



W3C HCLS ISWC Workshop
Semantic Web Applications in Scientific
Discourse, Washington DC



26th October 2009

Enabling Semantic Publication and Integration of Scientific Information



David Shotton

Image BioInformatics Research Group

Department of Zoology

University of Oxford, UK

<http://ibrg.zoo.ox.ac.uk>

e-mail: david.shotton@zoo.ox.ac.uk



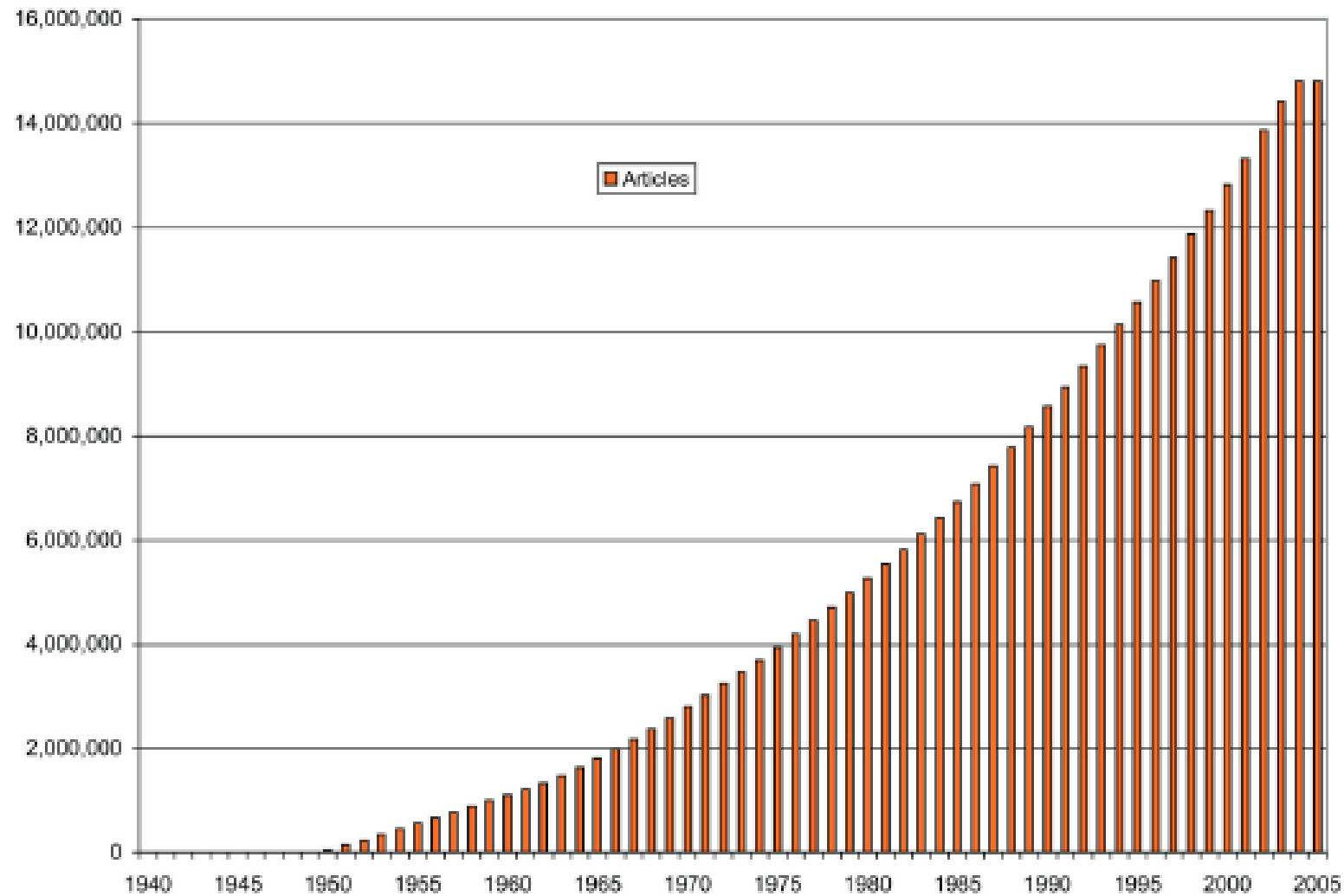
Outline

- Semantic Publishing - exemplar semantic enhancements of a research article
 - CiTO, the Citation Typing Ontology
 - MIIDI, Minimal Information for an Infectious Disease Investigation
 - OpenFlyData, a semantic data web for *Drosophila* gene expression data
-
- N.B. This full version of the presentation contains additional slides I did not have time to present during the workshop, that give more details about our work and the work of others in this area

Exemplar semantic enhancements to an article from
PLoS Neglected Tropical Diseases

<http://dx.doi.org/10.1371/journal.pntd.0000228.x001>

The number of research articles is growing . . .

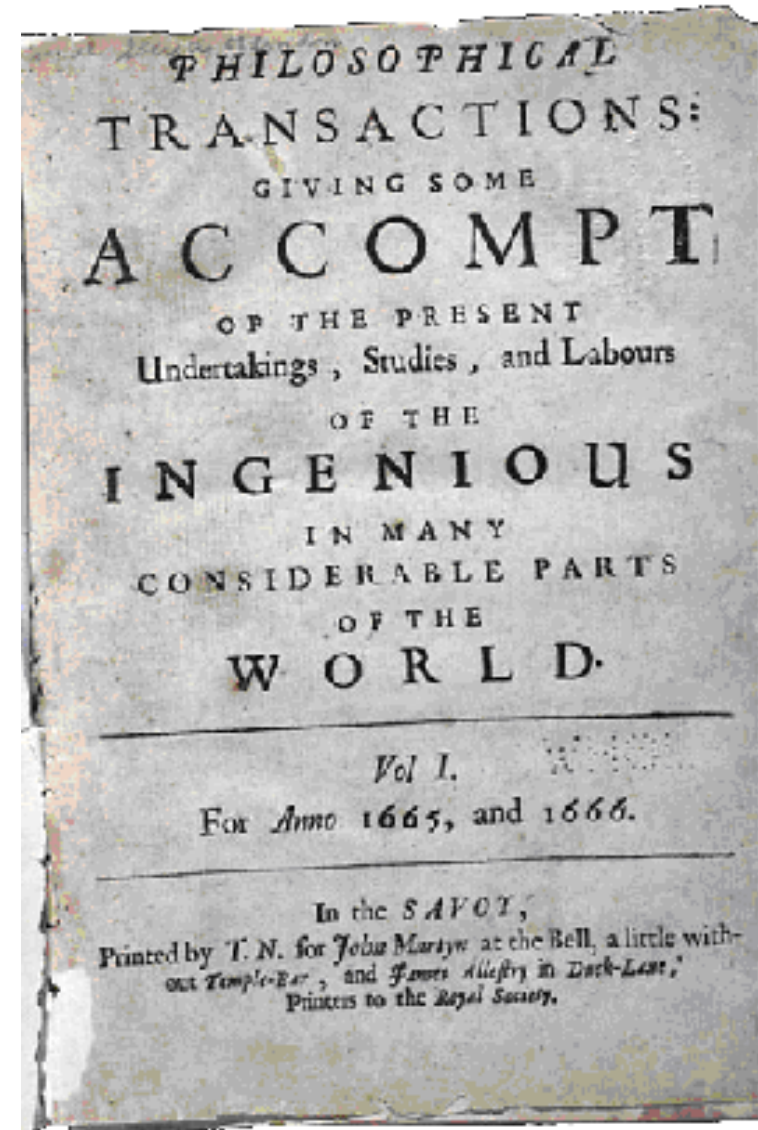


DOI: 10.1371/journal.pbio.0030065.g001

Figure 1. Medline Article Deluge

... but research publishing has changed very little

- We still have a **linear narrative**, with references
- The norm is to publish the online journal article as a static file mimicking the printed page
- This is **totally antithetical** to the spirit of the Web, and ignores its great potential
- Rather, we need **lively** journal content
 - Semantic mark-up of text
 - Interactive figures
 - Links between papers and datasets
 - Actionable numerical data




First a definition: What is Semantic Publishing?

The use of simple Web and Semantic Web technologies

- to enhance the meaning of on-line published research articles
- to provide access to published data in actionable form
- to link articles with their cited references
- to link articles to the research datasets that underpin them
- to provide machine-readable summaries of an article's content
- to facilitate integration of semantically related scientific information from heterogeneous distributed resources

so that data, information and knowledge can more easily be found,
extracted, combined and reused

The exemplar article we chose to semantically 'enliven'

**PLOS** **NEGLECTED
TROPICAL DISEASES**
a peer-reviewed open-access journal published by the Public Library of Science


Home Browse Articles About For Readers For Authors and Reviewers

RESEARCH ARTICLE

OPEN  ACCESS

Impact of Environment and Social Gradient on *Leptospira* Infection in Urban Slums

Renato B. Reis^{1#}, Guilherme S. Ribeiro^{1#}, Ridalva D. M. Felzemburgh¹, Francisco S. Santana^{1,2}, Sharif Mohr¹, Astrid X. T. O. Melendez¹, Adriano Queiroz¹, Andréia C. Santos¹, Romy R. Ravines³, Wagner S. Tassinari^{3,4}, Marília S. Carvalho³, Mitermayer G. Reis¹, Albert I. Ko^{1,5*}

 Download Article XML
 Download Article PDF
 Download Citation
 E-mail this Article
 Order Reprints
 Print this Article

Citation: PLoS Negl Trop Dis 2(4): e228. 2008 doi:10.1371/journal.pntd.0000228
Received: January 22, 2008; **Accepted:** March 27, 2008; **Published:** April 23, 2008

Features of the original *PLoS NTD* article

Good

- Article published as **XML** under a **Creative Commons attribution license**
- Internal **navigation links** to individual sections of the paper
- The figures and the table all have unique **DOIs**, making them citable
- The article contained a **rich variety of data types**
 - geospatial, disease incidence, serological assay, and questionnaire presented in **formats amenable to semantic enrichments**
 - maps, bar charts, tables, graphs and scatter plots

Poor

- **No direct links to the cited articles** from items in reference list
- **No hyperlinks** to other useful sites
- Static figures and table - **no interactivity**
- While figures and table can be downloaded, they can only be so as **images** !
 - **The numerical data are not directly available in actionable form**

Our motivation for semantic enhancement

- Our purpose was to create a compelling **existence proof** of the possibilities of semantic publication, using a single exemplar research article
- We first scoped possible enhancements, identifying those that were
 - **easy, moderately difficult** and **hard** to implement
 - **essential, desirable** and **peripheral** to our primary purpose
- Within the limited resources available for this unfunded project, we then implemented on the *PLoS NTD* article:
 - all those enhancements that were **easy**,
 - all those that we judged to be **essential**, whatever the difficulty
 - most of those that were **desirable but moderately difficult**
- These can be seen at <http://dx.doi.org/10.1371/journal.pntd.0000228.x001>

The enhanced *PLoS NTDs* paper by Reis *et al.* (2008):

<http://dx.doi.org/10.1371/journal.pntd.0000228.x001>

[turn all highlighting on](#) [date](#) [disease](#) [habitat](#) [institution](#) [organism](#) [person](#) [place](#) [protein](#) [taxon](#)

[Top](#) | [Abstract](#) | [Author Summary](#) | [Introduction](#) | [Methods](#) | [Results](#) | [Discussion](#) | [Supporting Information](#) | [Acknowledgements](#) | [References](#) | [Data Fusion Supplements](#)

SEMANTICALLY ENHANCED VERSION OF A RESEARCH ARTICLE FROM PLOS NEGLECTED TROPICAL DISEASES

Impact of Environment and Social Gradient on *Leptospira* Infection in Urban Slums

document summary

Renato B. Reis ^{1#}, Guilherme S. Ribeiro ^{1#}, Ridalva D. M. Felzemburgh ¹, Francisco S. Santana ^{1, 2}, Sharif Mohr ¹, Astrid X. T. O. Melendez ¹, Adriano Queiroz ¹, Andréia C. Santos ¹, Romy R. Ravines ³, Wagner S. Tassinari ^{3, 4}, Marília S. Carvalho ³, Mitermayer G. Reis ¹, Albert I. Ko ^{1, 5*}

¹ Centro de Pesquisas Gonçalo Moniz, Fundação Oswaldo Cruz, Ministério da Saúde, Salvador, Brazil ² Secretária Estadual de Saúde da Bahia, Salvador, Brazil ³ Escola Nacional da Saúde Pública, Fundação Oswaldo Cruz, Ministério da Saúde, Rio de Janeiro, Brazil ⁴ Universidade Federal Rural do Rio de Janeiro, Rio de Janeiro, Brazil ⁵ Division of International Medicine and Infectious Diseases, Weill Medical College of Cornell University, New York, New York, United States of America

Abstract

Background

Leptospirosis has become an urban health problem as slum settlements have expanded worldwide. Efforts to identify interventions for urban leptospirosis have been hampered by the lack of population-based information on *Leptospira* transmission determinants. The aim of the study was to estimate the prevalence of *Leptospira* infection and identify risk factors for infection in the urban slum setting.

