



Elsevier Health Science and Life Science Editors' Conference, Brussels

11th October 2009



The Future of Semantic Publishing



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

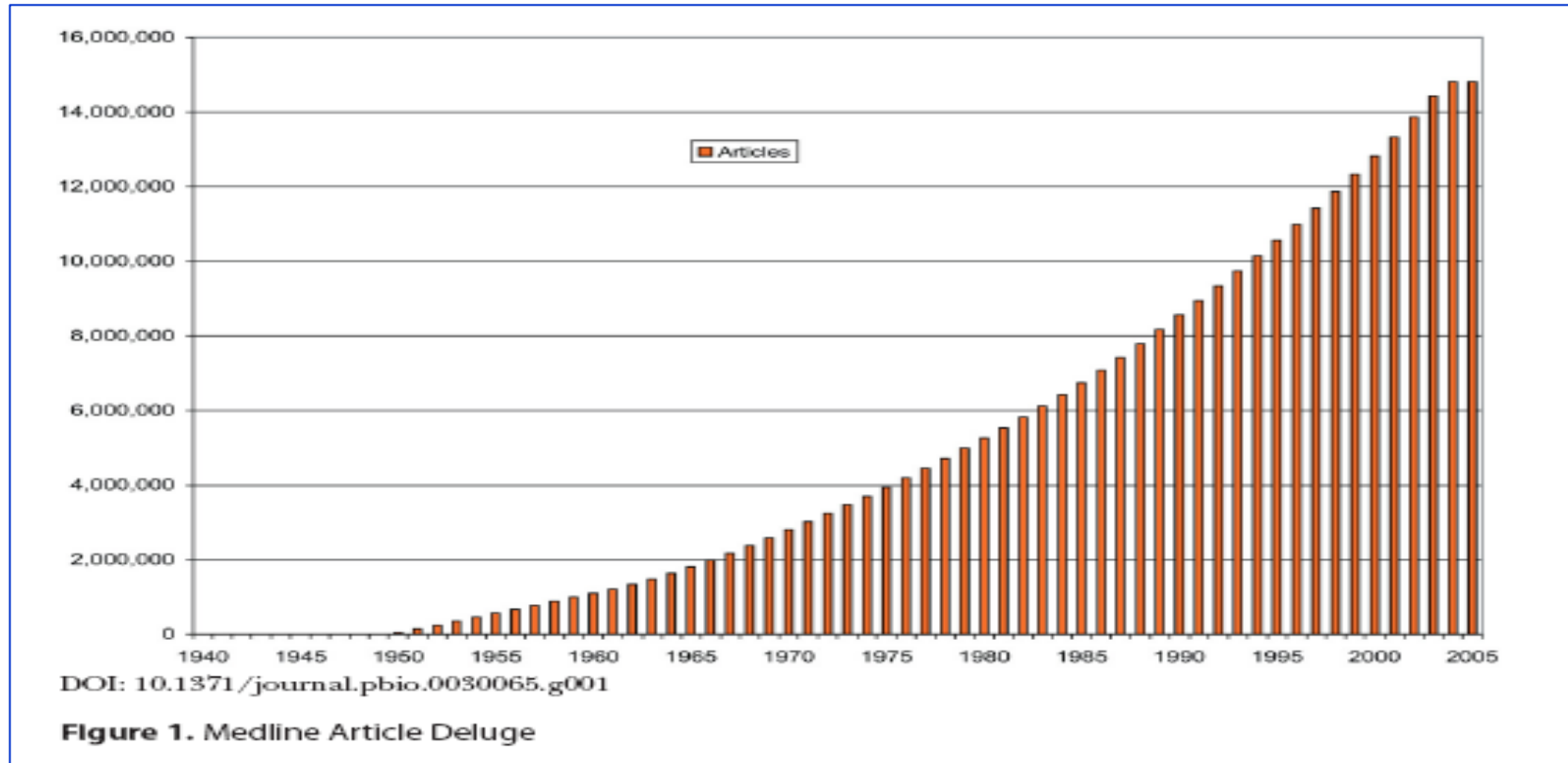


Outline

- Semantic Publishing - exemplar semantic enhancements of a research article
- CiTO, the Citation Typing Ontology
- MIIDI, Minimal Information for an Infectious Disease Investigation
- Changing the nature of biomedical publishing
- Take-home lessons

