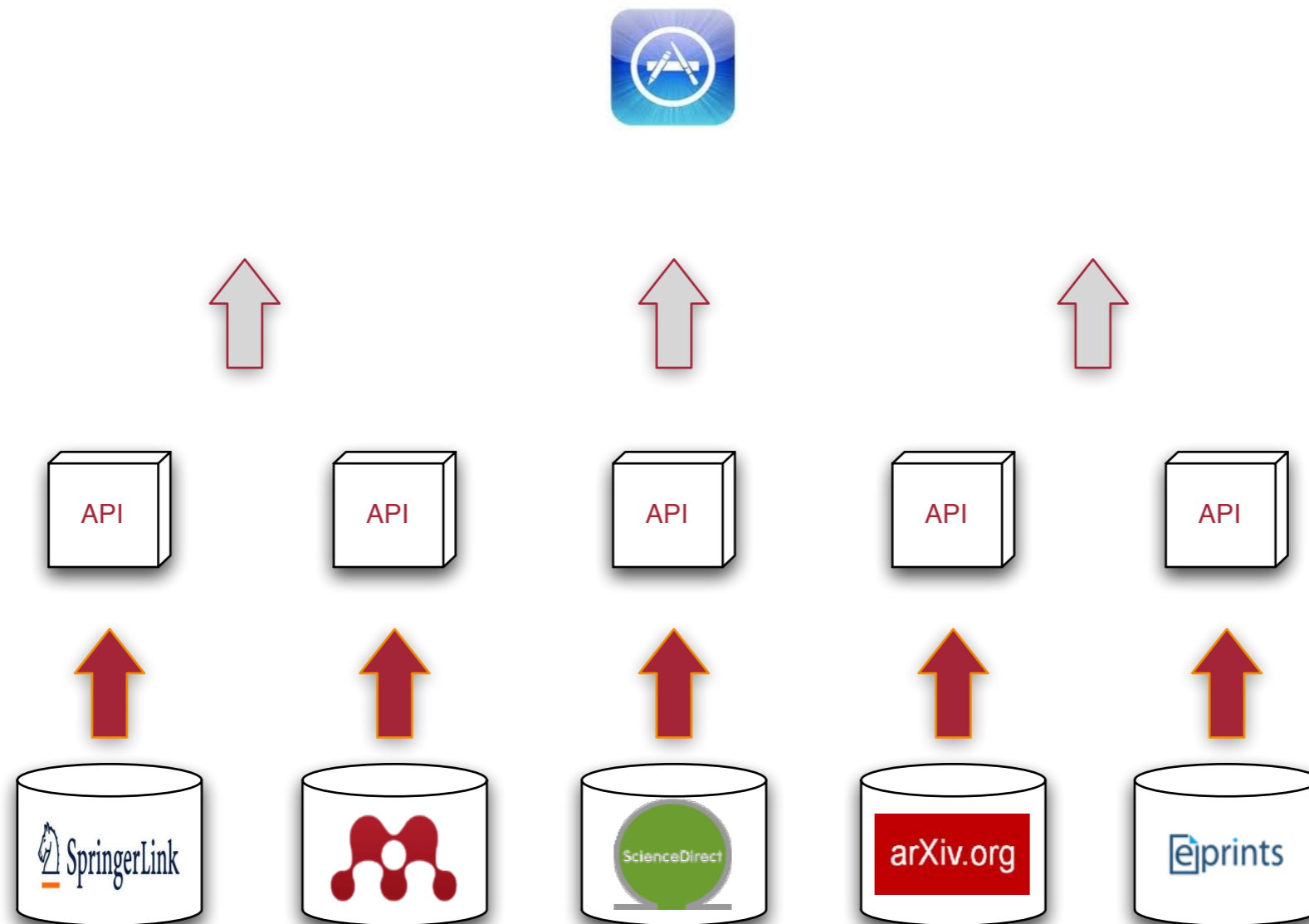
eprints.cs.univie.ac.at/cgi/exportview/creators/Haslhofer=3ABernhard=3A=3A/Atom/Haslhofer=3ABernhard=3A=3A.xml

Click entry for XPath. Double-click to collapse/expand. Enter XPath or XML string then click XPath

Source Options XPath/Render XML

```
<feed xmlns="http://www.w3.org/2005/Atom">
  <title>Faculty of CS Repository: Metadata Visibility matches "Always Show" AND Creators is "Haslhofer, Bernhard". Results ordered -Date, Title. </title>
  <link href="http://eprints.cs.univie.ac.at/" />
  <link rel="self" href="https://eprints.cs.univie.ac.at/cgi/exportview/creators/Haslhofer=3ABernhard=3A=3A/Atom/Haslhofer=3ABernhard=3A=3A.xml" />
  <updated>2011-09-15T15:02:48Z</updated>
  <id>tag:eprints.cs.univie.ac.at,2011:feed:feed-title</id>
  <entry>
    <title>Quality Criteria for Controlled Web Vocabularies</title>
    <link href="http://eprints.cs.univie.ac.at/2923/" />
    <summary>Mader, Christian and Haslhofer, Bernhard (2011) Quality Criteria for Controlled Web Vocabularies. In: TPDL 2011. </summary>
    <updated>2011-08-11T08:28:06Z</updated>
    <id>tag:eprints.cs.univie.ac.at,2011-08:item:/2923</id>
    <author>
      <name>Christian Mader</name>
    </author>
    <author>
      <name>Bernhard Haslhofer</name>
    </author>
  </entry>
  <entry>
    <title>A Retrospective on Semantics and Interoperability Research</title>
    <link href="http://eprints.cs.univie.ac.at/2921/" />
    <summary>Haslhofer, Bernhard and Neuhold, Erich J. (2011) A Retrospective on Semantics and Interoperability Research. In: Foundations for the Web of Information and Services. Springer, Berlin Heidelberg, pp. 3-27.</summary>
    <updated>2011-07-25T13:04:14Z</updated>
    <id>tag:eprints.cs.univie.ac.at,2011-07-24:item:/2921</id>
    <author>
      <name>Bernhard Haslhofer</name>
    </author>
    <author>
      <name>Erich J. Neuhold</name>
    </author>
  </entry>
</feed>
```