Methods and Findings

We performed a community-based survey of 3,171 slum residents from Salvador, Brazil. *Leptospira* agglutinating antibodies were measured as a marker for prior infection. Poisson regression models evaluated the association between the presence of *Leptospira* antibodies and environmental attributes obtained from Geographical Information System surveys and indicators of socioeconomic status and exposures for individuals. Overall prevalence of *Leptospira* antibodies was 15.4% (95% confidence interval [CI], 14.0–16.8). Households of subjects with *Leptospira* antibodies clustered in squatter areas at the bottom of valleys. The risk of acquiring *Leptospira*

Our semantic enhancements to this *PLoS NTD* paper

Better integration of the paper into the Web

- Provision of hyperlinks to relevant Web sites
- Live DOI links to full text of cited papers
- Machine-readable metadata and reference files (RDF N3 and RDFa)

Additions to the paper

- The datasets in the table and figures downloadable in actionable form
- Semantic mark-up of terms in the text, with links to authorities
- Enhanced Portuguese Abstract; Re-orderable reference list
- Interactive figures, and the Supporting Claims Tooltip (exemplars)

Analysis of the content of the paper

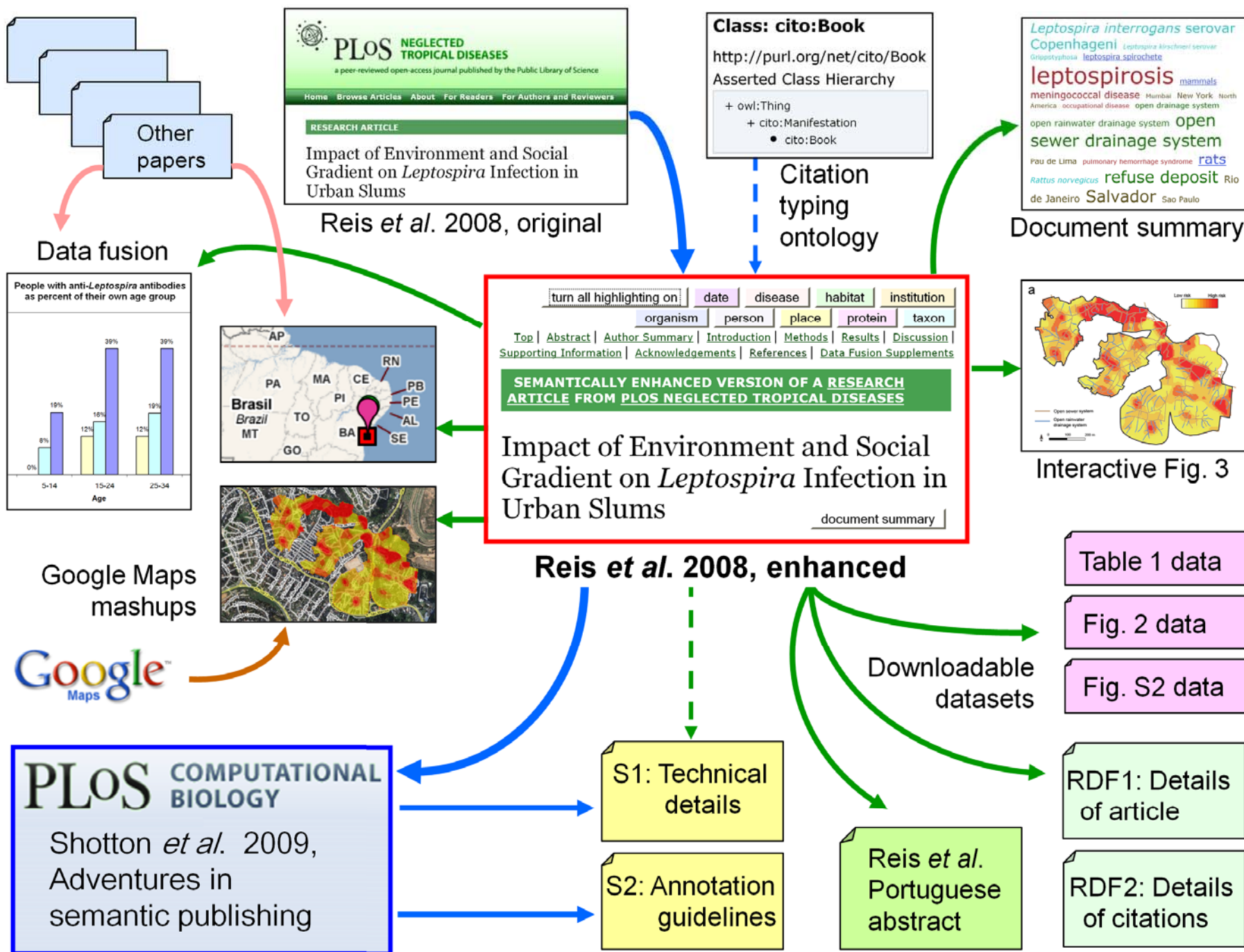
- Document summarization, including tag cloud and study summary
- Citation analysis, both of citation frequency and citation type

Data fusion (mashup) services

- Geo-temporal mashups with Google Maps
- Integration with relevant disease incidence data in other publications

Techniques used and effort involved

- What we did to the *PLoS NTD* paper is **not rocket science**. It involved
 - additions to the **XHTML** in which the paper was obtained from PLoS
 - application of standard **XHTML markup** for hyperlinks, etc.
 - standard use of **CSS** (Cascading Style Sheets) for format styling
 - use of simple **JavaScript**, and of the **Yahoo! User Interface (YUI) Library** of utilities and controls, to create interactivity
 - use of the **Google Maps API** to create geospatial data fusions
 - metadata in **RDF**, the W3C standard for encoding linked Web data
- The work was undertaken in an eight-week period in summer 2008 by one developer (Katie Portwin), with myself steering the development, and other members of my group occasionally providing ideas and feedback
- Most of that time was spent figuring out what we wanted to do, and then experimenting with how best to achieve our goals
- Knowing what we know now, the work could be done much more quickly
- We then described what we did in a *PLoS Computational Biology* paper



Criticism of our work – we only went part of the way

- The only serious constructive criticism of our work has come from Rod Page (<http://iphylo.blogspot.com/2009/04/semantic-publishing-towards-real.html>)
- He says it would have been better if we had provided **more RDF metadata**, e.g. by linked to DBpedia URIs than directly to Web pages, enabling our enhanced paper to become part of the *Linked Data* ecosystem
 - "I think that **real** integration by linking requires that the resources being linked are both computer and human readable, and that **both** resources know about the link. This would create much more powerful 'semantically enhanced' publications."
- Of course, I agree. We intend to enhance this article further, and invite others to participate
- Matthias Samwald has already created some exemplar <a>Tags

```
<span class="atag_atags">aTags: <a rel="sioc:topic"
href="http://dbpedia.org/resource/Leptospirosis">
<span about="http://dbpedia.org/resource/Leptospirosis"
property="rdfs:label">Leptospirosis</span></a>
```

Taking semantic publishing to prime time

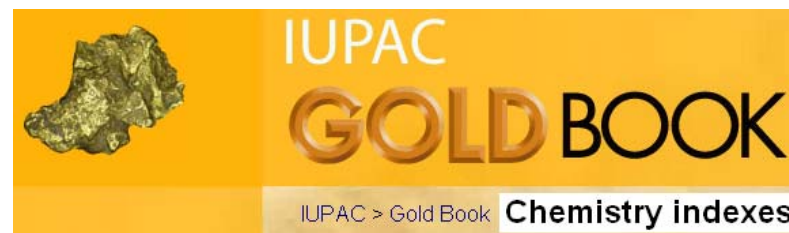
- Semantic authoring tools are coming
 - Microsoft's ontology-based plug-in for Word 2007
- Several journals are incorporating semantic enhancements
 - Royal Society of Chemistry's *Project Prospect*
 - *The Article of the Future* from Cell Press
- Post-publication semantic enhancement tools exist
 - *REFLECT*, winner of the Elsevier Grand Challenge, marks up gene, protein and chemical names
 - *Science Commons* has a text annotation service using *Whatizit* that outputs *aTags*
 - *ISACreator* is an annotation tool that permits selection of ontologies

Tools to assist pre-publication semantic enhancement

- Microsoft Research have released an open source **plug-in for MS Word 2007**, developed in collaboration with Lynn Fink in Philip Bourne's lab at UC San Diego, that permits **semantic mark-up of text**
 - it inserts XML tags based on selected domain ontologies
 - e.g. the word "**disease**", tagged using the Human Disease ontology

Semantic mark-up by journal editors: setting the standard

- The Royal Society of Chemistry's award-winning *Prospect Project* is being used to enhance many journals
 - e.g. *Molecular BioSystems* and *Integrative Biology*
 - Terms are marked up and linked to external resources
 - Chemical names
 - Gene Ontology terms
 - IUPAC Gold Book terms
- "Project Prospect is fantastic. I've just seen the future of the journal"
- Ed Pentz, Executive Director of CrossRef



- **SMILES:** [O-][n+]1ccccc1S
- **InChI:** InChI=1/C5H5NOS/c7-6-4-2-1-3-5(6)8/h1-4,8H

The *Article of the Future* prototype from Cell Press

(<http://beta.cell.com/index.php/2009/07/article-of-the-future/>)

... has a "visual abstract" of the paper's content

Cell **Article Prototype #1** [View Prototype #2](#) [Tell Us What You Think](#) [Share / Save](#)

May 2, 2008 - Volume 133, Issue 3, pp. 462-474 [PDF \(1,758 KB\)](#)

A Dynamic Pathway for Calcium-Independent Activation of CaMKII by Methionine Oxidation

Jeffrey R. Erickson¹, Mei-ling A. Joiner¹, Xiaoqun Guan¹, William Kutschke¹, Jinying Yang¹, Carmine V. Oddis⁵, Ryan K. Bartlett⁶, John S. Lowe¹, Susan E. O'Donnell², Nukhet Aykin-Burns³, Matthew C. Zimmerman³, Kathy Zimmerman⁹, Amy-Joan L. Ham^{7,8}, Robert M. Weiss^{1,9}, Douglas R. Spitz³, Madeline A. Shea², Roger J. Colbran⁷, Peter J. Mohler^{1,4,*} [Affiliations](#)

[Abstract](#) [Introduction](#) [Results](#) [Discussion](#) [Experimental Procedures](#) [Figures \(8+\)](#) [References \(51\)](#) [Authors](#) [Comments \(3\)](#) [Acknowledgements](#)