- Postscript: The SPIDER Project

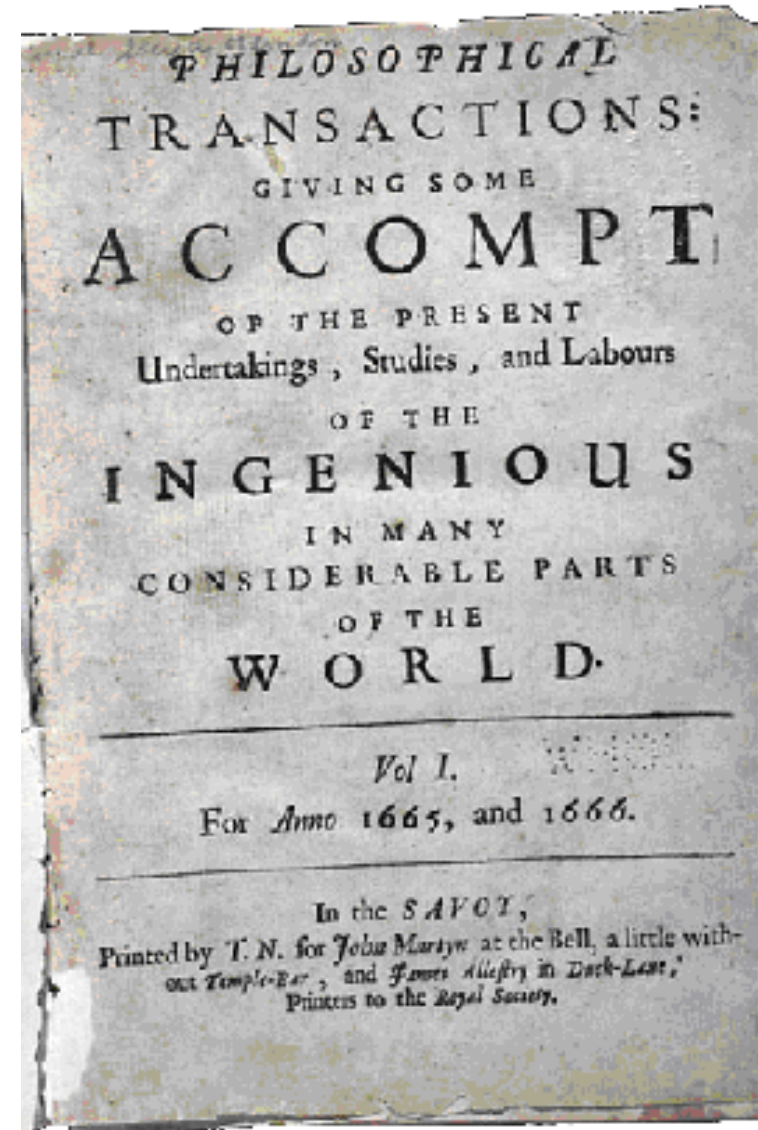
The number of research articles is growing . . .



- This increase is generally regarded as a threat, a challenge, a problem . . .
 - But it is *only* so if the assumption is that we have to **read** them all
- If, instead, they are seen as **resources to be mined**, this increase in the amount of information should be seen as a benefit and an opportunity

... however, 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



Exemplar **semantic enhancements** to an article from
PLoS Neglected Tropical Diseases

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

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

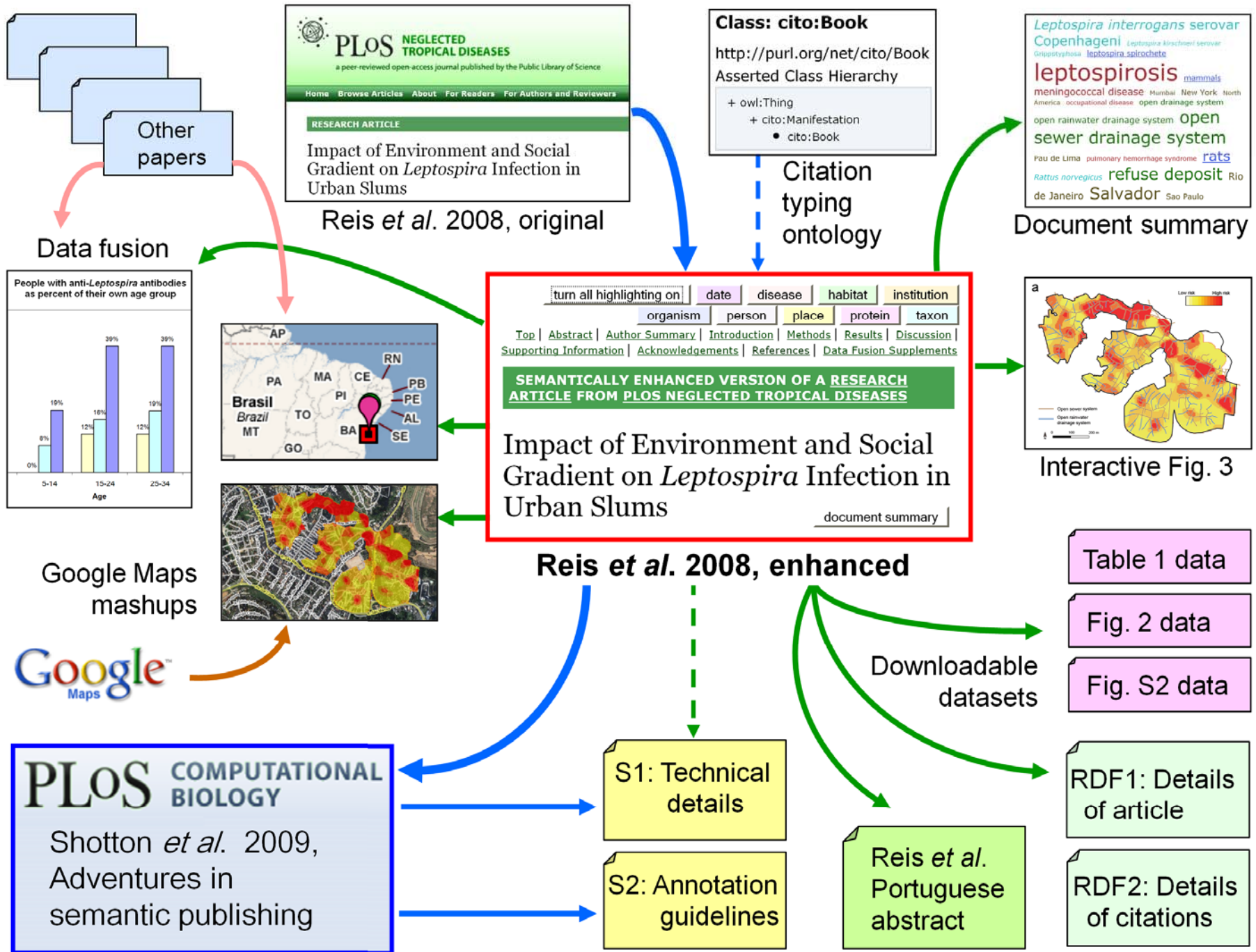
- Document summarization, including tag cloud and study summary
- Citation analysis, both of citation frequency and citation type