Web Services and Web APIs

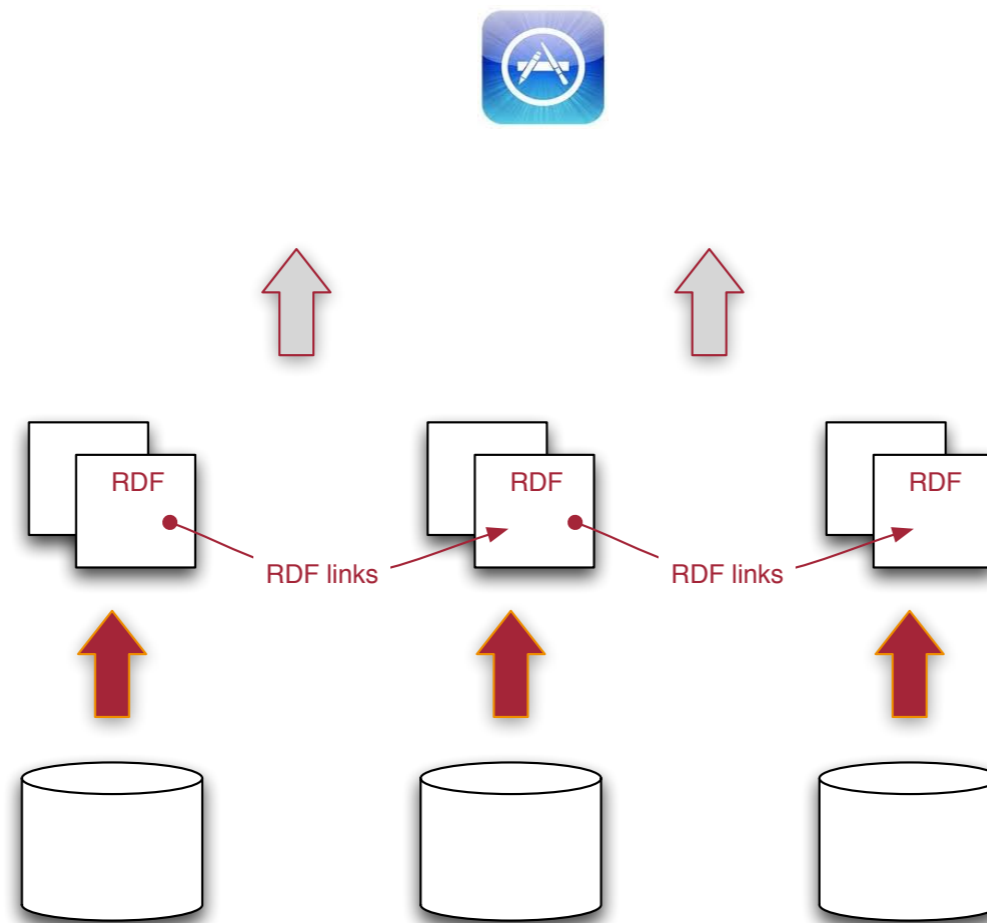


Web Services and Web APIs

- Each Web API has a proprietary interface
- Datasources must be known in advance
- Information entities (papers, authors, subjects, etc.) are not linked

Linked Data Vision

- Create a single global data space based on the Architecture of the World Wide Web



Linked Data Principles



- Use URIs as names for things
- Use **HTTP URIs** so that people can look up those names
- When someone looks up a URI, provide useful information, using standards (**RDF**, SPARQL)
- Include **links** to other URIs, so that they can discover more things

Web of Linked Data

- A set of simple standards
 - Uniform **data** model (RDF)
 - Uniform global (!) addressing (URI)
 - Uniform transportation (HTTP)
- RDF **links** connecting entities
- Forms a global dataspace and facilitates accessing and exchanging **data**

Web of Linked Data

- Data publishers
 - assign URIs to their information entities
 - publish RDF instance data about these entities
 - publish schemas / vocabularies on the Web
 - link their entities with other entities on the Data Web
 - can provide query access to their data

Linking Open Data Project

- **Open Data:** a philosophy, practice, or policy that data are freely available to everyone without restrictions from copyright, patents, a.s.o
- **Linked Data:** method / best practices for exposing, sharing, and connecting data using URI and RDF
- **Linking Open Data:** a W3C community project with the goal to extend the Web with a data commons by publishing various open data sets as RDF on the Web and by setting links between data items from different sources



LOD Cloud

[View](#) [History](#)[Subscribe](#)

Datasets:

There are 311 datasets in this group.

1 2 3 .. 7 Next »

2000 U.S. Census in RDF (rdfabout.com)

[meta/sitemap](#) [api/sparql](#) [example/rdf+xml](#)

2000 U.S. Census converted into over a billion RDF triples.

Not Openly Licensed

Population statistics at various geographic levels, from the U.S. as a whole, down through states, counties, sub-counties...

AEMET metereological dataset

[api/sparql](#) [RDF/XML](#), [HTML](#), [JSON](#) [example/rdf+xml](#) [example/turtle](#) [application/x-ntriples](#) [meta/rdf-schema](#) [mapping/owl](#) [meta/void](#)

Not Openly Licensed

AemetLinked Data (.es) is an open initiative of the Ontology Engineering Group (OEG) whose aim is to enrich the Web of Data with Spanish geospatial data. This initiative started off by...

AGROVOC

[api/sparql](#) [application/x-ntriples](#) [example/rdf+xml](#) [application/x-ntriples](#) [application/x-ntriples](#) [application/x-ntriples](#) [application/x-ntriples](#)[OPEN DATA](#)

AGROVOC Linked Open Data (LOD) is a project to turn the AGROVOC thesaurus into a multilingual, terminological backbone for agricultural digital goods. Hosted by research partner MIMOS it...

Amsterdam Museum as Linked Open Data in the Europeana Data Model

[api/sparql](#) [HTML](#) [api/cit](#) [example/rdf+xml](#) [example/rdf+xml](#) [HTML](#) -

This group catalogs data sets that are available on the Web as [Linked Data](#) and contain data links pointing at other Linked Data sets.

The descriptions of the data sets in this group are used to generate the [Linking Open Data Cloud diagram](#) at regular intervals. The descriptions are also used generate the statistics provided in the [State of the LOD Cloud](#) document.

If you publish a linked data set yourself, please add it to CKAN so that it appears in the next version of the LOD cloud diagram. Please describe your data set according to [Guidelines for Collecting Metadata on Linked Datasets in CKAN](#). Please tag newly added data sets with lod. The editors of this group will review your description afterwards and move your dataset into the lodcloud group.

Please also use the [CKAN LOD Validator](#) to check that the description of your data set is complete.

Administrators

- [Anja Jentzsch](#)

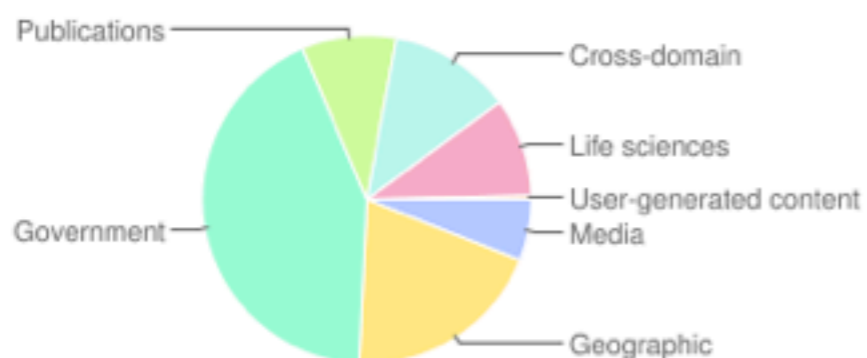
- [Anja Jentzsch](#)

1.2 Linked Data by Domain

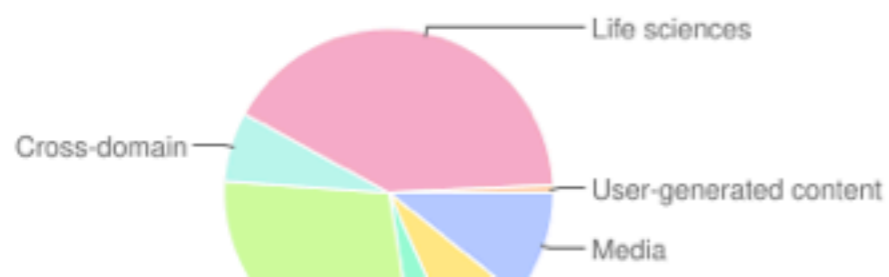
Linked Data technologies are being used to share data covering a wide range of different topical domains. The table below gives an overview of the amount of triples as well as the amount of RDF links per domain. The number of RDF links refers to out-going links that are set from data sources within a domain to other data sources.

Domain	Number of datasets	Triples	%	(Out-)Links	%
Media	27	1,855,413,060	6.01 %	50,491,015	10.72 %
Geographic	22	6,096,504,422	19.75 %	35,747,820	7.59 %
Government	39	13,229,470,882	42.85 %	19,261,998	4.09 %
Publications	82	2,868,088,257	9.29 %	135,336,031	28.72 %
Cross-domain	32	3,708,240,740	12.01 %	32,254,790	6.85 %
Life sciences	41	3,001,943,206	9.72 %	194,672,433	41.32 %
User-generated content	13	114,442,475	0.37 %	3,423,613	0.73 %
	256	30,874,103,042		471,187,700	

The diagram below shows the distribution of triples by domain.



The diagram below shows the distribution of links by domain.