Article Highlights

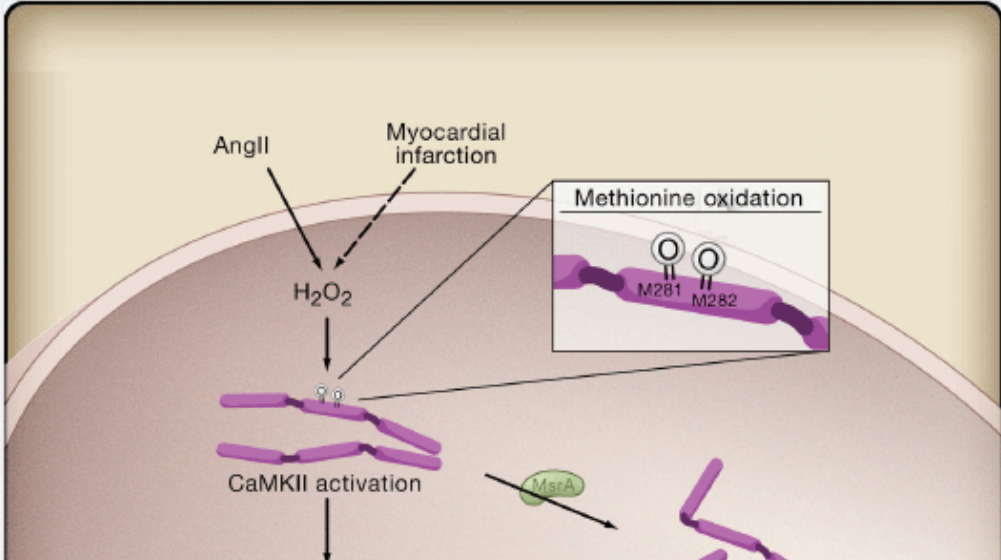
- Oxidation of methionine residues activates CaMKII
- Angiotensin II induces CaMKII oxidation leading to cardiomyocyte death
- CaMKII methionine oxidation is reversed by MsrA
- Elevated CaMKII oxidation impairs heart function and worsens ischemic injury

Author Interview

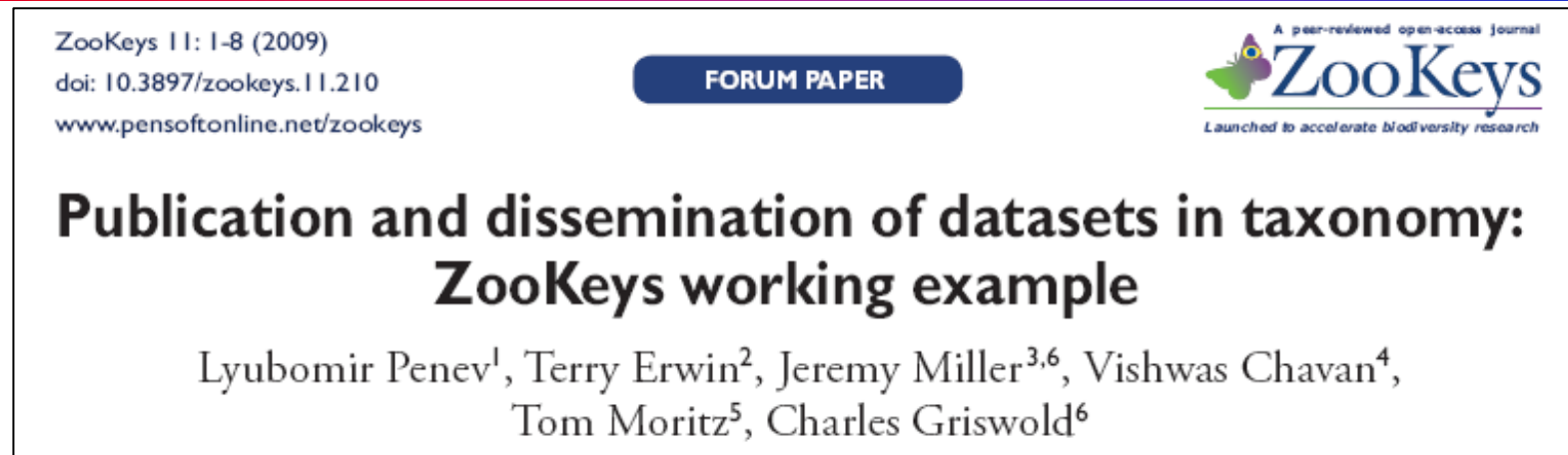
[«](#) [»](#) [»»](#) [🗨️](#) [🎧](#)

Abstract

Calcium/calmodulin (Ca²⁺/CaM)-dependent protein kinase II (CaMKII) couples increases in cellular Ca²⁺ to fundamental responses



Proper dataset publication by a journal publisher



- A new concept for data publication
- The biodiversity data are published as a dataset under a separate DOI
- The dataset is separately discoverable and accessible through the GBIF data portal (Global Biodiversity Information Facility; <http://data.gbif.org>)
- The dataset is also published as a KML (Keyhole Markup Language) file under a distinct DOI, to visualize species locations using Google Earth
- All new taxa are registered at ZooBank during the publication process
- All new taxa are provided to the Encyclopedia of Life through XML mark up on the day of publication

Post-publication semantic enhancement - REFLECT



- Created at the European Molecular Biology Laboratory by Sean O'Donoghue and his team, and available at <http://reflect.embl.de/>, Reflect was the **winning entry of the Elsevier Grand Challenge**
- Reflect uses a web service to send HTML text from any URL to the Heidelberg Reflect server, where simple text mining is used for identification of the names of genes, proteins and small molecules
- After matching to dictionary entries held in memory, the entities are semantically marked up, with links to appropriate databases and ontologies
- Clicking on an annotated element displays a pop-up window that gives information about the term, and allows the user to link quickly to more detailed information

REFLECT markup of genes and proteins (<http://reflect.embl.de/>)

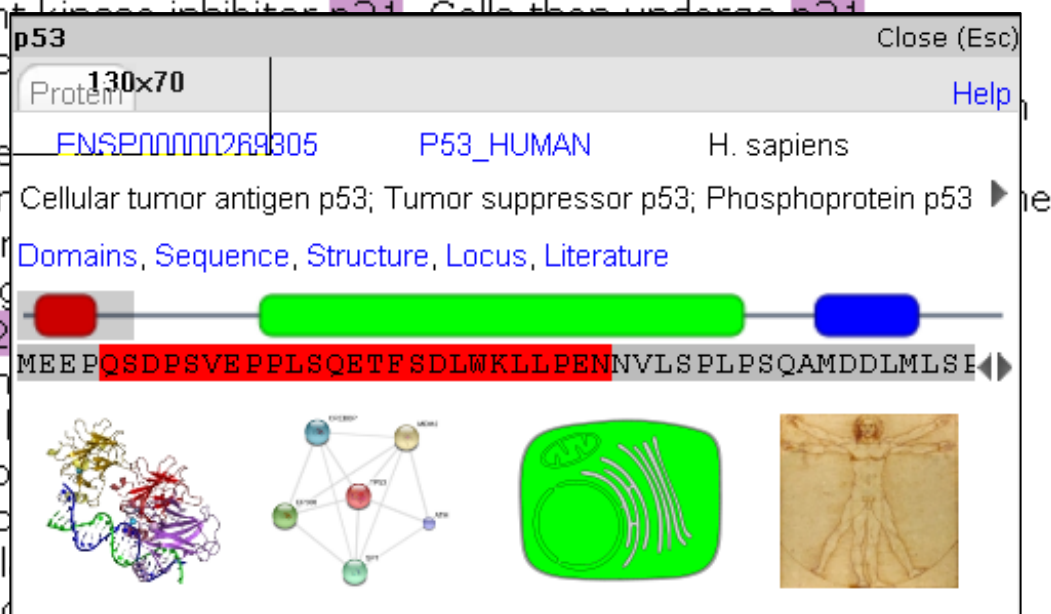
1: [J Exp Clin Cancer Res.](#) 2004 Dec;23(4):625-31.

P21 and Bax expression in cutaneous malignant melanomas: correlation with histologic prognostic parameters.

[Poyraz A](#), [Akyürek N](#), [Gönül II](#), [Erdem O](#).

Dept. of Pathology, Gazi University Medical School, Ankara, Turkey. aylarpoyraz@yahoo.com

In response to DNA damage, p53 accumulates and regulates expression of several genes, including cyclin-dependent kinase inhibitors p21. Cells that undergo p21 dependent cell cycle arrest, which is a death promoter member of the p53 pathway and commitment to programmed cell death. p53 is a major factor in predicting prognosis for cutaneous malignant melanoma. In a study of the role of cyclin dependent kinase inhibitors in cutaneous malignant melanoma we investigated the expression of p21 in 100 cases. We observed that significant high p21 expression was observed in thick tumours (p=0.006). Similarly, thick tumours expressed significantly high p21 (p=0.006). Results suggested that the role of p21 in cutaneous skin cancers and abnormal regulation of p53 occurred in the development and progression of cutaneous malignant melanoma. To understand the role of bax expression in thick malignant melanomas and invasion biology, comparative analytic studies with other apoptosis regulators are needed.



... but there are also Semantic Web equivalents

Science Commons text annotation service

powered by [EBI Whatizit](#)

This service recognizes entities and relations from biomedical texts and returns documents in Semantic Web formats (OWL, RDFa). You can find more information on the [documentation page](#).

Choose input

- ☐ PubMed query [i](#)
☒ PubMed ID (PMID) [i](#)

What should be recognized? [i](#)

Chemicals and proteins [v](#)

PubMed ID: 15743033

Output format

aTag [v](#)

[Submit PubMed ID](#)

- Developed by Matthias Samwald and Alan Ruttenberg
 - <http://whatizit.neurocommons.org/>

Science Commons text annotation service results

pmid:15743033 [HTML] [XML]

A Poyraz, N Akyürek, I I Gönül, O Erdem

Journal of experimental & clinical cancer research : CR

Dec 2004

P21 and Bax expression in cutaneous malignant melanomas: correlation with histologic prognostic parameters <a>.

In response to DNA damage, p53 accumulates and regulates expression of several genes, including cyclin-dependent kinase inhibitor p21 <a>. Cells then undergo p21 dependent cell cycle arrest, which allows DNA damage repair and apoptosis <a>. Bax is a death promoter member of the bcl-2 family which plays a central role in the regulation and commitment to programmed cell death <a>. Breslow thickness is the most important factor in predicting prognosis for cutaneous malignant melanoma. In order to define the role of cyclin dependent kinase inhibitors and apoptosis regulators in invasion of malignant melanoma we investigated the expression of p21 and bax proteins <a>. We observed that significant high p21 expression was associated with increasing Breslow thickness (Spearman correlation analysis, p=0.01) <a>. Additionally, Clark level I and II tumours expressed significantly lower p21 positivity than Clark level III, IV and V (p=0.006) <a>. Similarly, thick tumors showed a higher bax expression (p=0.012) <a>. Our results suggested that the role of p21 expression is more complicated in melanocytic skin cancers and abnormal regulation or abnormal function of cell cycle regulators occurred in the development and progression of malignant melanoma <a>. In order to understand the role of bax expression in thick malignant melanomas and invasion biology, comparative analytic studies with other apoptosis regulators are needed <a>.

- More complete mark-up, and results in RDFa rather than XML

Ideally, one also wants the ability to choose ontologies . . .

ISACREATOR <http://isatab.sf.net>

ONTOLOGY SEARCH

ontologylookup recenthistory

☐ recommended search ☒ all ontologies

✓ term search
accession

- + PATO:rate
- + BTO:blood plasma
- + OBI:labeling
- + CHEBI:chemical compound
- + OBI:mRNA extraction
- + UO:microliter
- + CHEBI:rapamycin
- + NEWT:Mus musculus (Mouse)
- + OBI:protein extraction

selected term(s) NEWT:Mus musculus (Mouse) close ok

Depending on your configuration (set via the ISAconfigurator), certain fields prompt a pop-up browser to search and select terms from ontologies that are accessed in real time via the [Ontology Lookup Service](#) and [BioPortal](#).

Post-publication enhancement - Citations in Context

Citation Sensitive In-Browser Summarisation of Cited Documents: A Research Prototype for Browsing Academic and Scholarly Literature

Stephen Wan
CSIRO / ICT Centre

Cécile Paris
CSIRO / ICT Centre

- This **finalist in the Elsevier Grand Challenge** used their **CSIBS** text mining system over the Elsevier life science corpus to automate the creation of 'citations in context'
- By clicking on the in-text citation of Dekker *et al.* 2002, four sentences of relevance to the context are pulled back from the cited paper

Dekker *et al.*, 2002 J. Dekker, J.W. Rossen, H.A. Buller and A.W. Einerhand, The MUC family: an obituary, *Trends Biochem. Sci.* **27** (2002), pp. 126–131. [Article](#) | [PDF \(72 K\)](#) | [View Record in Scopus](#) | [Cited By in Scopus \(119\)](#)

Additional Information from CSIBS:

⊕ Author Affiliation:

⊕ Abstract:

⊖ Citation-Sensitive Preview:

Defining mucins: family values:

- There are two approaches to the **definition** of mucins but both are unsatisfactory when it comes to defining the relationships of the mucin-encoding genes.
- Using this criterion to define mucins would be similar to conflating all lipoproteins based on their modification with lipid moieties and calling the encoding **genes** ?LIP-number?.