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



CiTO, the Citation Typing Ontology

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

Reference list annotations using CiTO

Sort by:

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

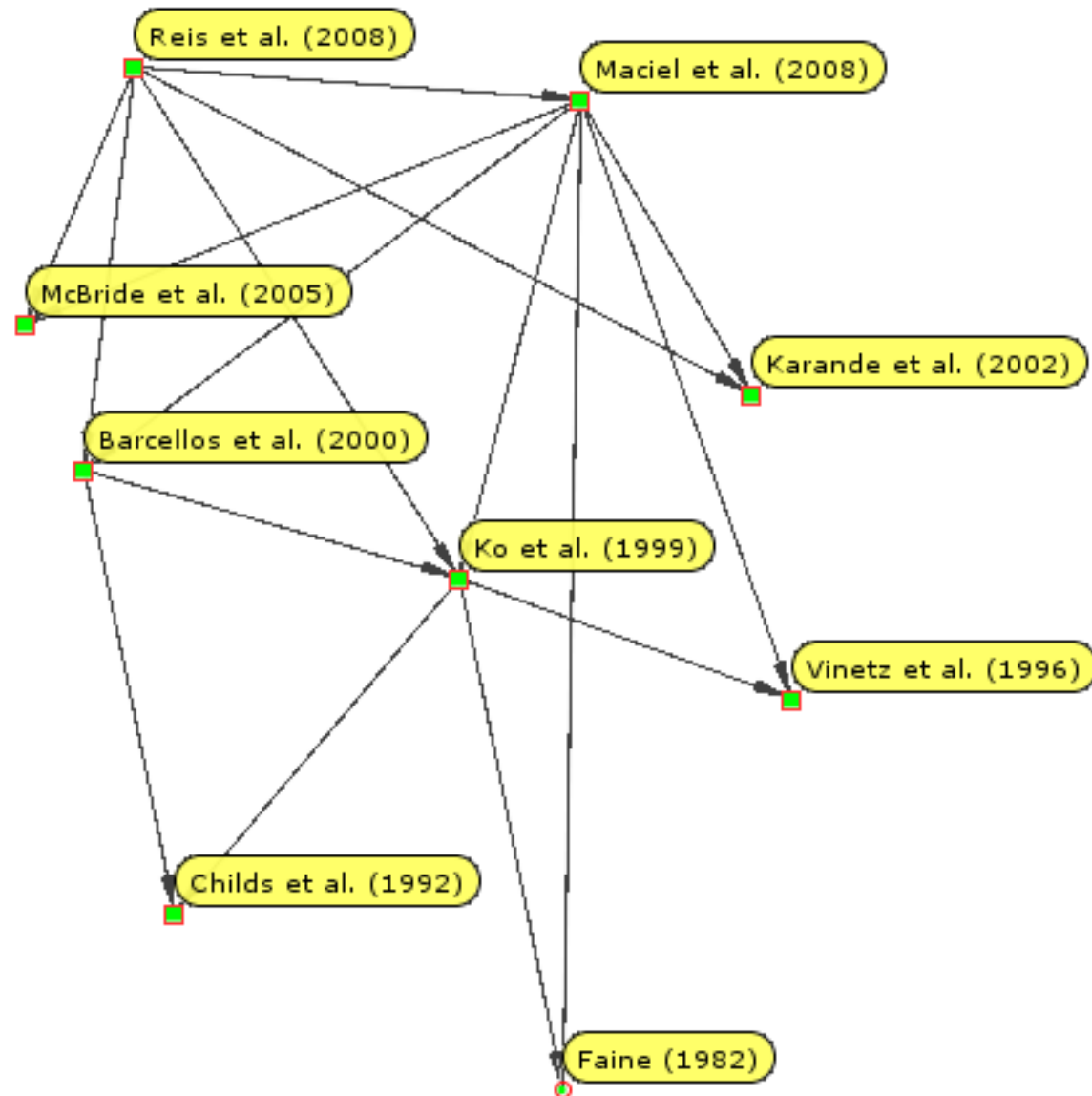
```
<http://example2.com/citedwork> cito:isCitedBy  
<http://example1.com/citingwork> .
```