Linked Data in the Industry



BBC Music - Green Day

http://www.bbc.co.uk/music/artists/084308bd-1654-436f-ba03-df6697104e19

Music BETA

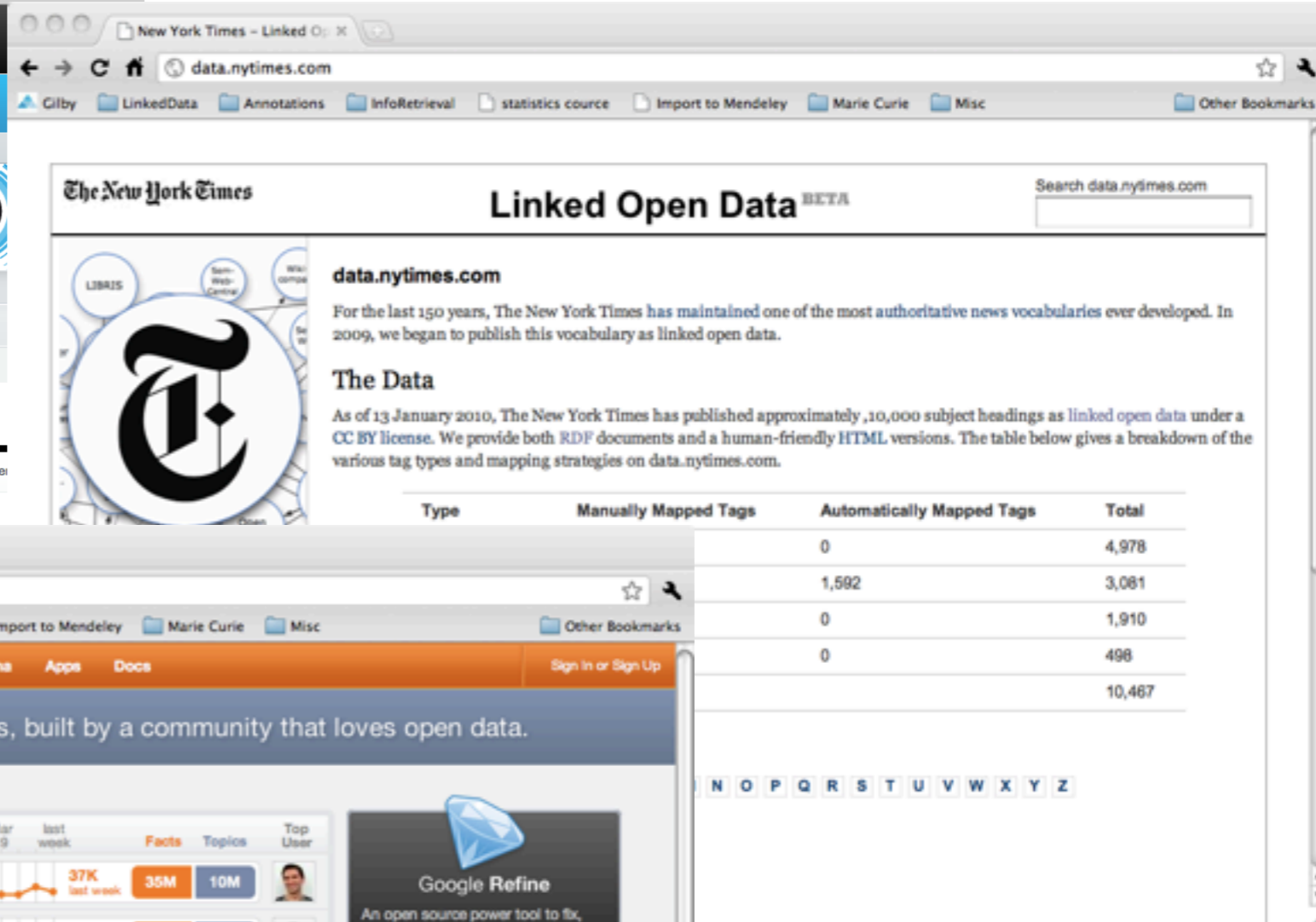
Green Day
Group. Formed 1989.

MOST PLAYED ON BBC RADIO 1

Played By
Since December 2008

- Scott Mills
BBC Radio 1
Go home with Mills as he enters
- Edith Bowman

Latest News Stories
NEWS FROM THE BBC
Tour latest
Mon 27 Apr 2009 15:25



The New York Times

Linked Open Data BETA

data.nytimes.com

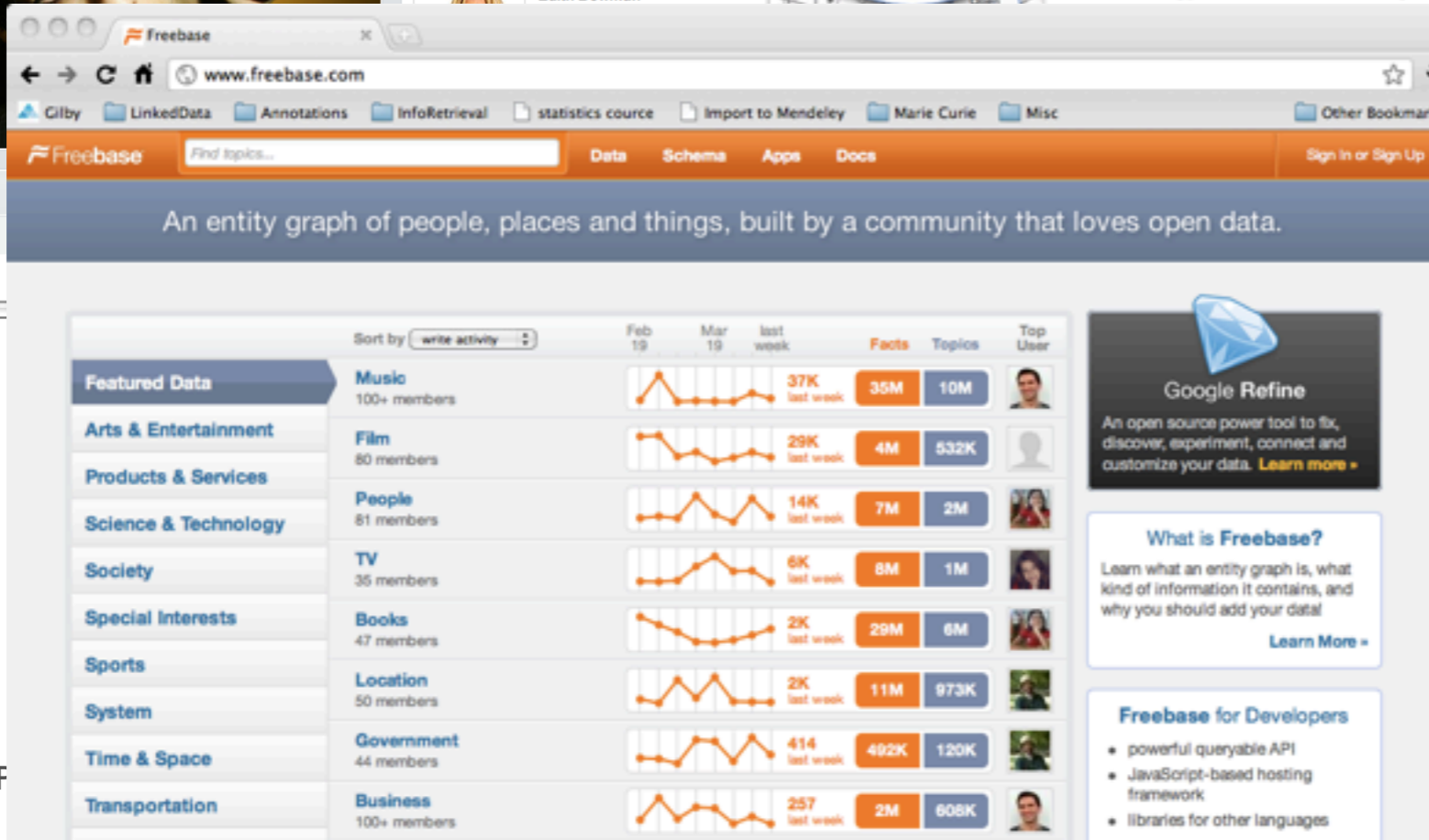
For the last 150 years, The New York Times has maintained one of the most authoritative news vocabularies ever developed. In 2009, we began to publish this vocabulary as linked open data.