All in the family?:

- MUC3 was one of the first **MUC** proteins found, in 1990 [4], but it has recently been discovered that there are, in fact, two closely related and adjacent **genes** (MUC3A and MUC3B) with 98% homology [26].

Conclusions: families and orphans:

- Based on sequence homology, two families of mucins can be distinguished: (1) the mucin **genes** at locus 11p15, which probably encode mucus-forming mucins; and (2) the mucin **genes** at loci 7q22, 3q and 1q21, presumably encoding membrane-bound mucins.

CiTO, the Citation Typing Ontology

<http://purl.org/net/cito/>

Reference list annotations using CiTO

Sort by:

alphabetical order

publication year

frequency of citation (within this paper)

number (original)

Turn citation typing off

1. United Nations Human Settlements Programme (2003) The challenge of slums: Global report on human settlements 2003. London: Earthscan Publications Ltd. [Link](#) (CiTO: *obtains background from, Report, Book, Online Document, not peer reviewed*)

2. Riley LW, Ko AI, Unger A, Reis MG (2007) Slum health: Diseases of neglected populations. BMC Int Health Hum Rights 7: 2. [DOI PubMed PubMedCentral](#) (CiTO: *obtains background from, shares authors with, Opinion, Journal Article, peer reviewed*)

3. Sclar ED, Garau P, Carolini G (2005) The 21st century health challenge of slums and cities. Lancet 365: 901–903. [DOI PubMed](#) (CiTO: *obtains background from, Opinion, Journal Article, peer reviewed*)

- The first three references from the reference list of the enhance version of Reis et al. (2008), with the citation typing display turned on.
- Above the references are buttons to re-order the references, and to turn off the citation typing display.

The first purpose of CiTO, the citation typing ontology

- To permit the **existence** of a citation between the citing work and the cited work to be recorded in RDF

```
<http://example1.com/citingwork> cito:cites  
<http://example2.com/citedwork> .
```

- And reciprocally, we can say

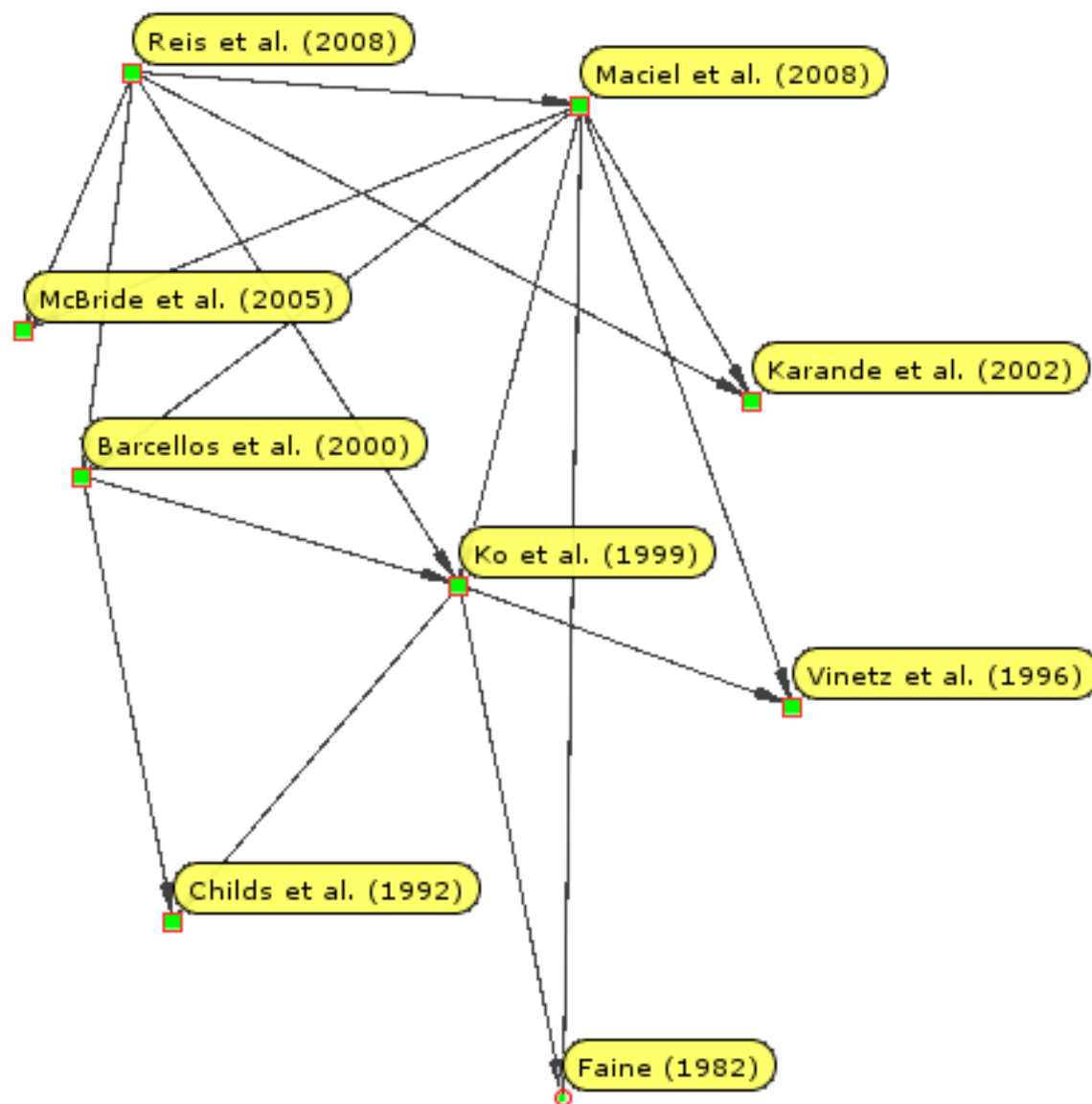
```
<http://example2.com/citedwork> cito:isCitedBy  
<http://example1.com/citingwork> .
```

which is useful, despite the logical redundancy

- Even this simple statement that a citation exists opens significant possibilities, for example in enabling the easy creation of **citation networks** simply by combining the RDF citation lists from several papers

A selected citation network from Reis *et al.* 2008

- Network is constructed automatically, by integrating the RDF citation data from Reis *et al.*, Maciel *et al.*, Barcellos *et al.* and Ko *et al.*, then visualized it using **Welkin**
- The nodes are arranged along a vertical temporal axis



The second purpose of CiTO

- To permit the **nature of a citation** between the citing work and the cited work to be characterized, **both factually and rhetorically**
 - An author will cite an article for one of several reasons, usually to acknowledge its importance, but sometimes to critique or refute it
 - CiTO makes it possible to capture and publish such distinctions in metadata describing the citation, quite distinct from descriptions of the cited work itself
- CiTO relationships between citing and cited document:
 - *cites, citesForInformation, confirms, corrects, credits, critiques, disagreesWith, discusses, extends, isCitedBy, obtainsBackgroundFrom, obtainsSupportFrom, refutes, reviews, sharesAuthorsWith, updates, usesDataFrom, usesMethodIn*

e.g. <http://example1.com/citingwork>

cito:cites <http://example2.com/citedwork> ;

cito:usesMethodIn <http://example2.com/citedwork> ;

cito:extends <http://example2.com/citedwork> ;

cito:sharesAuthorsWith <http://example2.com/citedwork> ; .

The third purpose of CiTO

- The third purpose of CiTO is to permit **citation frequencies** to be recorded, of two different types, **local and global**
 - **First**, the frequency of citation **within the text** of the citing work
 - If Paper A cites Paper B once, but cites Paper C ten times at different points in the text, then, *from the point of view of the citing paper*, Paper B is more significant, irrespective of its global citation frequency
 - **Second**, the frequency of citation **by the scholarly community** as a whole, as assessed by ISI Web of Knowledge or Google Scholar
 - Such global citation frequencies provide proxy estimates of the importance of each cited paper to the academic community

Encoding citation frequencies using CiTO

Citing document information

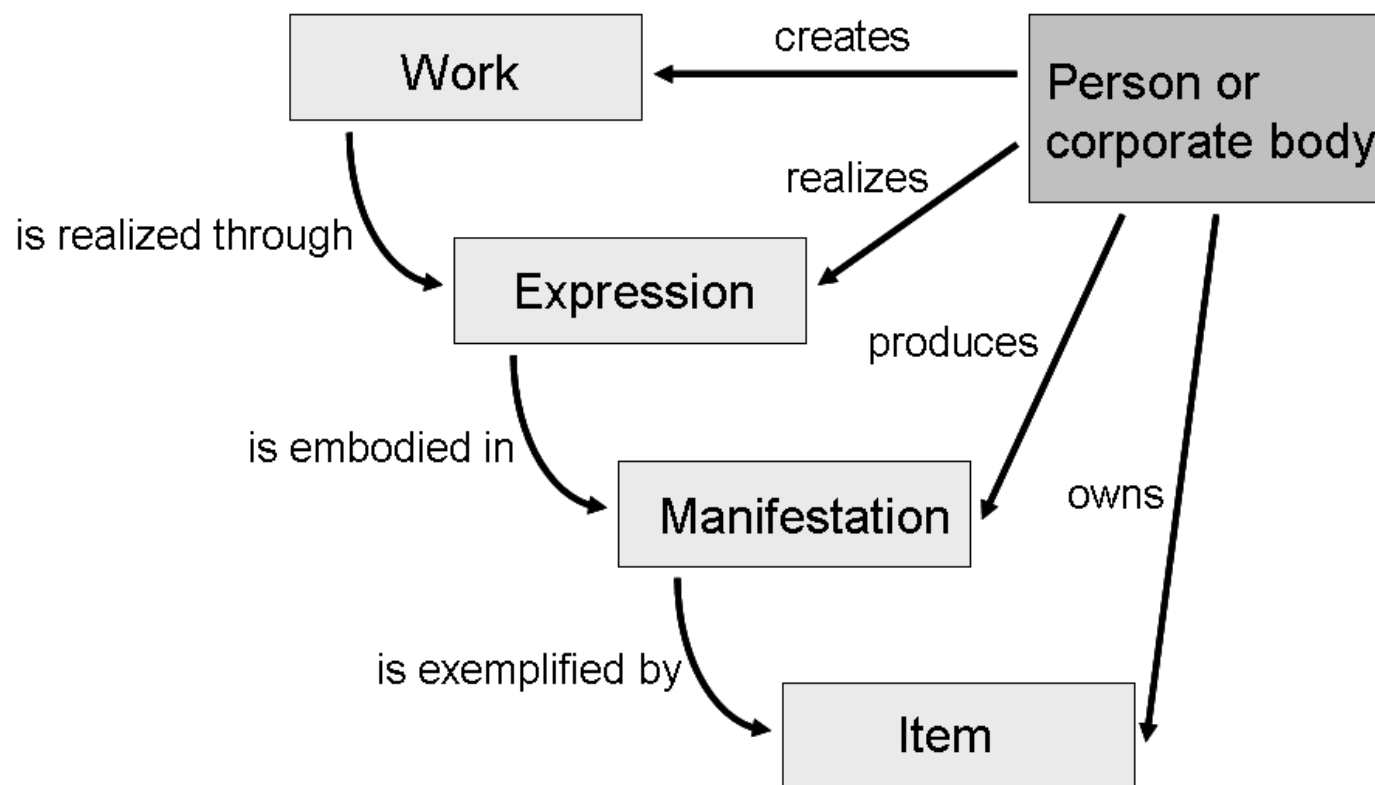
```
<http://example1.com/citingwork>
  cito:cites <http://example2.com/citedwork> ;
  cito:inTextCitationFrequency [
    a cito:InTextCitationCount ;
    cito:inTextCountValue "10"^^xsd:integer ;
    cito:inTextCitationTarget <http://example2.com/citedwork> ;
  ] ; .
```