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



Other uses of CiTO, the citation typing ontology

- To permit the **nature of a citation** between the citing work and the cited work to be characterized, **both factually**
 - *reviews, sharesAuthorsWith, usesMethodIn*, etc**and rhetorically**
 - *confirms, corrects, refutes*, etc
- To permit **citation frequencies** to be recorded
 - both **local**: "Paper A cites Paper B once, but cites Paper C ten times"
 - **and global**: "Paper C is cited 234 times according to Scopus"
- To **characterize the cited works** themselves using the FRBR "work, expression, manifestation" entity model (<http://www.loc.gov/cds/FRBR.html>)
 - Sub-classes of Work in CiTO include *Discussion* *ResearchPaper*
 - Sub-classes of Expression include *Blog* *JournalArticle*
 - Sub-classes of Manifestation include *OnlineDocument* *PrintDocument*

Applications of CiTO

- CiTO annotation of references within the reference lists of journal articles could be made standard
- Citation networks could be created to help readers navigate fields
- A tool is required that helps authors to input their CiTO metadata alongside their references at the time of paper writing
 - following Laura Hassink's proposal yesterday for a new tool to enable authors to enter reference metadata in a well-defined format
 - ?? adaption of the new Word 2007 plug-in for ontology-based semantic annotation

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

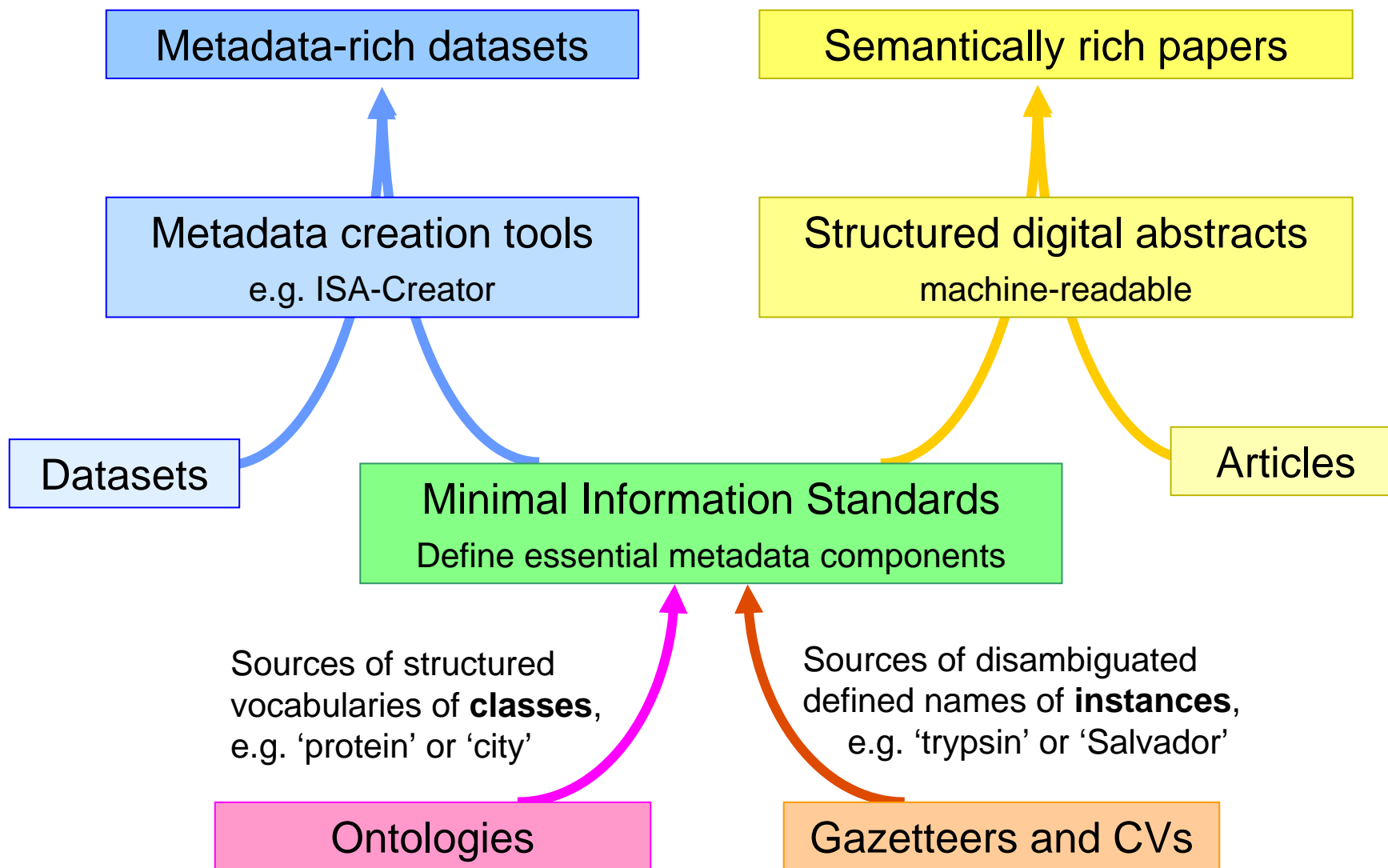
Infectious disease studied:	Leptospirosis
Pathogen (causative agent of disease):	Various species of the <i>Leptospira</i> spirochete bacterium
Primary animal vector of disease pathogen:	Rat (<i>Rattus norvegicus</i>)
Pathogen host subjected to study:	Human (<i>Homo sapiens</i>)
Number of subject individuals 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