The Data

As of 13 January 2010, The New York Times has published approximately 10,000 subject headings as linked open data under a CC BY license. We provide both RDF documents and a human-friendly HTML versions. The table below gives a breakdown of the various tag types and mapping strategies on data.nytimes.com.

Type	Manually Mapped Tags	Automatically Mapped Tags	Total
	0	4,978	
	1,592	3,081	
	0	1,910	
	0	498	
			10,467

N O P Q R S T U V W X Y Z



Freebase

www.freebase.com

Find topics... Data Schema Apps Docs Sign In or Sign Up

An entity graph of people, places and things, built by a community that loves open data.

Sort by	write activity	Feb 19	Mar 19	last week	Facts	Topics	Top User
Featured Data	Music	100+ members	37K last week	35M	10M		
Arts & Entertainment	Film	80 members	29K last week	4M	532K		
Products & Services	People	81 members	14K last week	7M	2M		
Science & Technology	TV	35 members	6K last week	8M	1M		
Society	Books	47 members	2K last week	29M	6M		
Special Interests	Location	50 members	2K last week	11M	973K		
Sports	Government	44 members	414 last week	492K	120K		
System	Business	100+ members	257 last week	2M	608K		

Google Refine

An open source power tool to fix, discover, experiment, connect and customize your data. [Learn more](#)

What is Freebase?

Learn what an entity graph is, what kind of information it contains, and why you should add your data! [Learn More](#)

Freebase for Developers

- powerful queryable API
- JavaScript-based hosting framework
- libraries for other languages

Linked Data in Libraries

- Strong uptake in recent years
 - Library of Congress (vocabularies)
 - German / Swedish / Hungarian National Library (authorities and/or catalogue data)
 - Europeana (metadata about ~4M items)
- W3C Library Linked Data Incubator Group

Overview

- Linked Data Vision and Goals
- **Enabling Technologies (by example)**
- Publishing and Consuming Linked Data
- The SciLink Project

Uniform Resource Identifiers (URIs)

- **Name** and **identify** things (resources)
- **Dereferencable** HTTP URIs

[http://arxiv.org/authors/
Haslhofer_B](http://arxiv.org/authors/Haslhofer_B)

[http://springerlink.com/
author/xyzafbea](http://springerlink.com/author/xyzafbea)

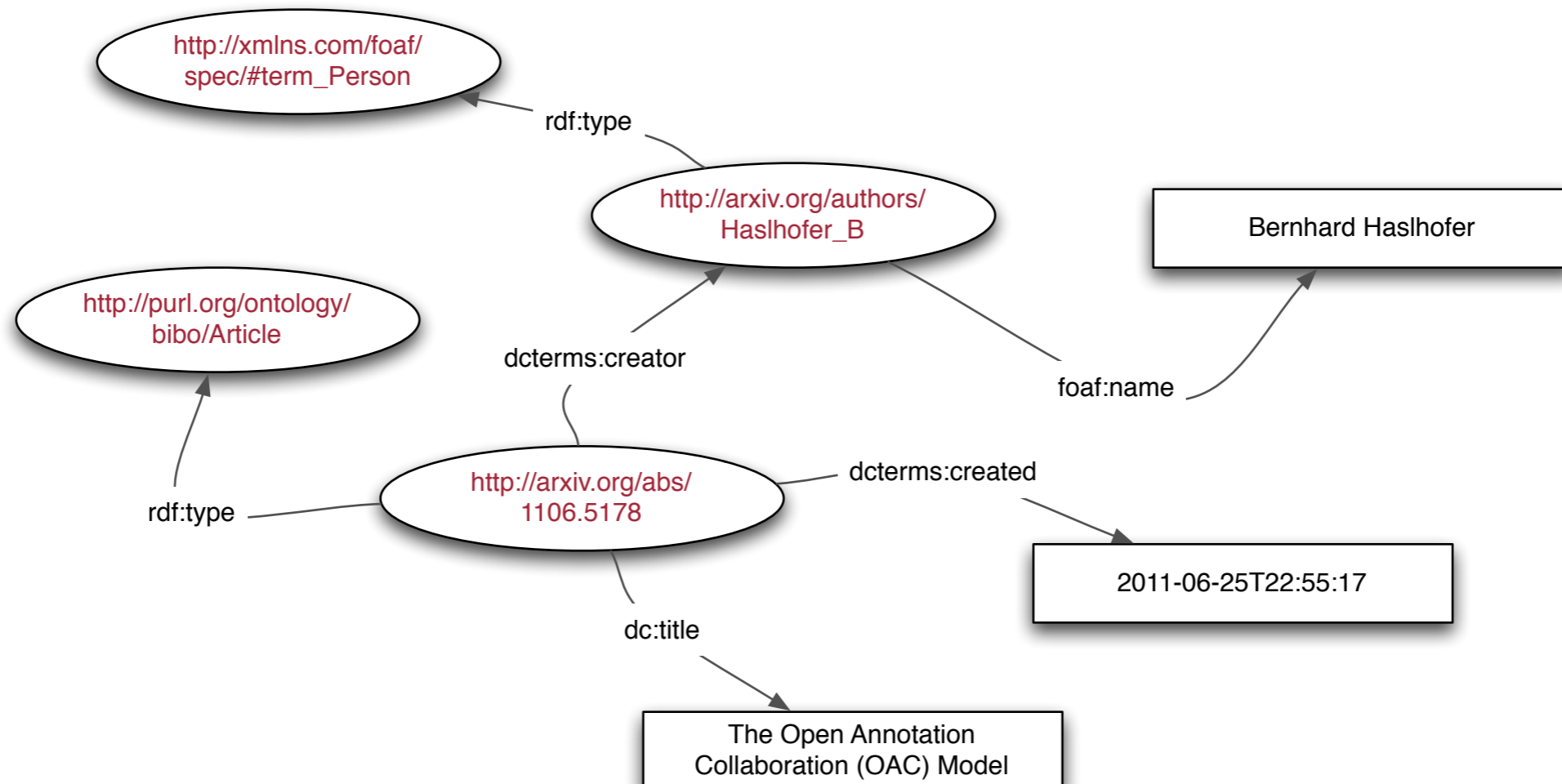
[http://arxiv.org/abs/
1106.5178](http://arxiv.org/abs/1106.5178)

[http://eprints.cs.univie.ac.at/
creator/xyzafbea](http://eprints.cs.univie.ac.at/creator/xyzafbea)

[http://eprints.cs.univie.ac.at/
2871](http://eprints.cs.univie.ac.at/2871)

Resource Description Framework (RDF)

- A **data model** for representing data on the Web
- Sets of **statements** (triples) form a **graph**



RDF Serialization (RDF/XML, Turtle, N-Triples, RDF/JSON)