Cited document information

```
<http://example2.com/citedwork>
  cito:isCitedBy <http://example1.com/citingwork> ;
  cito:globalCitationFrequency [
    a cito:GlobalCitationCount ;
    cito:globalCountValue "206"^^xsd:integer ;
    cito:globalCountSource <http://scholar.google.com>;
    cito:globalCountDate "2009-03-11"^^xsd:date ;
  ] ; .
```

The fourth purpose of CiTO

- The fourth purpose of CiTO is to **characterize the cited works** themselves
- In doing so, I have adopted the FRBR entity model



FRBR: Functional Requirements for Bibliographic Records, developed by the US Library of Congress (<http://www.loc.gov/cds/FRBR.html>)

Sub-classes of Work in CiTO

- *BookReview, Catalogue, Dataset, Discussion, Editorial, Explanation, GrantApplication, Image, Message, Model, MovingImage, NewsItem, Ontology, Opinion, Patent, Protocol, ReferenceWork, Report, ResearchPaper, Review, ScholarlyText, Software, Specification, StillImage, Taxonomy, WorkingPaper*

Sub-classes of Expression in CiTO

- *Blog, Book, BookChapter, BookSection, ConferencePaper, ConferencePoster, Database, Email, Figure, JournalArticle, JournalItem, PatentDocument, Preprint, Presentation, ReportDocument, Spreadsheet, Table, TextFile, Thesis*

Sub-classes of Manifestation in CiTO

- *DigitalMediaObject, OnlineDocument, PrintDocument, WebPage*

Citation information for Reference 2 in RDF (extracts)

Citing document information

#2

<<http://dx.doi.org/10.1371/journal.pntd.0000228>>

cito:cites <<http://dx.doi.org/10.1186/1472-698X-7-2>> ;

cito:obtainsBackgroundFrom <<http://dx.doi.org/10.1186/1472-698X-7-2>> ;

cito:sharesAuthorsWith <<http://dx.doi.org/10.1186/1472-698X-7-2>> ;

.

Cited document information

#2

<<http://dx.doi.org/10.1186/1472-698X-7-2>>

cito:isCitedBy <<http://dx.doi.org/10.1371/journal.pntd.0000228>> ;

dcterms:bibliographicCitation "Riley LW, Ko AI, Unger A, Reis MG (2007). Slum health: Diseases of neglected populations. BMC Int Health Hum Rights 7: 2.";

dcterms:issued "2007-03-07";

rdf:type cito:Opinion ; # work

rdf:type cito:JournalArticle ; # expression

cito:peerReviewed "true"^^**xsd:boolean** ; # peer review status

.

Overlap of CiTO with other ontologies: cited work

- In its characterization of the **cited work**, CiTO shares classes with both the SWAN Discourse Ontology and with BIBO, the Bibliographic Ontology
- Despite the clumsiness of this FRBR nomenclature, and the occasional seemingly redundant terminology that results from its use:

- *Work: Report* Expression: *ReportDocument*

this level of granularity avoids ambiguities of meaning present in these other ontologies, for example

- CiTO *Work: ResearchPaper* Expression: *JournalArticle*

- BIBO *AcademicArticle*, which conflated these concepts

- Effort is now required to harmonize these ontologies, for example:

- CiTO Expression: *JournalArticle*

- SWAN Manifestation: *JournalArticle*

- CiTO Expression: *Book*

- SWAN PublicationEnvironment: *Book*

Overlap of CiTO with other ontologies: the citation itself

- The other ontologies also contain a few object properties that could be used to characterize the nature of **the citation itself**
- The SWAN Discourse Relationship Ontology, although its primary purpose is wider than just citations, includes the relationships:
 - *agreesWith, arisesFrom, cites, consistentWith, disagreesWith, discusses, inconsistentWith, motivatedBy*
- BIBO, the Bibliographic Ontology, has only:
 - *affirmedBy, annotates, reviewOf, translationOf*
- Effort is required to harmonize these terms with those in CiTO
- I propose that we
 - (a) strengthen CiTO to do the job of describing citations,
 - (b) ensure BIBO is adequate for fully describing cited works, and
 - (c) use the SWAN Discourse Ontology to describe the wider rhetorical structures such as *Statement derivedFrom JournalArticle*, or *Statement refersTo Gene*

MIIDI

Minimal Information for an Infectious
Disease Investigation

The need for better research data descriptions

- Historically, we relied on printed Tables of Content and manual searching and browsing
- With the advent of on-line databases and bibliographic resources came free-text keyword Web searches
- With the ever-accelerating growth of biomedical data and literature, we now need to **automate methods of resource discovery and integration**
- This, in turn, requires **more principled methods of data description**
 - creating metadata adhering to community-agreed standards
 - publication of these metadata on the Web in machine-readable form

The *PLoS NTD* article Study Summary

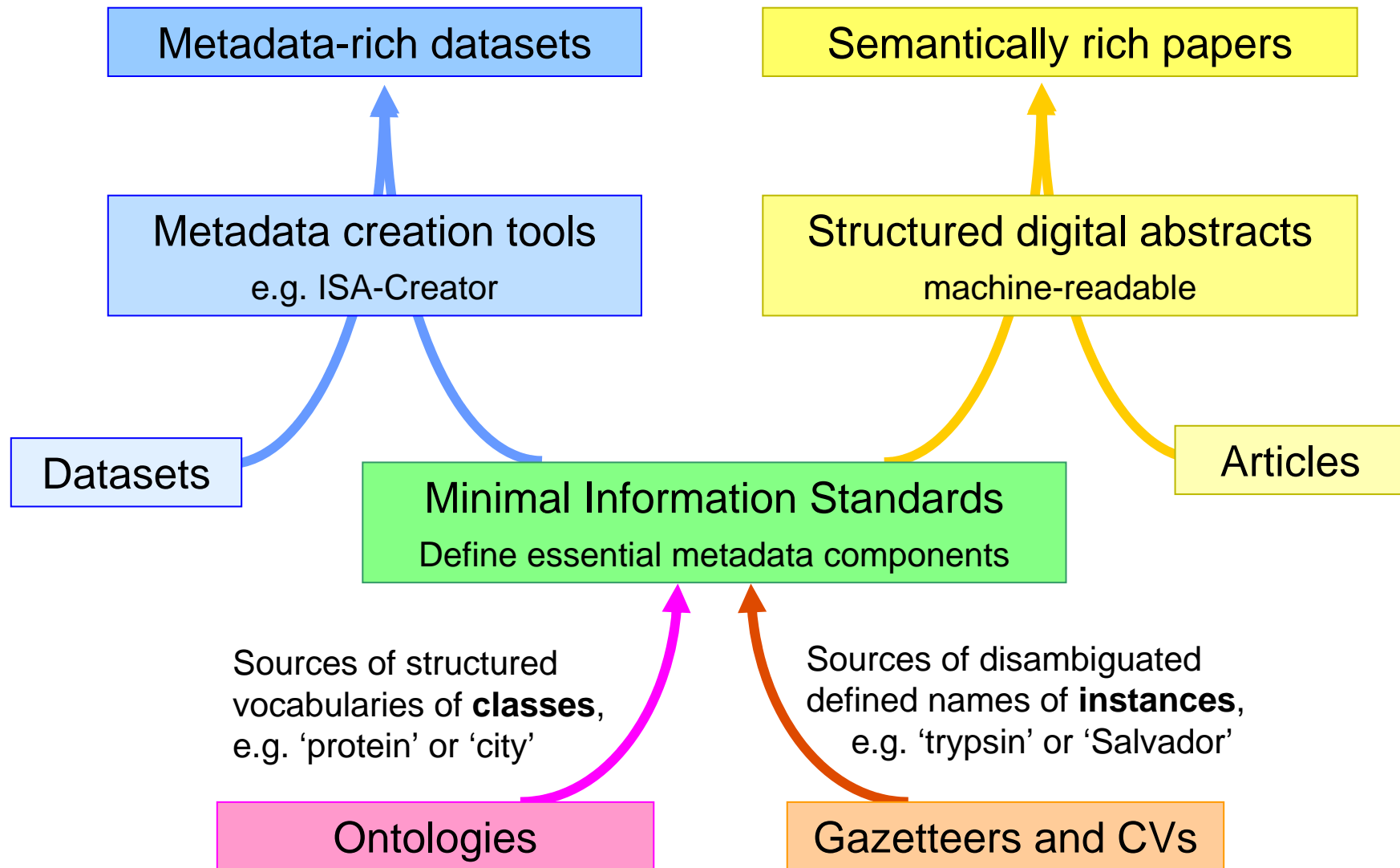
Infectious disease studied:	Leptospirosis
Pathogen (causative agent of disease):	Various species of the <i>Leptospira</i> spirochete bacterium
Primary animal vector of disease pathogen:	Rat (<i>Rattus norvegicus</i>)
Pathogen host subjected to study:	Human (<i>Homo sapiens</i>)
Number of subject individuals in study:	3,171
Number of control individuals in study:	None. This was a whole population study
Indicator of infection:	Presence of <i>Leptospira</i> agglutinating antibodies in blood
Assay used:	Microscopic agglutination test (MAT)
Location of study site (place name):	Pau da Lima, Salvador, Bahia, Brazil

- The Study Summary of our chosen *PLoS NTD* article:
 - was **specific** to that individual paper
 - was **not in machine-readable form**
- What was required was a proper machine-readable metadata standard that could be used to summarize any infectious disease investigation

MIIDI and other MIBBI standards

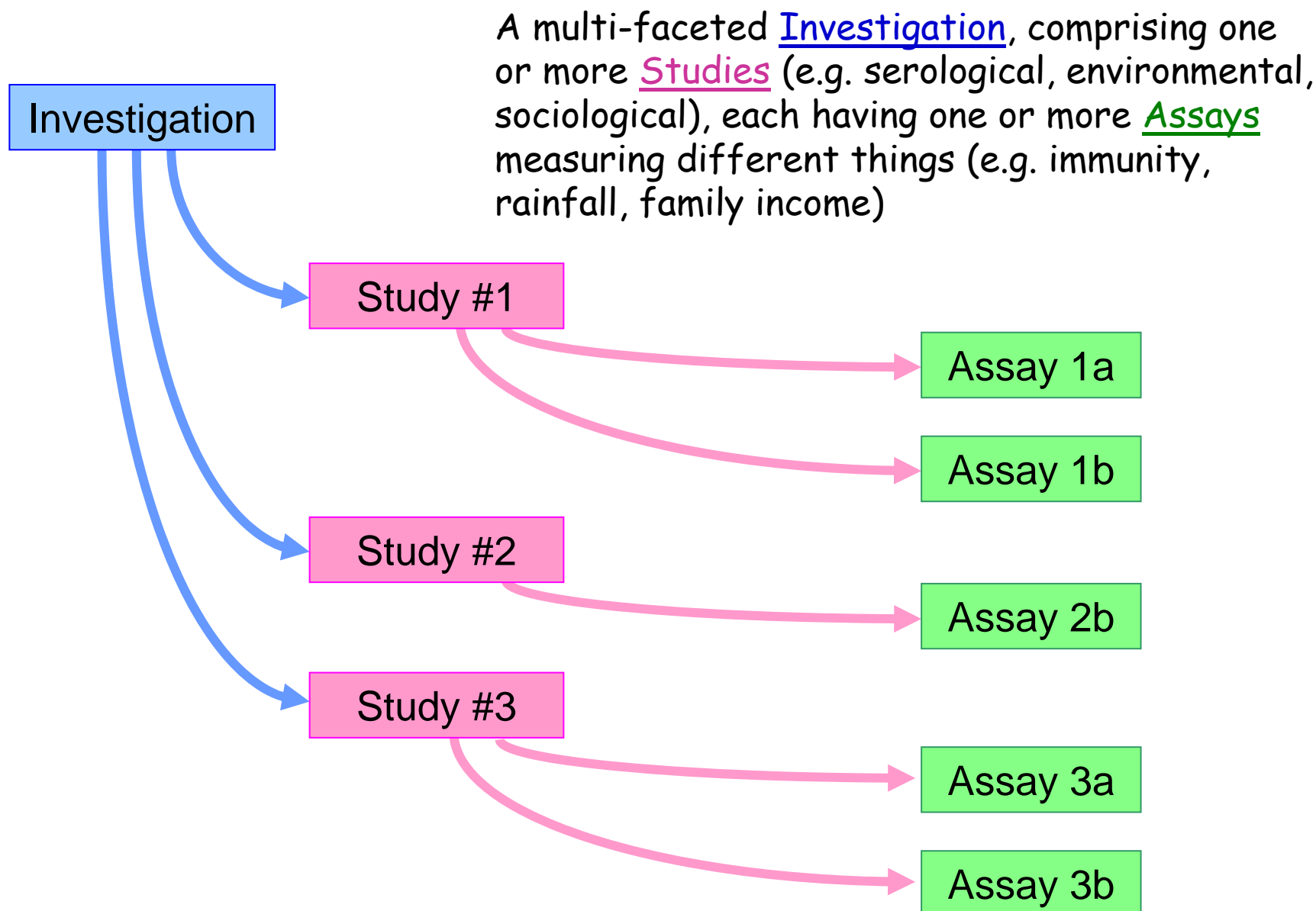
- So now we are developing MIIDI, a Minimal Information standard for reporting an Infectious Disease Investigation
- MIIDI is designed to provides a metadata check list for a wide diversity of investigation relevant to infectious diseases
- MIIDI extends the scope of previous MIBBI standards (Minimum Information for Biological and Biomedical Investigations), which are largely focused on metadata for research datasets of laboratory origin
 - MIIDI is designed for use in describing both datasets and publications
 - For the latter, it has items not found in any other MIBBI standard
 - e.g. investigation conclusions
- The MIIDI concept is generic, and can re-purposed to meet the requirements of other disciplines