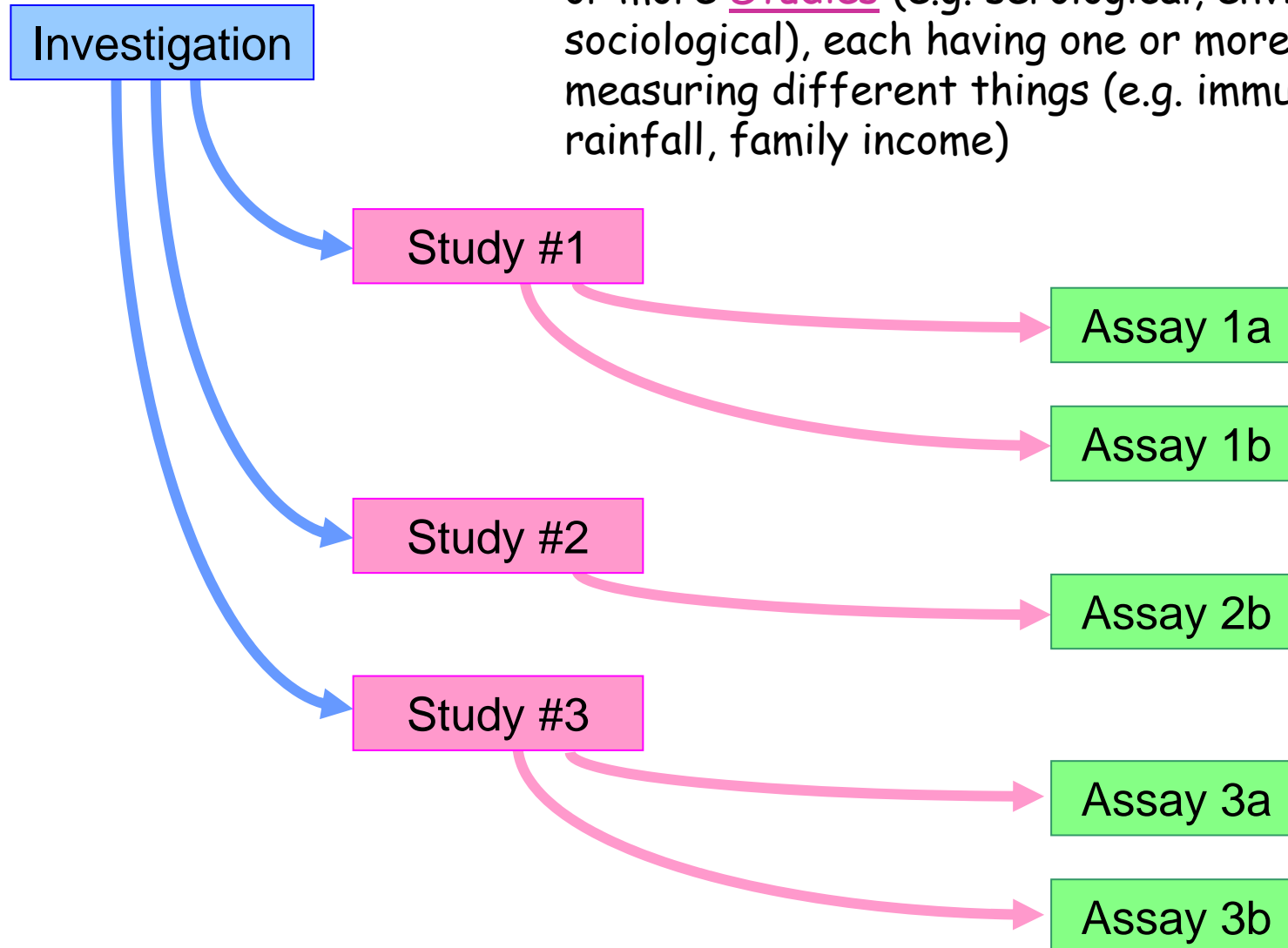
- So now we are developing **MIIDI**, a **Minimal Information standard for reporting an Infectious Disease Investigation**
- MIIDI is designed to provides a **metadata check list** for a wide diversity of investigation relevant to infectious diseases
- MIIDI extends the scope of previous **MIBBI** standards (**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

