Dryad-UK Workshop
Wolfson College, Oxford
12 September 2011

Minimal Metadata Standards and MIIDI Reports

David Shotton, Silvio Peroni and Tanya Gray

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

Dryad metadata

- At present, Dryad imposes minimal metadata requirements
- The DataCite Mandatory Metadata Properties required for DOI assignment:
  - Identifier (the DOI)
  - Creator (i.e. authors)
  - Title
  - Publisher (i.e. repository name “Dryad Data Repository”)
  - Publication Year
- Since a Dryad title is “Data from *Journal Article Title*”, and this and the authors’ names are harvested automatically from the journal, nothing extra is required
- In Work Package 5 of the Dryad-UK Project “Metadata standards for data annotation, deposition and citation”, we set out to investigate whether we could enable the creation of richer metadata without too much effort, providing better data descriptions that would assist discovery and reuse
- We also wanted to enable the publication of Dryad metadata as as Open Linked Data, encoded in RDF, the machine-readable data description language used on the Web
- The particular focus for enhanced metadata was infectious disease data

Mapping DataCite metadata to RDF

- Using appropriate ontology elements, from
  - from standard vocabularies - Dublin Core, FOAF, FRBR and PRISM
  - from our SPAR ontologies CiTO, CiTO4Data and FaBiO
  - and adding two missing properties in a new DataCite Ontology

  Silvio Peroni and I mapped all the metadata elements defined in the DataCite Metadata Kernel v2.0 (Jan 2011) to RDF
- This mapping document is available at http://bit.ly/jG0wt1

- We also created exemplar mappings of Dryad metadata to RDF, shown in a document **RDF for a Dryad repository holding, using DataCite terms**
- This mapping document is available at http://bit.ly/qpmUBU

- It would be good if Dryad could start to embed such RDF metadata in its landing pages as RDFa, so that it could enter the web of open linked data

RDF for a Dryad data package, using DataCite terms

<http://datadryad.org/handle/10255/dryad.8684> # The Dryad data package

dcterms:bibliographicCitation "Vijendravarma RK, Narasimha S, Kawecki TJ (2011) Data from: Plastic and evolutionary responses of cell size and number to larval malnutrition in Drosophila melanogaster. Dryad Digital Repository. 10.5061/dryad.8684" ;

datacite:hasPrimaryIdentifier [a prism:doi ;

literal:hasLiteralValue "10.5061/dryad.8684"] ;

datacite:hasAlternateIdentifier [a fabio:hasHandle ;

literal:hasLiteralValue "10255/dryad.8684"] ;

dcterms:creator [a foaf:Person ; foaf:name "Vijendravarma, Roshan K"] ;

dcterms:creator [a foaf:Person ; foaf:name "Narasimha, Sunitha"] ;

dcterms:creator [a foaf:Person ; foaf:name "Kawecki, Tadeusz J"] ;

dcterms:title "Data from: Plastic and evolutionary responses of cell size and number to larval malnutrition in Drosophila melanogaster" ;

dcterms:publisher [a foaf:Organization ;

foaf:name "Dryad Data Repository" ;

foaf:homepage <http://datadryad.org/>] ;

fabio:hasPublicationYear "2011"^^xsd:gYear ;

datacite:hasRelatedIdentifer [a prism:doi ;

literal:hasLiteralValue "10.1111/j.1420-9101.2010.02225.x" ; # DOI of the related journal article

literal:isLiteralOf [dcterms:relation <http://dx.doi.com/doi/10.1111/j.1420-9101.2010.02225.x>]] ;

frbr:supplementOf <http://dx.doi.com/doi/10.1111/j.1420-9101.2010.02225.x> .

Enhancing metadata – the Reis *et al*. (2008) exemplar

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

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*

Factual metadata in the Study Summary

Study Summary

Infectious disease studied:	Leptospirosis
Pathogen (causative agent of disease):	Various species of the *Leptospira* spirochete bacterium
Primary animal vector of disease pathogen:	Rat (*Rattus norvegicus*)
Pathogen host subjected to study:	Human (*Homo sapiens*)
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 *Leptospira* agglutinating antibodies in blood
Assay used:	Microscopic agglutination test (MAT)
Location of study site (place name):	Pau da Lima, Salvador, Bahia, Brazil
Northern limit of study site:	12 degrees 55 minutes 15.20 second South
Southern limit of study site:	12 degrees 55 minutes 42.90 second South
Eastern limit of study site:	38 degrees 25 minutes 51.20 seconds West
Western limit of study site:	38 degrees 26 minutes 26.70 seconds West
Starting date of study:	April 2003
Ending date of study:	May 2004