Minimal Information Standards, Ontologies and Tools

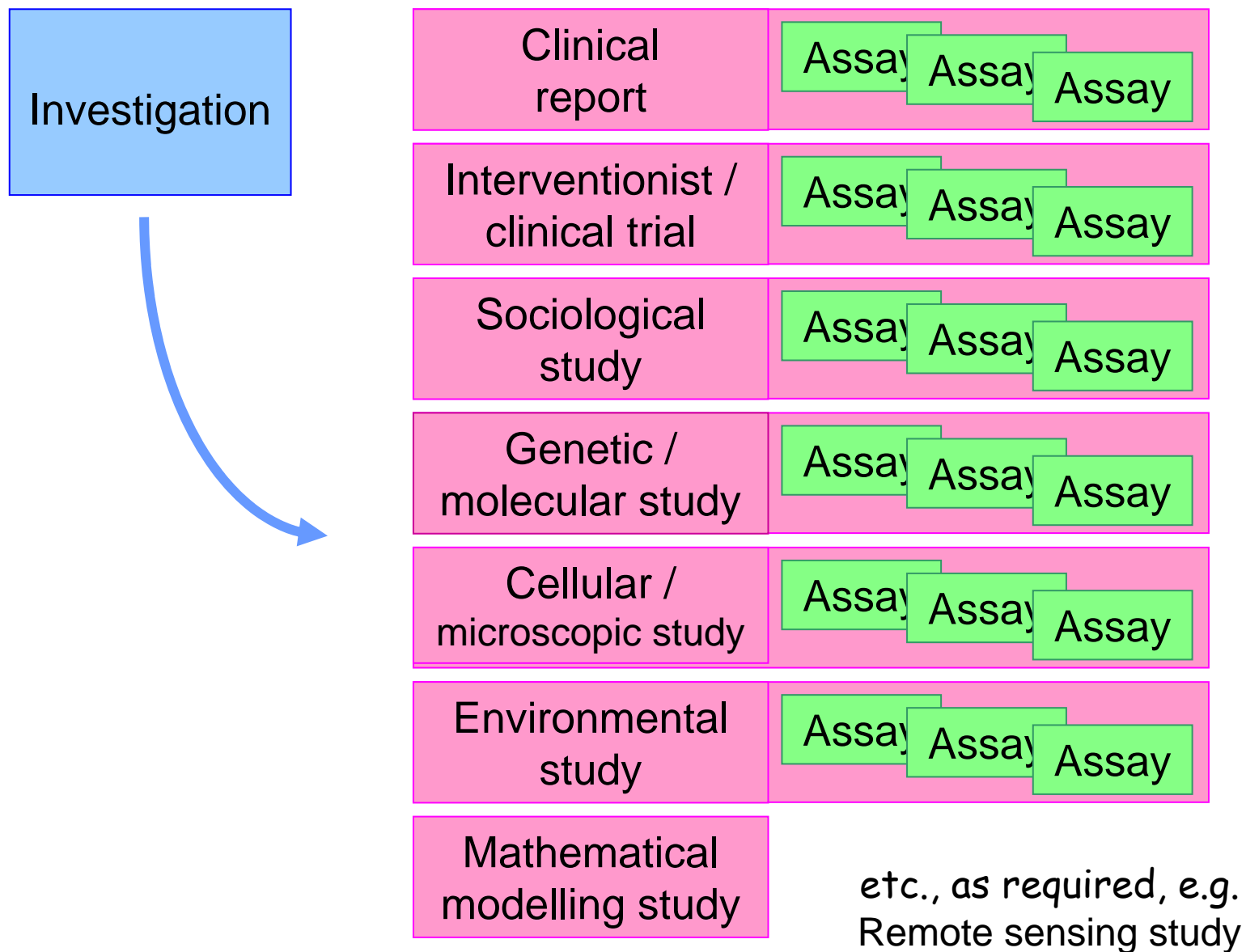


MIIDI adopts the ISA hierarchy

(<http://isatab.sourceforge.net/>)



The MIIDI Study Types capture domain-specific details



Generic MIIDI investigation metadata

INVESTIGATION DETAILS

Research project name Investigation purpose
Principal and other investigator(s) and their institution(s)
Funding agency/agencies and grant number(s)

DATASET DETAILS

Nature of stored data
Names of data submitters
Database or data location (name and URL)
Deposition date Accession number / ID
Open source license details
Access restrictions (if any)

and
/
or

ARTICLE DETAILS

Authors
Date of publication
Bibliographic details
Peer review status
DOI (or URL)
PubMed identifier

SELF-REFERENTIAL PROVENANCE INFORMATION

Nature of this MIIDI metadata document
Authors of this MIIDI metadata document
Date of this MIIDI metadata document
DOI of this MIIDI metadata document

Domain-specific MIIDI investigation metadata

DISEASE INVESTIGATED

Disease name

Subclass / type / severity

Host species name

Vector species name

Animal reservoir species name

Pathogen / parasite species name

Disease transmission source and route

STUDY TYPES EMPLOYED

(check one or more, as applicable)

Outbreak investigation / Clinical report

Epidemiological observational study

Interventionist investigation / Clinical trial

Mathematical modelling study

Cellular or developmental study

Molecular or genetic study

Environmental study

Systematic review or meta-analysis

Other (please specify)

INVESTIGATION CONCLUSIONS

(free text)

Principal conclusion 1

Principal conclusion 2

Principal conclusion 3

Principal conclusion 4

Principal conclusion 5

Principal conclusion 6

KEYWORDS

(MESH terms)

e.g. a clinical report - MIIDI Investigation metadata

INVESTIGATION DETAILS

Investigation purpose To characterize a measles outbreak in New Zealand
Principal investigator Brunton C
PI's institution Community and Public Health, Christchurch, New Zealand

ARTICLE DETAILS

Author Brunton C
Date of publication 14 July 2009
Bibliographic details Measles - New Zealand (06): Alert.
ProMED-mail 20090714.2512, page 1
Peer review status Not peer reviewed
DOI (or URL) <http://www.promedmail.org/>

DISEASE INVESTIGATED

Disease name Measles
Host species name *Homo sapiens* (Man)
Pathogen name Measles virus

STUDY TYPES

Outbreak report Yes

KEYWORDS: Measles, MMR

SELF-REFERENTIAL PROVENANCE METADATA

Nature Metadata recorded in conformity with MIIDI, the Minimal Information standard for reporting an Infectious Disease Investigation
Author of this MIIDI document Shotton DM
Date of this MIIDI document 31-08-2009

e.g. a clinical report - MIIDI Study metadata

STUDY DETAILS

Purpose of study To gather information on a current outbreak
of measles in Christchurch, New Zealand
Nature of study Clinical report

ORGANISM UNDER STUDY

Organism name *Homo sapiens* (Human)
Disease role Host

SUBJECT DETAILS

Subjects Human children
Inclusion criteria Symptoms of measles
Age range 9 months to 22 years
Total number 26
Number of males 21

STUDY PLACE AND TIME

Location - country New Zealand
Location - city / town Christchurch
Location (lat. & long.) 43° 34' S; 172° 39' E
Study start date 04/06/2009
Study end date 14/06/2009

e.g. a clinical report - MIIDI **Assay** metadata

ASSAY 1

Type of assay Sociological

Purpose To investigate any social connections between patients

Assay results 14 of the cases attend a Boys' High School in Christchurch, and 3 are in the same class.

The remainder of the cases are spread over the city with no obvious geographical connections.

ASSAY 2

Type of assay Medical

Purpose To determine previous vaccination history

Assay results 7 patients had received MMR triple vaccine in 2 doses: at 15 months and at 4 years of age

ASSAY 3

Type of assay Serological

Purpose To confirm measles infection in patients showing symptoms

Assay results 16 cases confirmed serologically

Confirmation awaited for a further 8 cases
(2 cases have refused blood tests)

How will MIIDI help?

- MIIDI, the Minimal Information standard for reporting an Infectious Disease Investigation, has six potential uses:
 - It can act as a content checklist for authors, editors and reviewers
 - It can underpin *machine-readable* Structured Digital Abstracts
 - It can ensure metadata for a research dataset is adequate
 - It can underpin tools for metadata creation (e.g. ISA-Creator)
 - It can aid resource discovery by providing consistent semantically defined search terms
 - Machine-readable MIIDI metadata files can facilitate *automated* data integration
 - Machine-readable MIIDI Structured Digital Abstracts can facilitate *automated* publication selection
 - e.g. of clinical trial reports, for systematic reviews

OpenFlyData and CLAROS

Two exemplar data webs,
integrating heterogeneous data
from distributed sources
on the fly

<http://openflydata.org/>

<http://www.clarosnet.org/>

The challenges of data integration

- Syntactic differences between data sources
 - Data are stored in incompatible formats within different DBMSs
 - Solved by converting all data to RDF
- Semantic differences between data sources
 - One person's "author" is another person's "creator"
 - Solved by mapping to a common data schema or ontology
- The co-reference problem
 - The same entity - for example a particular gene - is known by different names in different databases
 - Solved by creating a co-reference service to disambiguate synonyms

OpenFlyData sources: *Drosophila* gene expression data

FlyTED: the *Drosophila* Testis Gene Expression Database

Home Browse by Gene Name Browse by Strain Browse by Expression Location Links

Welcome to FlyTED

schuy Simple Search

- * Browse by Gene Name
- * Browse by Strain
- * Browse by Expression Location
- * Related Links
- * About the Dataset

5. Gene schuy in wt 6. Gene schuy in wt 7. Gene schuy in wt

University of Glasgow logo

bbsrc logo

BDGP In situ homepage - Mozilla Firefox

File Edit View History Delicious Bookmarks Tools Help

Patterns of gene expression in *Drosophila* embryogenesis

Release 2 (March 2007)

FAQ User feedback People

Project overview

We use high throughput 96-well plate RNA *in situ* protocol to determine patterns of gene expression during embryogenesis for *Drosophila* genes represented in non-redundant sets of *Drosophila* ESTs DGC1 and DGC2. At the end of production pipeline gene expression patterns are documented by taking a large number of digital images of individual embryos. The quality and identity of the captured image data are verified by independently derived microarray time-course analysis of gene expression using Affymetrix GeneChip technology (download array data from [here](#)). Gene expression patterns are annotated with controlled vocabulary for developmental anatomy of *Drosophila* embryogenesis. Image, microarray and annotation data are stored in a modified version of Gene Ontology database and the entire dataset is available on the web in browsable and searchable form (see below) or mysql dump can be downloaded from [here](#). So far we examined expression of 6138 genes and documented 3418 of them with 74612 digital photographs. Summary page summarizes all genes that have *in situ* images.