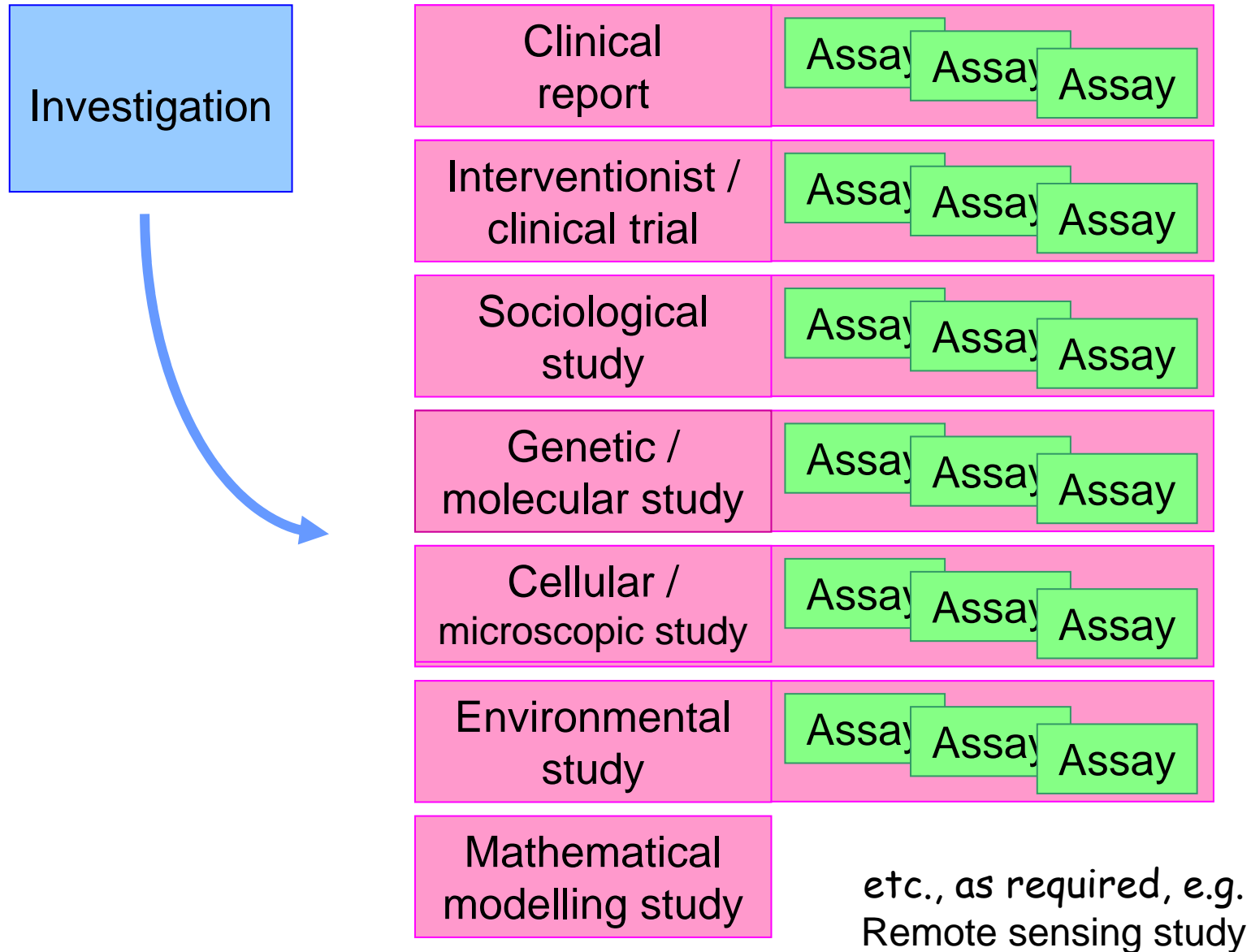


MIIDI adopts the ISA hierarchy [\(http://isatab.sourceforge.net/\)](http://isatab.sourceforge.net/)

A multi-faceted Investigation, comprising one or more Studies (e.g. serological, environmental, sociological), each having one or more Assays measuring different things (e.g. immunity, rainfall, family income)



The MIIDI Study Types capture domain-specific details



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

Where do we go from there?

Bringing Semantic Publishing into main stream biomedical journal production



Going forward from our start - the need for automation

- Our semantic enhancements were hand-coded, to provide an exemplar
- For semantic publishing to become main-stream, both human effort and **cost-effective automation** will be required
- The effort for this will need to be shared:
 - Pre-publication by authors
 - During publication by editors and publishers
 - Post-publication by users and third parties
- Improved software tools and better data standards will make the task progressively easier
- Everything does not have to be done at once
 - improvements can be introduced gradually
 - First harvest the low-hanging fruit
- The important thing is to make a start !

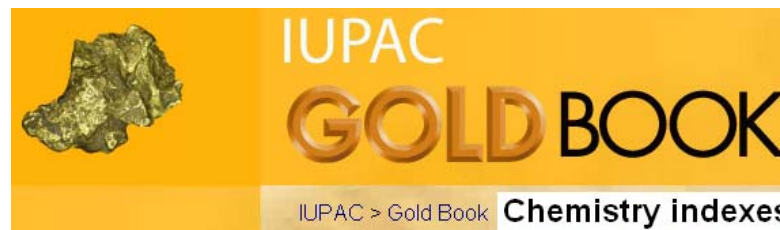


Tools to assist pre-publication semantic enhancement

- Microsoft Research have published a **plug-in for MS Word 2007** 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
- "Given the ubiquity of Microsoft Office software, as more and more users leverage the tool, we hope to see a notable impact on our ability to accelerate scientific discovery" Lynn Fink, Bourne Lab, UC San Diego
- The code for the ontology plug-in for MS Word 2007 is Open Source
 - Users can improve the plug-in or even to port it to other publishing systems, facilitating the growth of scientific semantic publishing

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*
- Technical editors mark up terms and link them to external resources, e.g.
 - **Chemical names** link to chemical structural formulae, lists of synonyms, IUPAC International Chemical Identifiers (InChI), XML descriptions in Chemical Mark-up Language, and patent searches for this chemical
 - **Gene Ontology terms** link to definitions, the GO ID numbers, list of synonyms and lists of other RSC articles referencing this term
 - **IUPAC Gold Book terms**