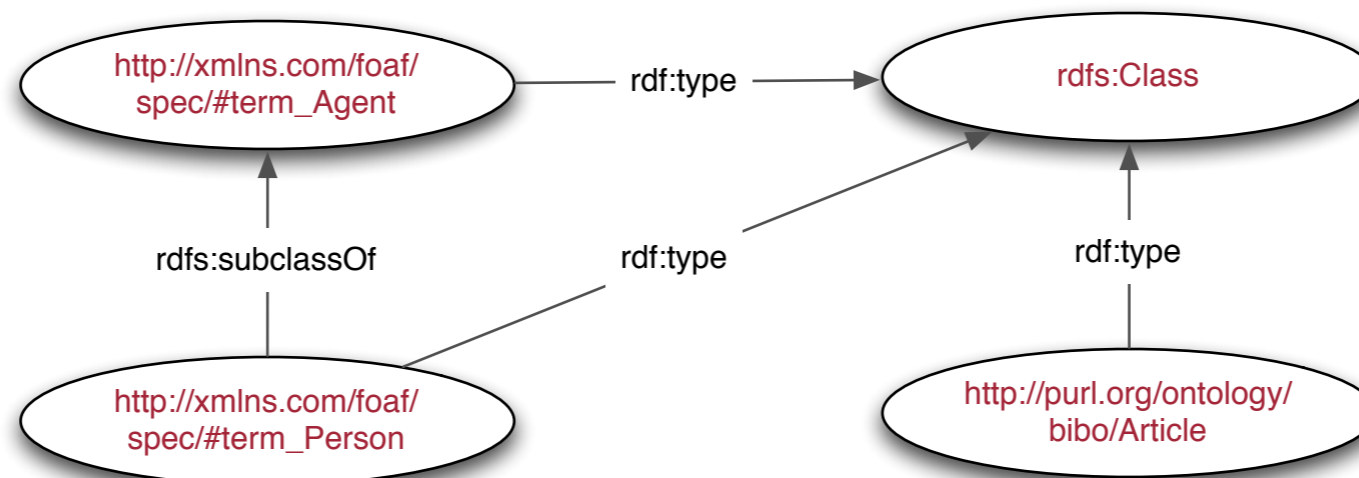
- **Data formats** for serializing RDF graphs
- Used to transfer RDF data between apps

```
<http://arxiv.org/abs/1106.5178>  
  dc:title "The Open Annotation Collaboration Model" .
```

...

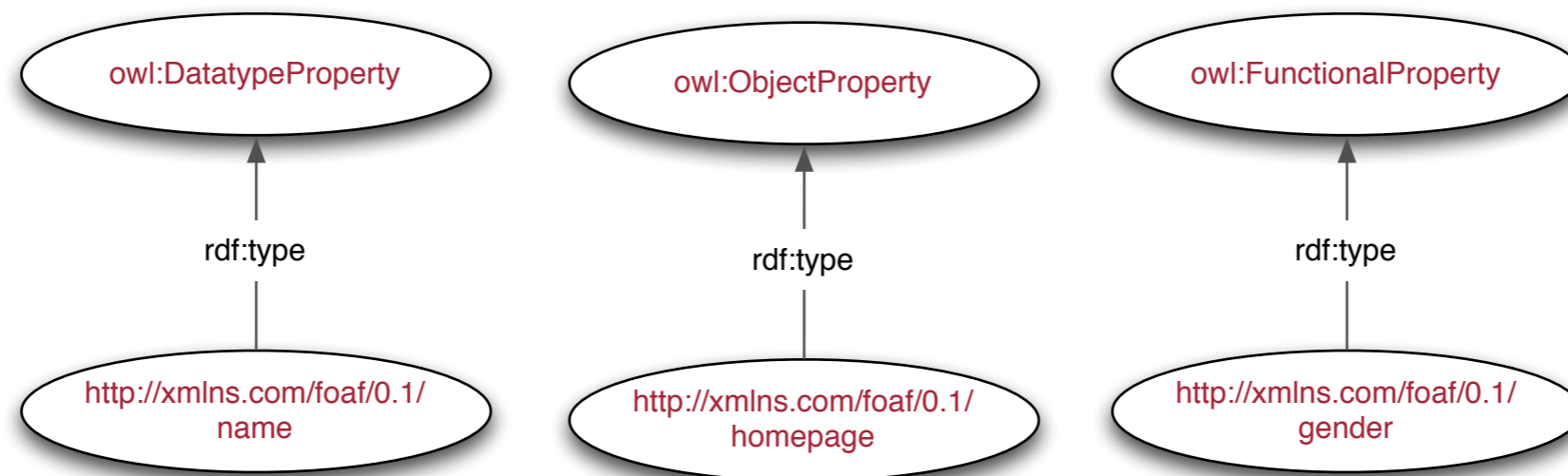
RDF Vocabulary Description Language (RDFS)

- A **language** for describing the syntax and semantics of **vocabularies** in a machine understandable way



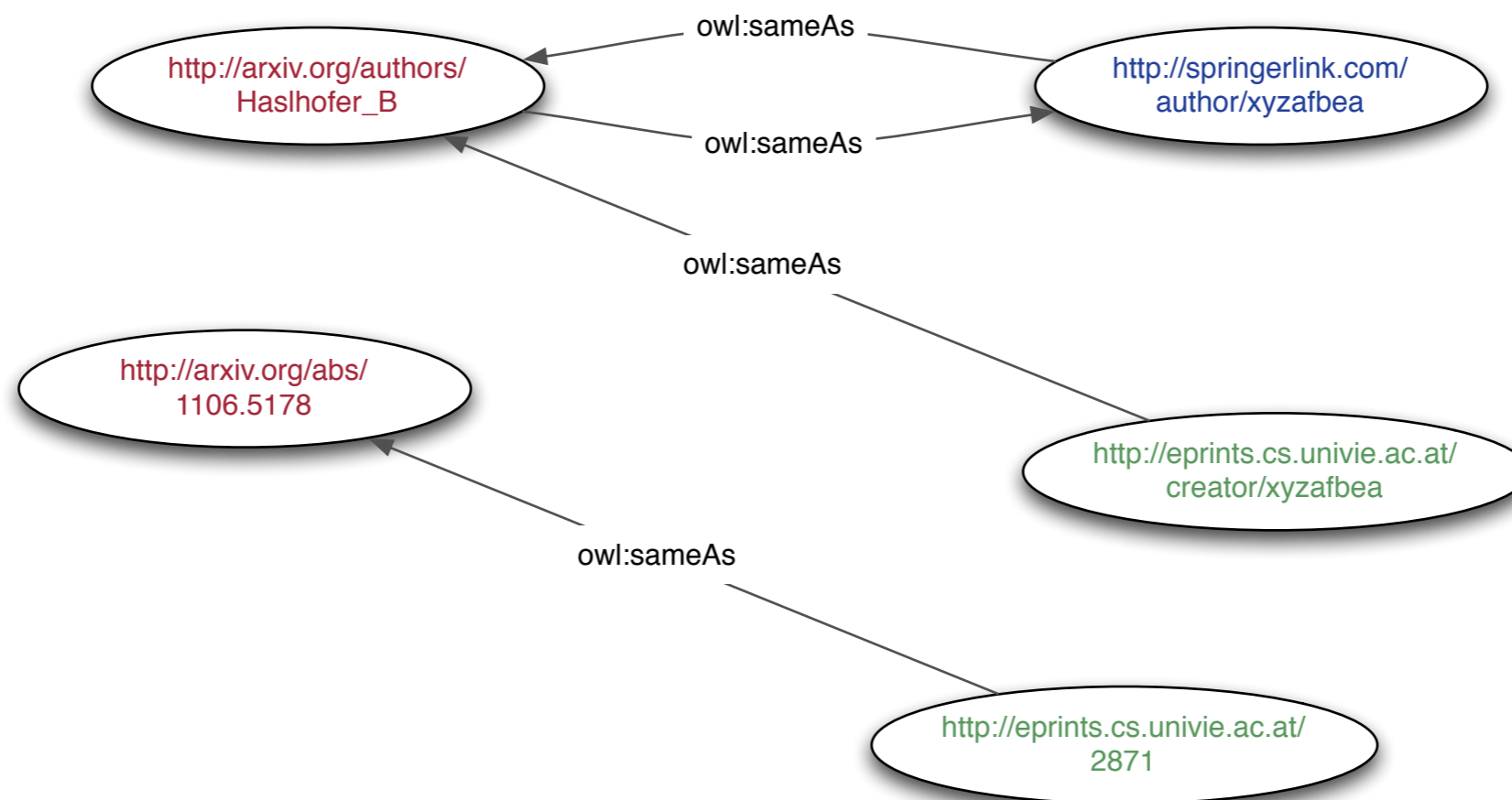
OWL Web Ontology Language

- A more expressive (formal) **language** for defining the syntax and semantics of **vocabularies**