Search tools

Gene: Quick Search Enter a gene name to search for *in situ* patterns, use "*" as wild card

Basic Search Search form for *in situ* patterns by gene name, body part, function, cytology, protein domains

lmsGO Browse controlled vocabulary used to annotate *in situ* expression patterns

File Edit View History Delicious Bookmarks Tools Help

String to look for? schuy e.g. vha, cell adhesion, receptor, aquaporin, adenylate, CG1147, pnt

Searching for SCHUY through 18783 annotations produced 1 hits

schumacher-levy (schuy) FBgn0036925; FBgn0036925; 1626032 at

Accessions: NP_649166.1; NM_140909.

BDGP gene expression has embryonic *in situ* data for CG17736: 11 pix of staining in 2 body parts.

VDRG has 2 RNAi stocks (viable lethal) for FBgn0036925 available for purchase.

Tissue	mRNA Signal	Present Call	Enrichment	Affy Call
Brain	1 ± 0	0 of 4	0.00	Down
Head	0 ± 0	0 of 4	0.00	Down
Thoracoabdominal ganglion	1 ± 0	0 of 4	0.00	Down
Salivary gland	3 ± 1	0 of 4	0.04	Down
Crop	0 ± 0	0 of 4	0.00	Down
Midgut	0 ± 0	0 of 4	0.00	Down
Tubule	1 ± 0	0 of 4	0.00	Down
Hindgut	0 ± 0	0 of 4	0.00	Down
Ovary	0 ± 0	0 of 4	0.00	Down
Testis	1140 ± 36	4 of 4	13.90	Up
Male accessory glands	0 ± 0	0 of 4	0.00	Down
Virgin spermatheca	0 ± 0	0 of 4	0.01	Down
Mated spermatheca	1 ± 0	0 of 4	0.01	Down
Adult carcass	1 ± 0	0 of 4	0.00	Down
Larval Salivary gland	0 ± 0	0 of 4	0.01	Down
Larval midgut	1 ± 0	0 of 4	0.01	Down
Larval tubule	0 ± 0	0 of 4	0.00	Down
Larval hindgut	0 ± 0	0 of 4	0.00	Down
Larval fat body	68 ± 43	3 of 4	0.80	None
Whole fly	81 ± 7	4 of 4		

University of Glasgow logo

FlyAtlas

FlyBase Gene Report: Dmel\schuy - Mozilla Firefox

File Edit View History Delicious Bookmarks Tools Help

FB0009_02, released February 20, 2009

Gene Dmel\schuy

Home Tools Files Species Documents Resources News Help Archives Jump to Gene Go

Profile Manager

General Information

Symbol	Dmel\schuy	Species	<i>D. melanogaster</i>
Name	schumacher-levy	Annotation symbol	CG17736
Feature type	protein_coding_gene	FlyBase ID	FBgn0036925
Gene Model Status	Current	Created / Updated	2003-12-01/2003-12-01
Characterization Status	Uncertain(3)		

Genomic Location

Chromosome (arm)	3L	Recombination map	
Cytogenetic map	76D8-76D9	Sequence location	3L:19,979,177..19,981,172 [+]

Genomic Maps

FlyBase GBrowse

Ensembl GBrowse

Decorated Fasta Get genome region

Gene region Get Fasta

Summary Information

Automatically generated summary

The gene *schumacher-levy* is referred to in FlyBase by the symbol *Dmel\schuy* (CG17736, FBgn0036925). It is a protein_coding_gene from *Drosophila melanogaster*. It has the cytological map location 76D8. Its sequence location is 3L:19,979,177..19,981,172. Its molecular function is described as zinc ion binding. The biological processes in which it is involved are not known. One allele is reported. No phenotypic data is available. It has one annotated transcript and one annotated polypeptide.

See sections below for more information

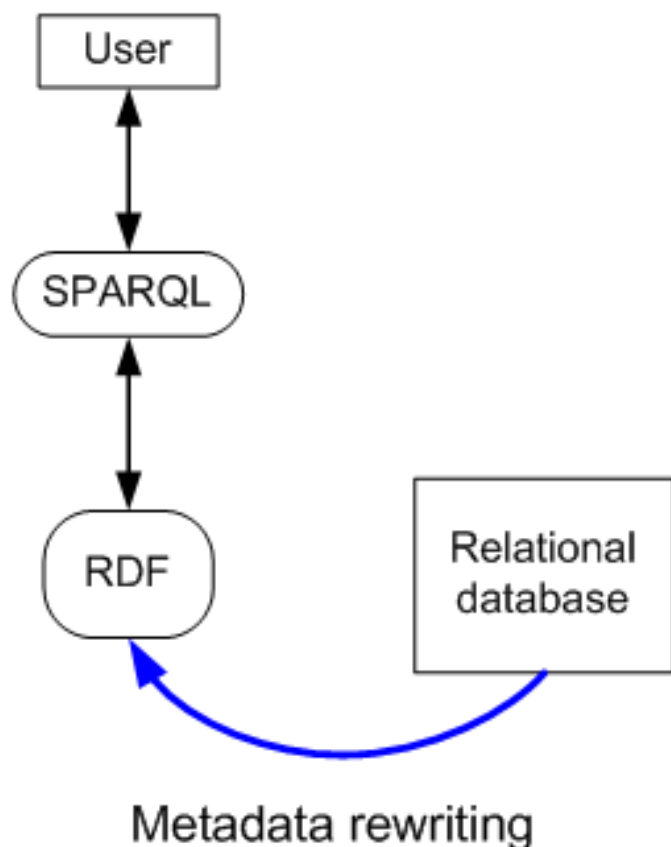
External Summaries

ihop

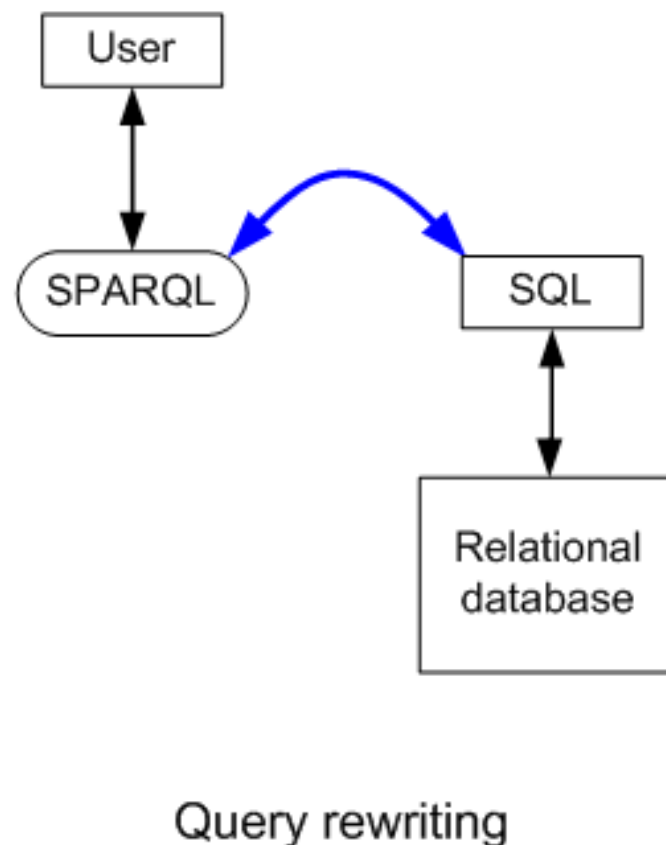
New functionalities developed to create OpenFlyData

- **Resource-specific ontologies** to support transformation of data from FlyBase, FlyAtlas, BDGP and FlyTED to RDF
 - available at <http://openflydata.googlecode.com>
- **SPARQL endpoints** for each data resource, from which RDF can be obtained in response to SPARQL queries
- **FlyUI** (based on the YUI library, <http://developer.yahoo.com/yui/>) - a library of JavaScript widgets providing re-usable user interface components for displaying *Drosophila* gene expression data
 - available at <http://flyui.googlecode.com>
 - these JavaScript applications run in a Web browser and fetch RDF data asynchronously over HTTP from the SPARQL endpoints
- **SPARQLite**, an implementation of the SPARQL protocol that avoids some performance problems when accessing the underlying triple store
 - available at <http://sparqlite.googlecode.com>
- Use of **FlyBase** (<http://flybase.org/>) to disambiguate gene names

Two methods of creating a SPARQL endpoint

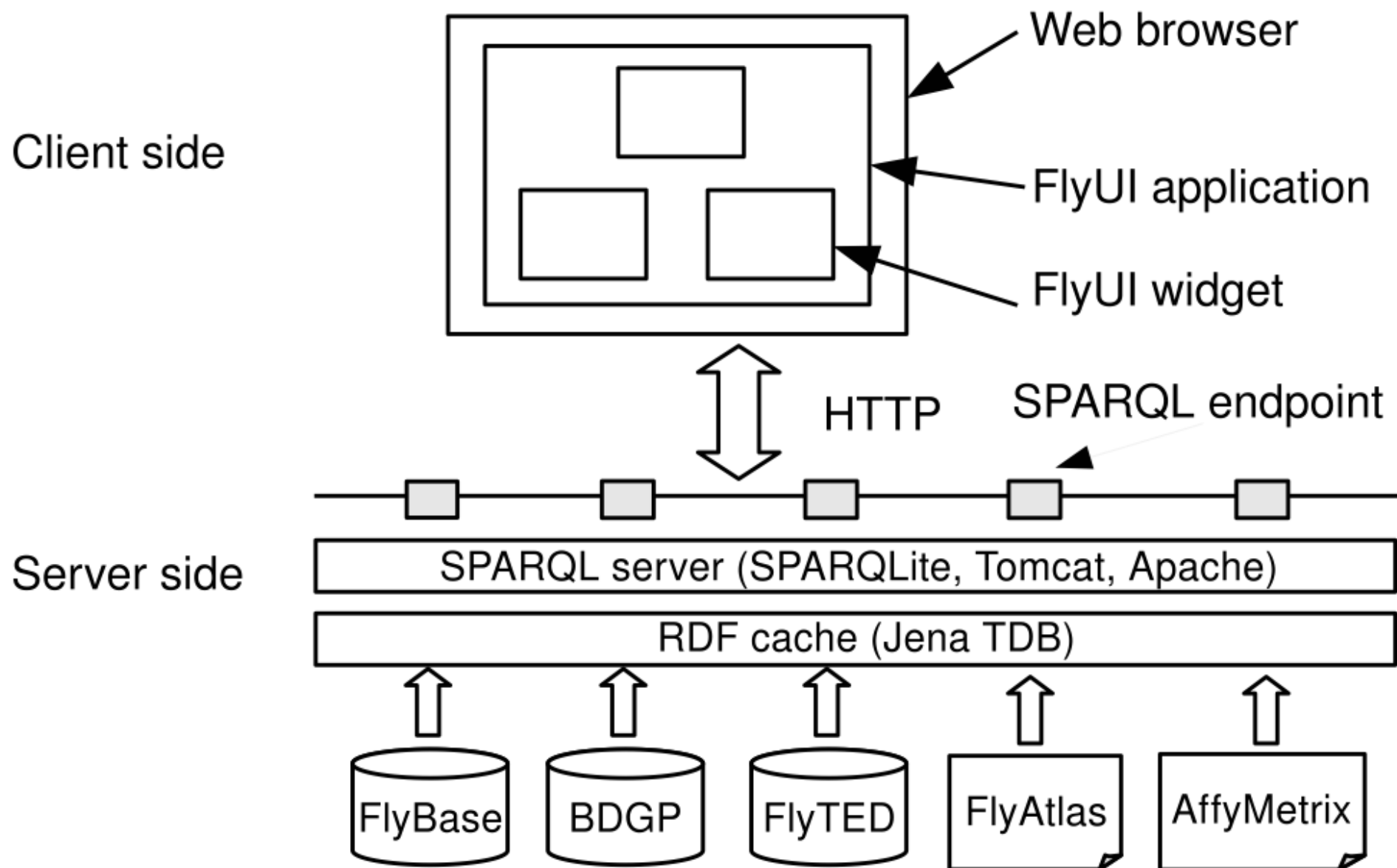


- Creation of a local RDF triplestore that **caches** selected source metadata, which are SPARQLeD - **"RDF caching"**