Rhetorical metadata in the Study Summary

Study Summary

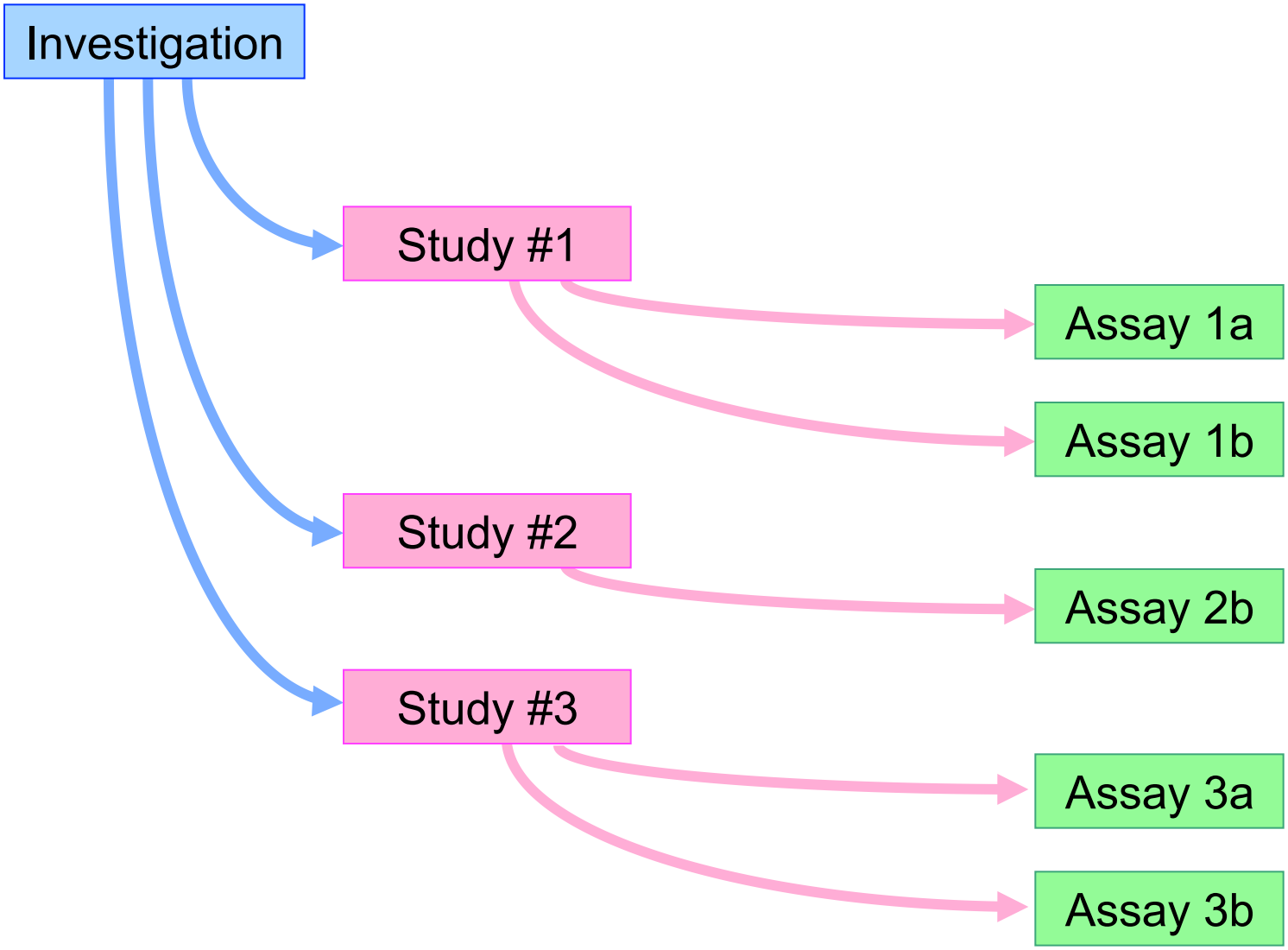
Purpose of study:	To quantify social and environmental risk factors for contracting leptospirosis in an urban slum
Principal finding 1:	Overall prevalence of *Leptospira* antibodies in the surveyed population was 15.4%.
Principal finding 2:	Disease risk was positively correlated with residence in flood-risk regions with open sewers, and with proximity of residence to accumulated refuse.
Principal finding 3:	Disease risk was positively correlated with sighting of rats and presence of chickens at the residence.
Principal finding 4:	Low income and black race were independent positive risk factors.
Principal finding 5:	An increase of US$1 per day in per capita household income was associated with an 11% decrease in infection risk.