- "Project Prospect is fantastic. I've just seen the future of the journal"
- Ed Pentz, Executive Director of CrossRef

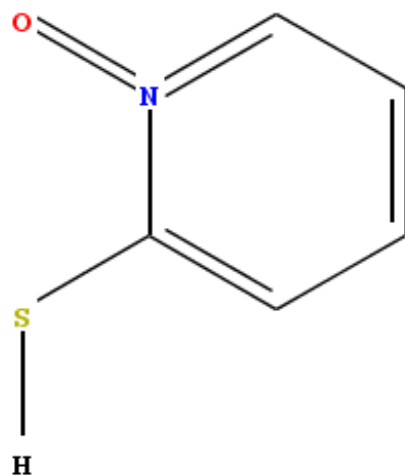
e.g. marked up chemicals, Gold Book and GO terms

Chemical genetics suggests a critical role for lysyl oxidase in zebrafish notochord morphogenesis†

Carrie Anderson^{† a}, Stephen J. Bartlett^{† b}, John M. Gansner^c, Duncan Wilson^a, Ling He^a, Jonathan D. Gitlin^c, Robert N. Kelsh^{*a} and James Dowden^{*bd}

As a result of a chemical genetic screen for modulators of metalloprotease activity, we report that 2-mercaptopyridine-*N*-oxide induces a con- of the *leviathan* mutant. The location of mutation led to the identification of a mutation in the *lysyl oxidase* gene, thus defining a narrow chemical sensit- led that notochord undulations appeared c- Notochord cells become swollen as we- tion of collagen fibrils in the surrounding sheath. *N*-oxide inhibits lysyl oxidase. Thus, we prov- r lysyl oxidase inhibition. Taken togeth- namic mechanisms of early morphogenesis at- er in zebrafish notochord formation.

2-mercaptopyridine-*N*-oxide



Synonyms

- pyridine-2-thiol N-oxide
- MCP
- 1-oxido-2-pyridinyl hydrosulfide

Chemical Information

- **SMILES:** [O-][n+]1cccc1S
- **InChI:** InChI=1/C5H5NOS/c7-6-4-2-1-3-5(6)8/h1-4,8H

Proper dataset publication by journal publisher Pensoft

ZooKeys 11: 1-8 (2009)
doi: 10.3897/zookeys.11.210
www.pensoftonline.net/zookeys

FORUM PAPER

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

**Publication and dissemination of datasets in taxonomy:
ZooKeys working example**

Lyubomir Penev¹, Terry Erwin², Jeremy Miller^{3,6}, Vishwas Chavan⁴,
Tom Moritz⁵, Charles Griswold⁶

- 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 mark-up of protein instances

□ 1: [Cell Death Differ.](#) 2009 Oct 2. [Epub ahead of print]

An **ARF/CtBP2** complex regulates **BH3** - only gene expression and **p53** - independent apoptosis.

[Kovi RC](#), [Paliwal S](#), [Pande S](#), [Grossman SR](#).

Department of Cancer Biology, University of Massachusetts Medical School, 364 Plantation Street, Worcester, MA 01605 USA.

The alternative reading frame (**ARF**) tumor suppressor exerts both **p53** - dependent and **p53** - independent functions. The **ARF** interacts with **ARF**, resulting in **prote** induce apoptosis in **p53** - null colon c interaction with **CtBP**. Bik was unique coordinately upregulated in colon car overexpression. Validating the array expression, and this activity required by **CtBP** deficiency was substantially silenced. An analysis of the Bik promoter interacting basic Kruppel - like factor repressed by **BKLF** and **CtBP2**, and A Chromatin immunoprecipitation anal promoter largely by **BKLF**. Expression expressing or **CtBP** - deficient cells revealed that Bik was uniquely regulated by

Close (Esc)

Protein [Help](#)




[ENSP00000354638](#) H. sapiens

Cyclin-dependent kinase 4 inhibitor A; CDK4I; p16-INK4; p16-INK4a; Mul ▶

[Domains](#), [Sequence](#), [Locus](#), [Literature](#)

RACGAEMGRGRCVGPSLQLRGQEWRCSPLVPKGGAAAELGPGGGEN

no structure available



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.

Take Home Lessons

Benefits of semantic publishing

- **Publishers** who are **early adopters** of semantic publishing will benefit by increasing the desirability and usage of their journals, and by having the potential to develop added-value services
 - Enhanced discoverability, increased interactivity, better integration with the Web, fuller access to data, better links to datasets
 - Increased manuscript submissions, of enhanced quality
 - Enlarged readership and improved journal citation index
- **Authors** will benefit by writing more useful and hence more cited papers, and by improved publication of their own datasets
- **Biomedical researchers** will get better access to both text and data
- Semantic publishing will enable STM publishers to fulfil their commitment in the Brussels Declaration on STM Publishing
 - to **change and innovation** that will make science more effective, and
 - to the **free availability of raw research data** submitted with a paper

A word of relevance from the past

*Come writers and critics who prophesize with your pen
And keep your eyes wide, the chance wont come again
And don't speak too soon for the wheels still in spin
And there's no tellin' who that its namin',
For the loser now will be later to win
For the times they are a-changin'.*

Bob Dylan, 1964



- Semantic publishing and open data publishing are already happening
- Continuing to do the same old thing is not an option at the present time!