- Use of **D2R Server** to rewrite the SPARQL query into the database native query language (SQL) - **"SPARQL virtualization"**

The OpenFlyData architecture diagram



Same data integrated in a single OpenFlyData window

- Query SPARQL endpoints established on cached RDF data

Search *D. melanogaster* Gene Expression Data by Gene

gene name:

E.g. schuy, CG17736 or FBgn0036925 (case doesn't matter)

gene expression levels by tissue

found 1 matching probe from [flyatlas.org](#) (retrieved on 2008-09-16) for gene **schuy** ...


probe	tissue	mRNA signal	present call	enrichment	affy call
1626032_at	whole	81.9 ± 7.9	4 of 4	-	-
	brain	1.2 ± 0.6	0 of 4	0	Down
	head	0.4 ± 0	0 of 4	0	Down
	crop	0.8 ± 0.2	0 of 4	0	Down
	midgut	0.6 ± 0	0 of 4	0	Down
	hindgut	0.6 ± 0.1	0 of 4	0	Down
	tubule	1.5 ± 0.7	0 of 4	0	Down
	ovary	0.5 ± 0.2	0 of 4	0	Down
	testis	1140.4 ± 36.8	4 of 4	13.9	Up
	acc	0.9 ± 0.1	0 of 4	0	Down
	l_tubule	0.6 ± 0.2	0 of 4	0	Down
	l_fatbody	68.5 ± 43.9	3 of 4	0.8	None
	ta_ganglion	1.4 ± 0.5	0 of 4	0	Down
	carcass	1.3 ± 0.3	0 of 4	0	Down
	sgland	3.5 ± 1.3	0 of 4	0.04	Down

in situ hybridisation in embryos

found 8 matching images from [fruitfly.org](#) (retrieved on 2008-10-30) for gene **schuy** (BDGP report: [CG17736](#)) ...


stage 11-12

no staining;



stage 13-16

gonad;



references (flybase)

found 8 references from [flybase.org](#) (FB2009_02) for gene **schuy** (FlyBase report: [FBgn0036925](#)) ...

Barreau et al., 2008, Development 135(11): 1897--1902
Post-meiotic transcription in Drosophila testes. [[FBrf0205264](#)]

Dickson et al., 2007.7.18, RNAi construct and insertion data submitted by the Vienna Drosophila RNAi Center
RNAi construct and insertion data submitted by the Vienna Drosophila RNAi Center [[FBrf0200327](#)]

Benson et al., 2006, A. Dros. Res. Conf. 47: 494A
The Drosophila testis gene expression database. [[FBrf0189116](#)]

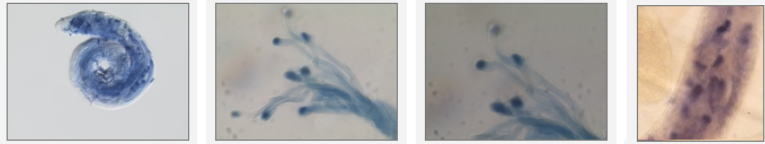
Benson et al., 2005, Europ. Dros. Res. Conf. 19: GG15
The Drosophila testis gene expression database. [[FBrf0184732](#)]

Benson, 2004.2.24, Helping FlyBase: ADRC-10142.
Helping FlyBase: ADRC-10142. [[FBrf0178789](#)]

Benson et al., 2004, A. Dros. Res. Conf. 45: 398B
Epigenetic Regulation by the aly-class Meiotic Arrest Genes. [[FBrf0173441](#)]

in situ hybridisation in testes

found 11 matching images from [www.fly-ted.org](#) (retrieved on 2008-12-03) for gene **schuy** (Fly-TED reports: [schuy](#))



schuy in wt schuy in wt schuy in wt CG17736/schuy in wt

CLAROS: The same technology repurposed for classical art

CLAROS

Classical Art Research Online Services

Search

Search



Home

CLAROS

Pottery

Gems

Sculpture

Iconography

Antiquaria

Dictionary

Tools

Virtual integration of digital assets on classical art

About us

Partner Institutions:

Beazley Archive,
Oxford

German Archaeological
Institute, Berlin

Lexicon
Iconographicum
Mythologiae Classicae
(LIMC) - Paris, Basel

Research Archive for
Ancient Sculpture,
Cologne

Lexicon of Greek
Personal Names



News

Beazley Archive / LGPN
Pilot Project

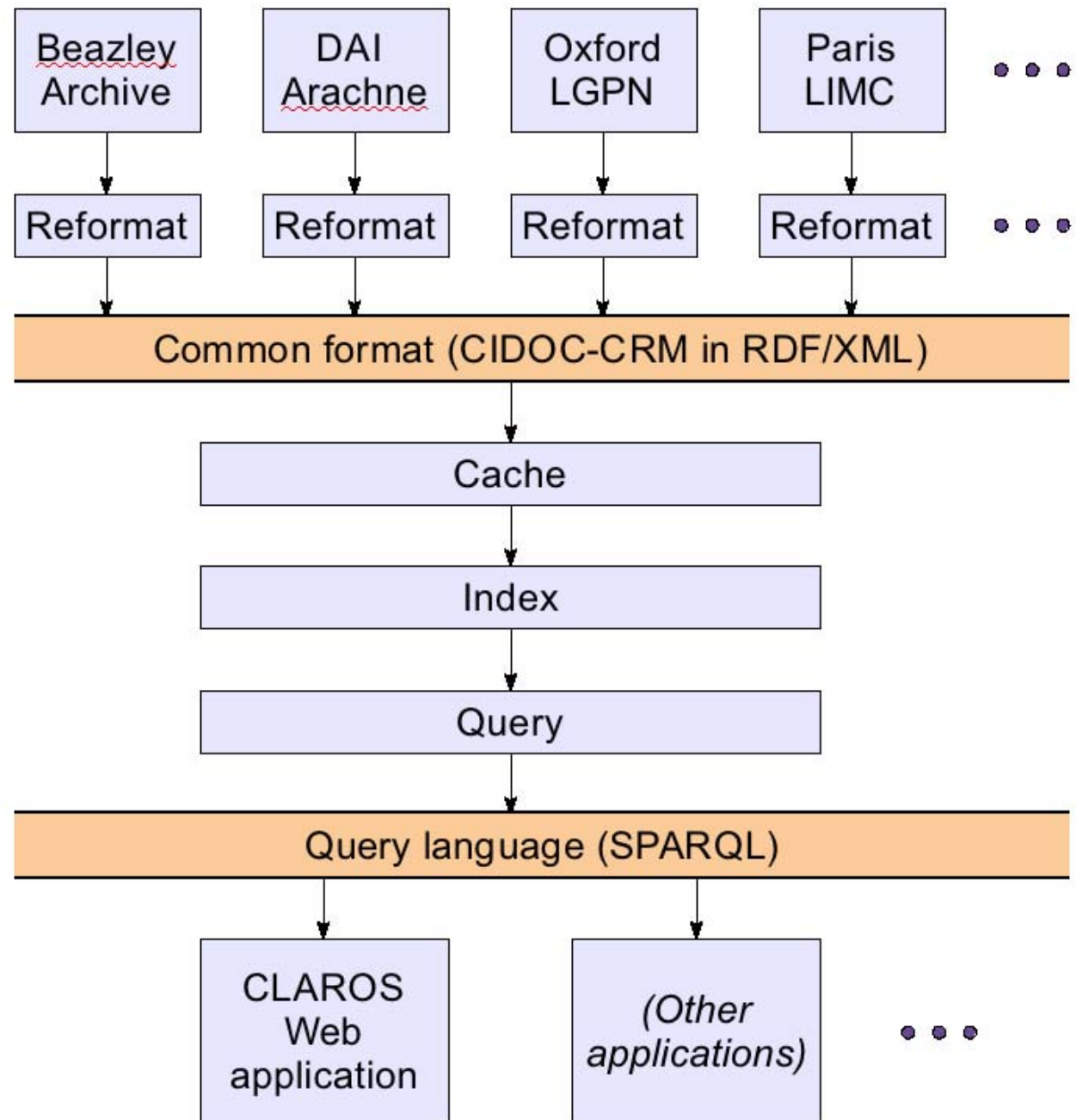
Berlin, FIEC Conference, 29
August: proof of concept
launch

Oxford Alumni Weekend,
25 September

CLAROS

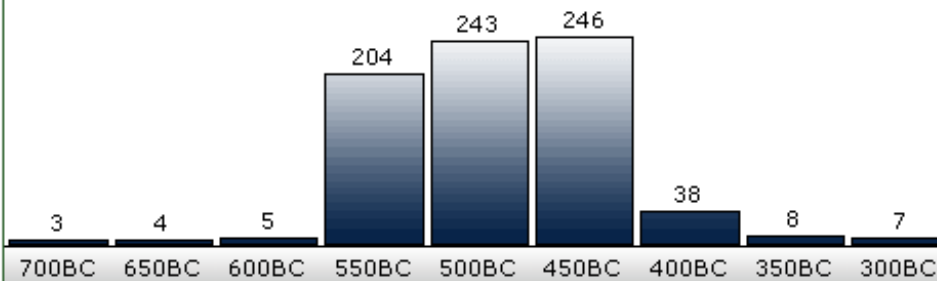
architectural diagram

- Data from all sources aligned to a single CIDOC CRM model
- Single CLAROS triple store cache contains about 10 million triples
- User interface is via the CLAROS Explorer
- Other data sources and new user applications can be added ...



The CLAROS Explorer faceted browse interface

The timeline shows the number of occurrences in each period. Click on the bar to show the distribution within the period or [click here to view the distribution for all periods](#).



Distribution of heron for all time periods

Click on the marker ▼ on the map and a balloon will pop up with the site name and number of occurrences of the name found at that site.



40008233, Edinburgh Tassie, 2203, EROS HOLDING A HERON BY THE NECK AND RESTING HIS BOW ON THE GROUND, Unpublished Tassie, TRAY 35.2



40000953, Poniatowski, T915, DIOMEDE AND ULYSSES PROCEEDING AS SPIES AT NIGHT TO THE TROJAN CAMP [CONDUCTED BY THE HERON WHICH MINERVA HAD SENT TO GUIDE THEM], KROMOU, Kromos, Prendeville, J.: Explanatory catalogue of the proof-impressions of the antique gems possessed by the late Prince Poniatowski and now in the possession of John Tyrrell, Esq. (1841), 915, Cornelian

Web page search results 1 to 4 of about 4 for heron

[Dexamenos - Classical gems - Gems](#)

... 21mm. GGFR no. 467. Blue chalcedony scaraboid, from Kerch (Crimea). A flying **heron**. ...

A **heron** preening and catching a locust with his foot. Signed Dexamenos. St. ...

<http://www.clarosnet.org/gems/styles/classical/dexamenos.htm> - 8k - 2008-12-11

Other Claros references for this period



Pottery 102155 records



Rundplastik 37905 records



Relief 25488 records

Acknowledgements

- My Oxford colleagues **Katie Portwin**, **Alistair Miles**, **Jun Zhao** and **Graham Klyne**, who worked with me on all these topics



- **The authors** of Reis *et al.* 2008, and the **Public Library of Science**, for being very supportive of our reuse of their published article
- **Lynette Hirschman**, for her excellent anonymous refereeing of our *PLoS Computational Biology* paper, and for then being gracious enough to reveal her identity
- **Anita de Waard** and **Philip Bourne**, for earlier work that inspired me
- **EPSRC** for funding the Ontogenesis Network that supported CiTO, the **RIN** for supporting MIIDI, and the **JISC** for supporting OpenFLyData

end