OWL Web Ontology Language

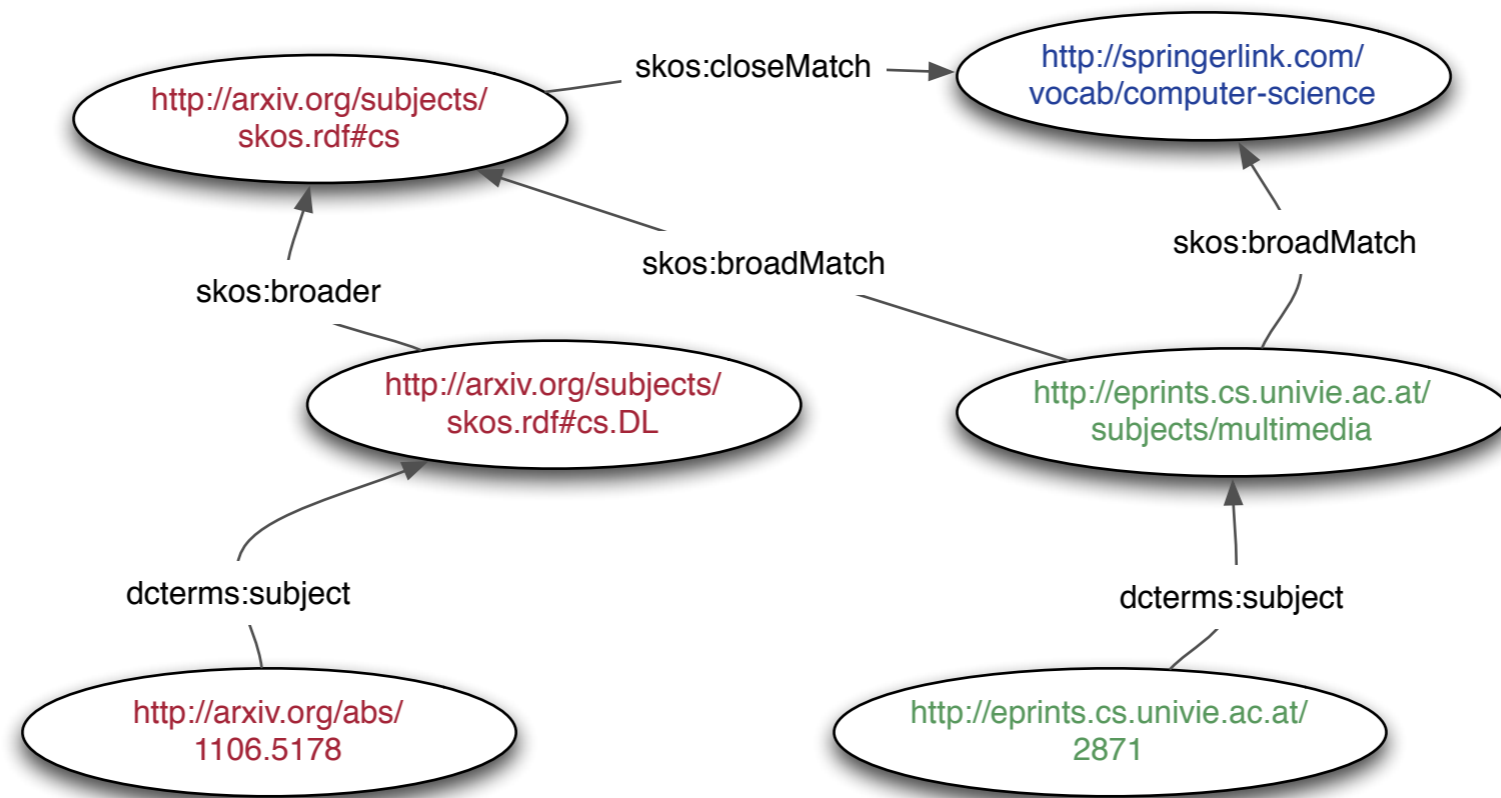
- Defines properties for **linking** resources



NOTE: owl:sameAs is NOT the only linking property.

Simple Knowledge Organization System (SKOS)

- A language for describing **controlled vocabularies** (taxonomies, thesauri, etc)



SPARQL

- A **query language** and **protocol** for accessing RDF data on the Web

```
SELECT DISTINCT ?x
```

```
WHERE { ?x skos:subject <http://arxiv.org/subjects/skos.rdf#cs> }
```

```
LIMIT 10
```

Overview

- Linked Data Vision and Goals
- Enabling Technologies (by example)
- Publishing and Consuming Linked Data
- The SciLink Project

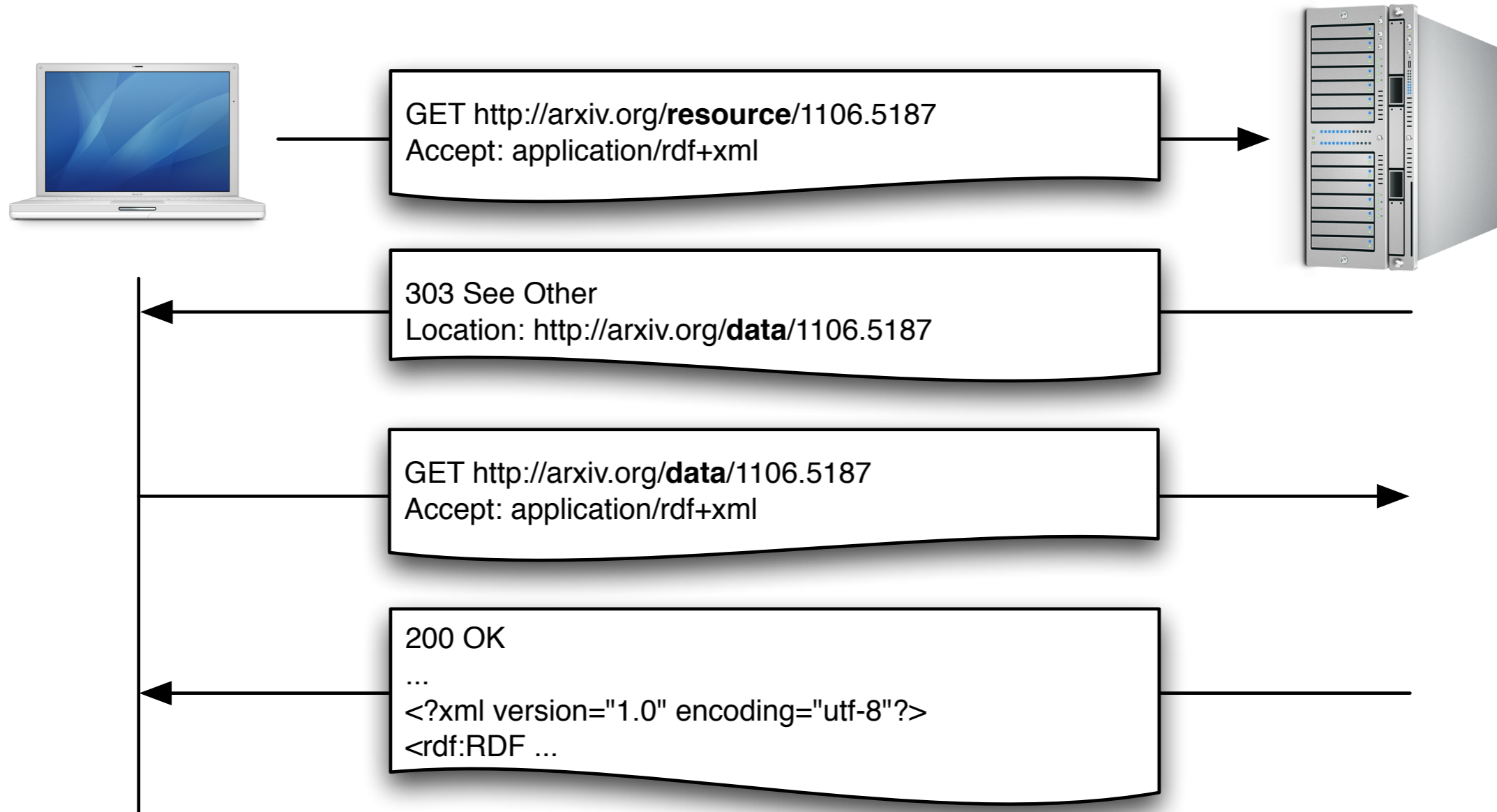
Information vs. Non-Information Resource