Acknowledgements

- My Oxford colleagues **Katie Portwin**, **Alistair Miles** and **Graham Klyne**, who worked with me on semantic enhancements to the *PLoS Neglected Tropical Diseases* paper



- **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

Postscript:

The SPIDER Project



Name: Semantic Publishing for Infectious Disease Epidemiology Research

Premise:

- Lives may depend upon **timely availability of reliable epidemiological data**

Objective:

- To change the world in terms of the **publication of infectious disease research results**, by working as a consortium of researchers, authors, pharma and software companies, knowledge management specialists, and journal editors and publishers, to use on-line journals to their full potential
 - To develop tools to enable authors to create semantic mark-up, and annotated reference lists using CiTO
 - To publish MIIDI-based Structured Digital Abstracts for articles in chosen infectious disease journals
 - To assist authors in publishing *citable* infectious disease datasets that are reciprocally linked to the papers that describe them
- We now need to establish this consortium, obtain funding and start work
- Journal editors who wish to participate are requested to contact me

end

Achieving change in publishing - a lesson from history

- The field of bioinformatics is distinguished by the fact that open publication of biological data in research databases is now routine
- The first bioinformatics databases started during the early years of my own research career
 - **GenBank** for nucleic acid sequences (www.ncbi.nlm.nih.gov/Genbank/)
 - **UniProt** for protein sequence (www.uniprot.org/), and
 - **The Protein Data Bank** (PDB) for 3D structures (www.rcsb.org/pdb/)
- Initially, researchers were wary of submitting data to open repositories
- However, the decision by leading journals such as *Nature* to **mandate the inclusion of database accession numbers** in papers submitted for publication caused a very rapid sea change in attitude
- The same situation prevails today for semantic publishing - as soon as **leading journals require** authors to provide structured digital abstracts, the whole community will adopt this practice
- The benefits will be very apparent once this is no longer a 'niche' activity

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

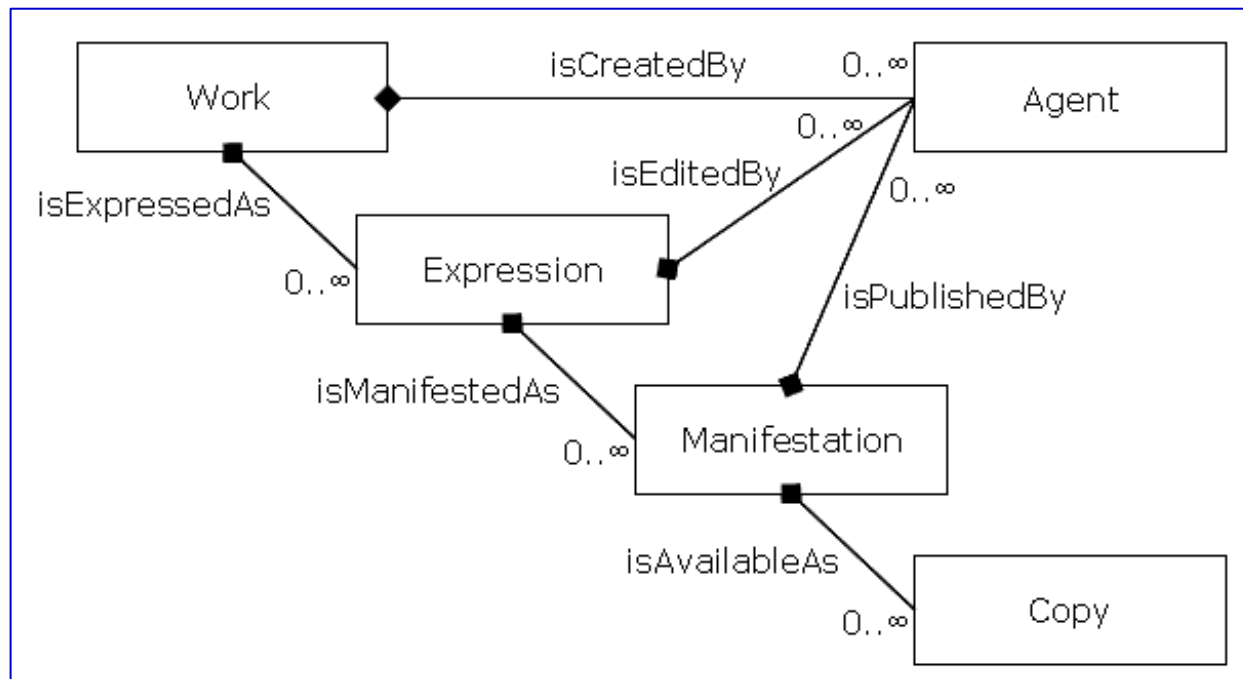
```
<http://example1.com/citingwork>
  cito:cites <http://example2.com/citedwork> ;
  cito:inTextCitationFrequency [
    a cito:InTextCitationCount ;
    cito:inTextCountValue "10"^^xsd:integer ;
    cito:inTextCitationTarget <http://example2.com/citedwork> ;
  ] ; .
```

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, created 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

.

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 (linkeddata.org/)
 - “The links are to web pages, so it will be hard to do computation on these”
 - “No reciprocal linking - the resource doesn't know it's a link target”
 - “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.”

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)

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

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

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)