- Information Resource
 - web pages, images, pdf documents, etc.
 - all their essential characteristics can be conveyed in a message
 - e.g., <http://arxiv.org/pdf/1106.5178v1>
- Non-Information Resource
 - other things such as people, this meeting, concepts
 - their essence is not information
 - e.g., http://arxiv.org/authors/Haslhofer_B

Publishing Data

- Distinguish between non-information and information resource
- Sample NIR
 - <http://arxiv.org/resource/1106.5178>
- Sample IRs
 - <http://arxiv.org/abs/1106.5178> (HTML)
 - <http://arxiv.org/data/1106.5178> (RDF)

Publishing Data



Publishing Data



RDFa / Microformats / Microdata

- Mechanisms for embedding structured data in Web documents
- Define or use a set of attributes to **augmented** presentation-oriented (**HTML**) documents **with structured data**
- User agents can extract data from Web documents

RDFa Example

```
<div typeof="bibo:Article" xmlns:bibo="http://purl.org/ontology/bibo">  
  <div xmlns:dc="http://purl.org/dc/elements/1.1/">  
    <h1 property="dc:title">The Open Annotation Collaboration (OAC) Model</h1>  
    ....  
  </div>  
</div>
```

Microformats Example

```
<head profile="http://www.w3.org/2006/03/hcard">
```

```
...
```

```
</head>
```

```
<div class="vcard">
```

```
  <div class="fn">Bernhard Haslhofer</div>
```

```
</div>
```

RDFa vs. Microformats

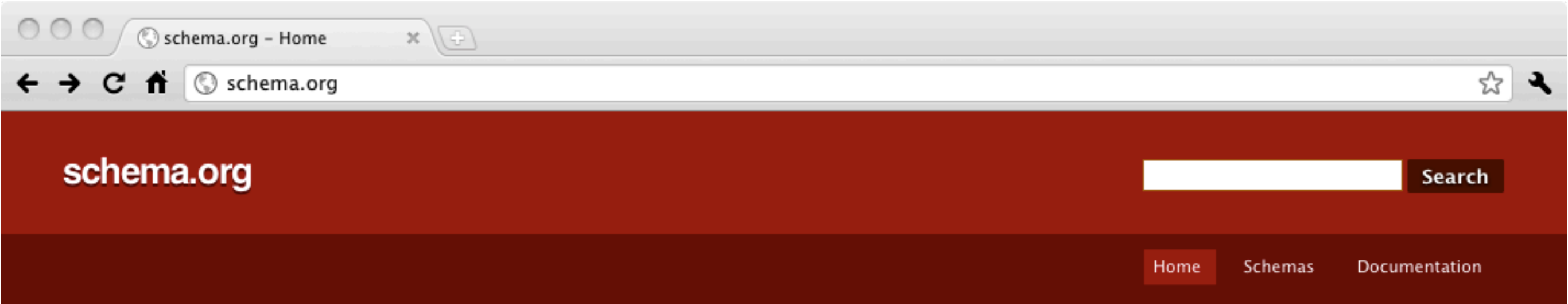
Microformats	RDFa
flat namespace	XML namespaces
support HTML4, XHTML 1.1, and HTML 5	support for XHTML 1.1, HTML 5
use latent HTML attributes	introduces new metadata attributes
vocabulary defined by one organization/community	open to any RDF-based vocabulary

Also see: http://evan.prodromou.name/RDFa_vs_microformats

Microdata

- A very young HTML proposition that extends Microformats and addresses its shortcomings

```
<div itemscope itemtype="http://schema.org/Article">
  <span itemprop="name">The Open Annotation Collaboration (OAC) Model</span>
  by <span itemprop="author">Bernhard Haslhofer</span>,
  <span itemprop="author">Rainer Simon</span>,
  <span itemprop="author">Robert Sanderson</span> and
  <span itemprop="author">Herbert Van De Sompel</span>
  This article has been tweeted 1 times and contains 0 user comments.
  <meta itemprop="interactionCount" content="UserTweets:1"/>
  <meta itemprop="interactionCount" content="UserComments:0"/>
</div>
```

What is Schema.org?

This site provides a collection of schemas, i.e., html tags, that webmasters can use to markup their pages in ways recognized by major search providers. Search engines including Bing, Google and Yahoo! rely on this markup to improve the display of search results, making it easier for people to find the right web pages.

Many sites are generated from structured data, which is often stored in databases. When this data is formatted into HTML, it becomes very difficult to recover the original structured data. Many applications, especially search engines, can benefit greatly from direct access to this structured data. On-page markup enables search engines to understand the information on web pages and provide richer search results in order to make it easier for users to find relevant information on the web. Markup can also enable new tools and applications that make use of the structure.

A shared markup vocabulary makes easier for webmasters to decide on a markup schema and get the maximum benefit for their efforts. So, in the spirit of sitemaps.org, Bing, Google and Yahoo! have come together to provide a shared collection of schemas that webmasters can use.

We invite you to [get started!](#)

Last Updated: 27 May 2011

[Terms and conditions](#)

Web Data Publishing Approaches

	Pro	Con
Linked Data Conneg IR / NIR	Separates data / presentation markup Web Architecture	Technical complexity Communication overhead
RDFa / Microformats / Microdata	Easy to implement Search Engines	Mixes data / presentation markup Becomes messy

Consuming Linked Data

```
curl -H "Accept: application/rdf+xml" http://arxiv.org/data/1105.5178
```

```
rapper http://arxiv.org/resource/1105.5178
```

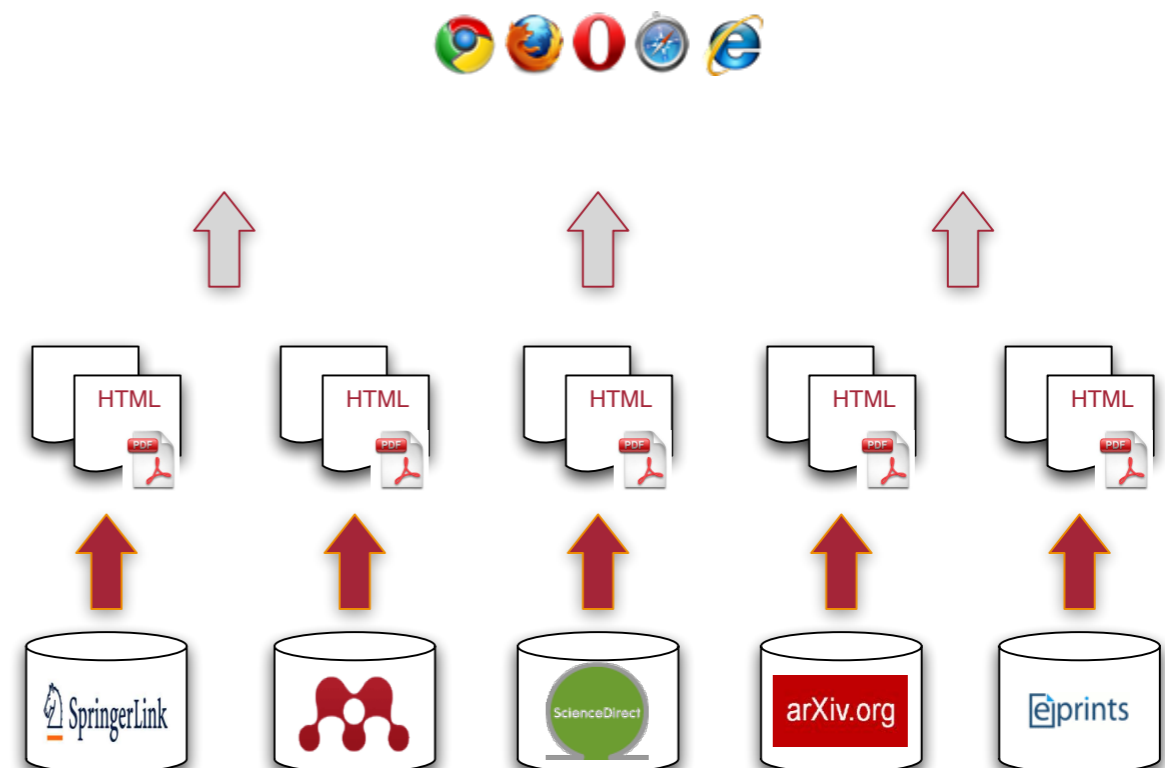
```
... any RDF library
```

Overview

- Linked Data Vision and Goals
- Enabling Technologies (by example)
- Publishing and Consuming Linked Data
- **The SciLink Project**

Motivation

- Current research articles
 - are mostly PDFs (a kind of BLOB)
 - point to supplemental or related information by textual references, DOIs, or embedded hypertext links



Motivation

- But a publication is more:
 - data sets, conference presentations
 - blog / wiki entries, software libraries, etc
- There is related information on the Web (e.g., Wikipedia)

Overall Goal

Publication Document (PDF)

Genome-wide analysis of Notch signaling in *Drosophila* by transgenic RNAi

Jennifer Mummery-Widmer,

Here we report the use of a library of ***Drosophila*** strains...we identified 6 new genes involved in **asymmetric cell division**.

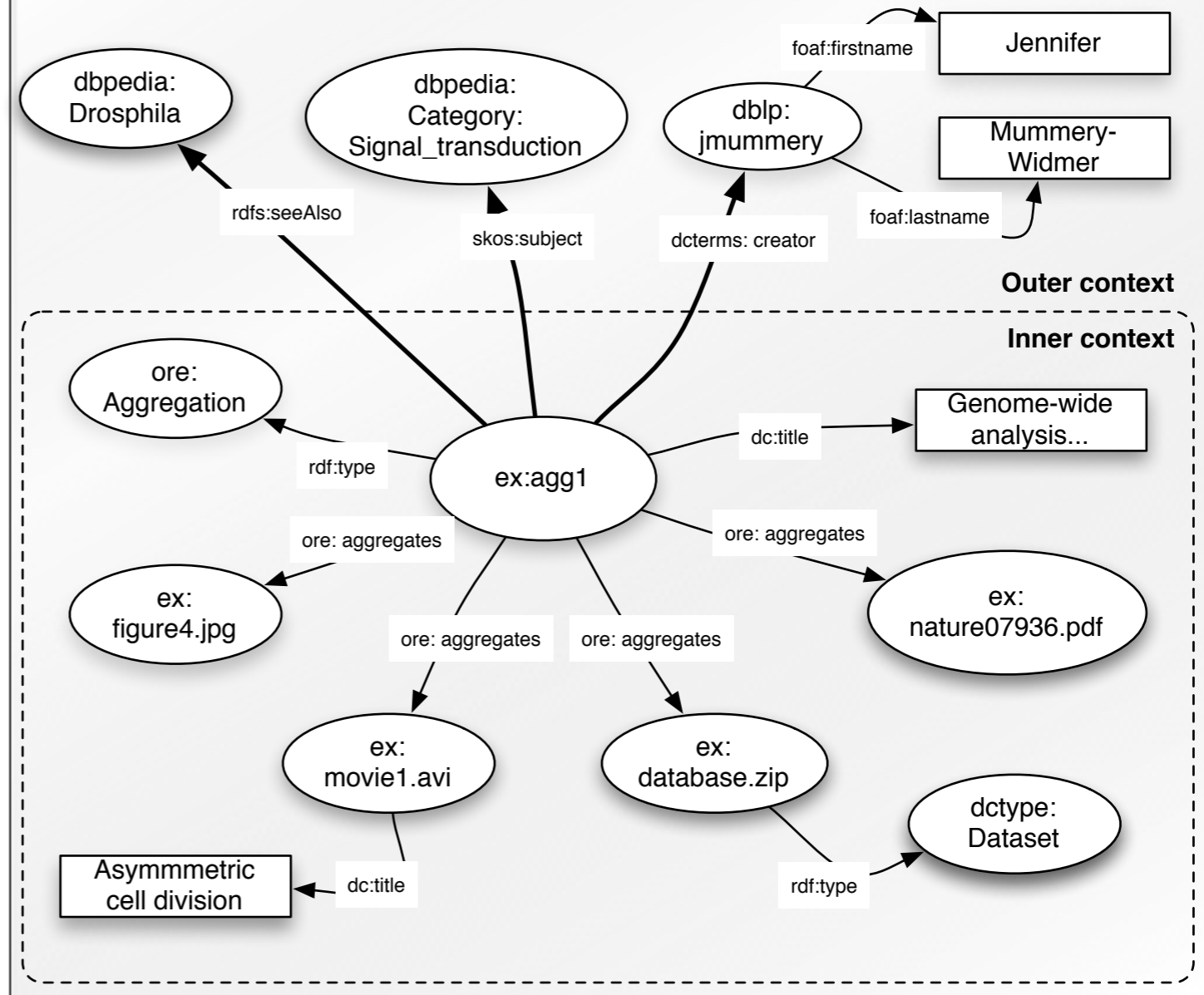
A total of 20,262 transgenic RNAi lines were screened...Ten flies were recorded in an online **database**...



Figure 4. An interaction network for Notch signalling

<http://www.nature.com/nature/journal/vaop/ncurrent/pdf/nature07936.pdf>

Aggregation for the Document



Methodology

- **Aggregation/Data Extraction:** create aggregations from available information (PDFs, metadata)
- **Analysis:** combine aggregations and learn about established research and publication practices in different areas (focus on linking)
- **Tool design and Implementation**
 - link discovery
 - resource synchronization / link maintenance
 - demonstrator applications operating on data graph

Aggregation / Data Extraction

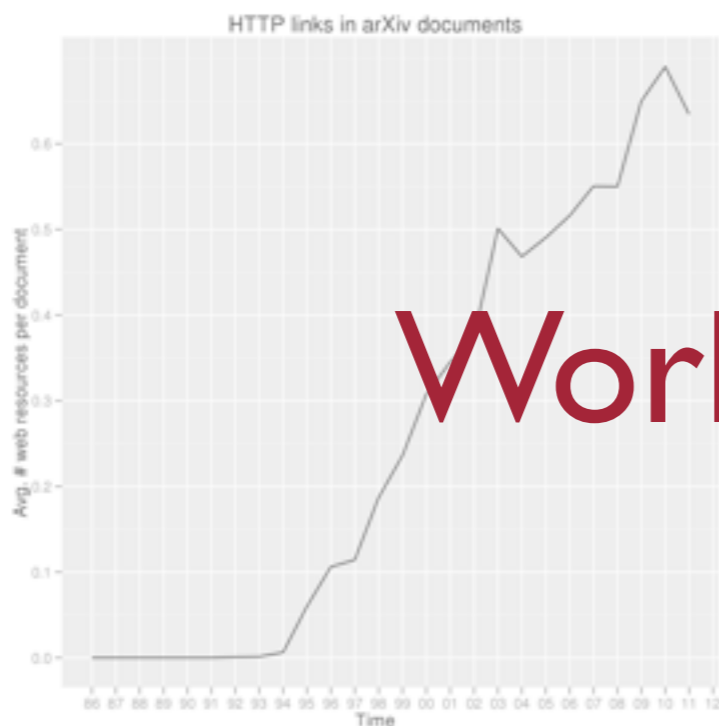
- **scilink-extract** library
 - takes PDFs and metadata as input
 - creates ORE Aggregations / CSV files
 - currently implemented for arXiv.org
 - <https://github.com/behass/scilink-extract>

Analysis

- How did HTTP linkage evolve in the past?
- Where in their papers do scholars make use of HTTP links?
- What are the linking to?
- Are the links still operational?
-

Status

- Extraction for arxiv.org finished (~ 700K) PDFs
- We need access to further corpora covering different research areas !!!



Work in Progress