- The problem with this summary is that
  - it is hand-crafted by a single individual
  - it is not backed by any recognised metadata standard
  - it is only human-readable, lacking an ontology-based machine-readable RDF representation

MIIDI http://www.miidi.org/

- MIIDI is a Minimal Information standard for reporting an Infectious Disease Investigation
  - A standard to formalise the Document Summary of Reis *et al*. 2008
- An international MIIDI workshop in September 2009 led to an initial draft
- In January 2011, Tanya Gray started work to develop MIIDI properly
- The MIIDI standard can be used *both* to create structured digital abstracts for journal articles, such as that by Reis *et al., **and also*** **to describe data sets, mathematical models, experimental workflows and software** relevant to an infections disease investigation, **to accompany Dryad submissions**
- Because the range of infectious disease investigations is large, MIIDI is specifically designed to deal with a diversity of investigation types and a variety of study types
  - It does so using the ISA – Investigation, Study, Assay hierarchy

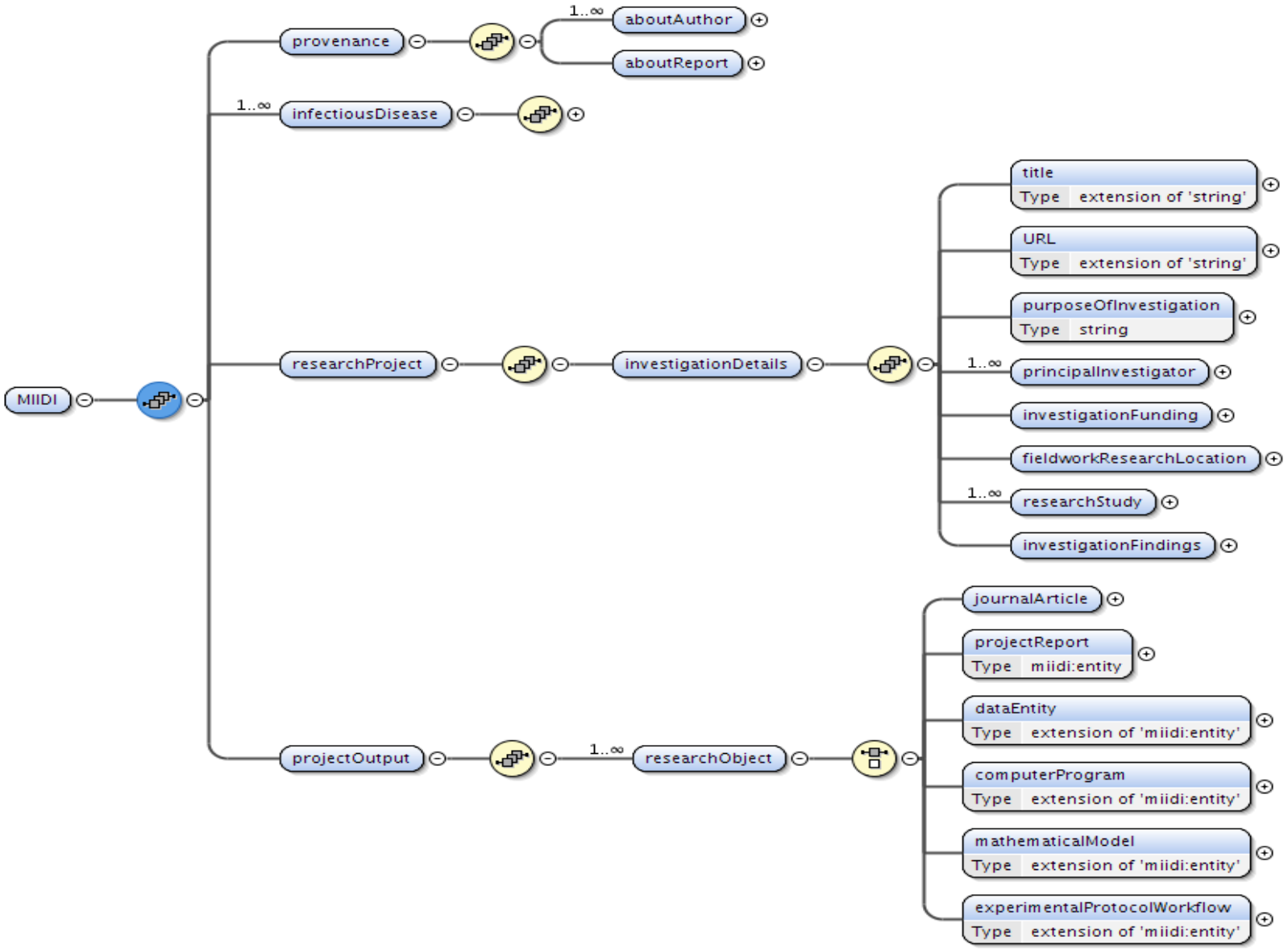
The Investigation – Study – Assay hierarchy of MIIDI



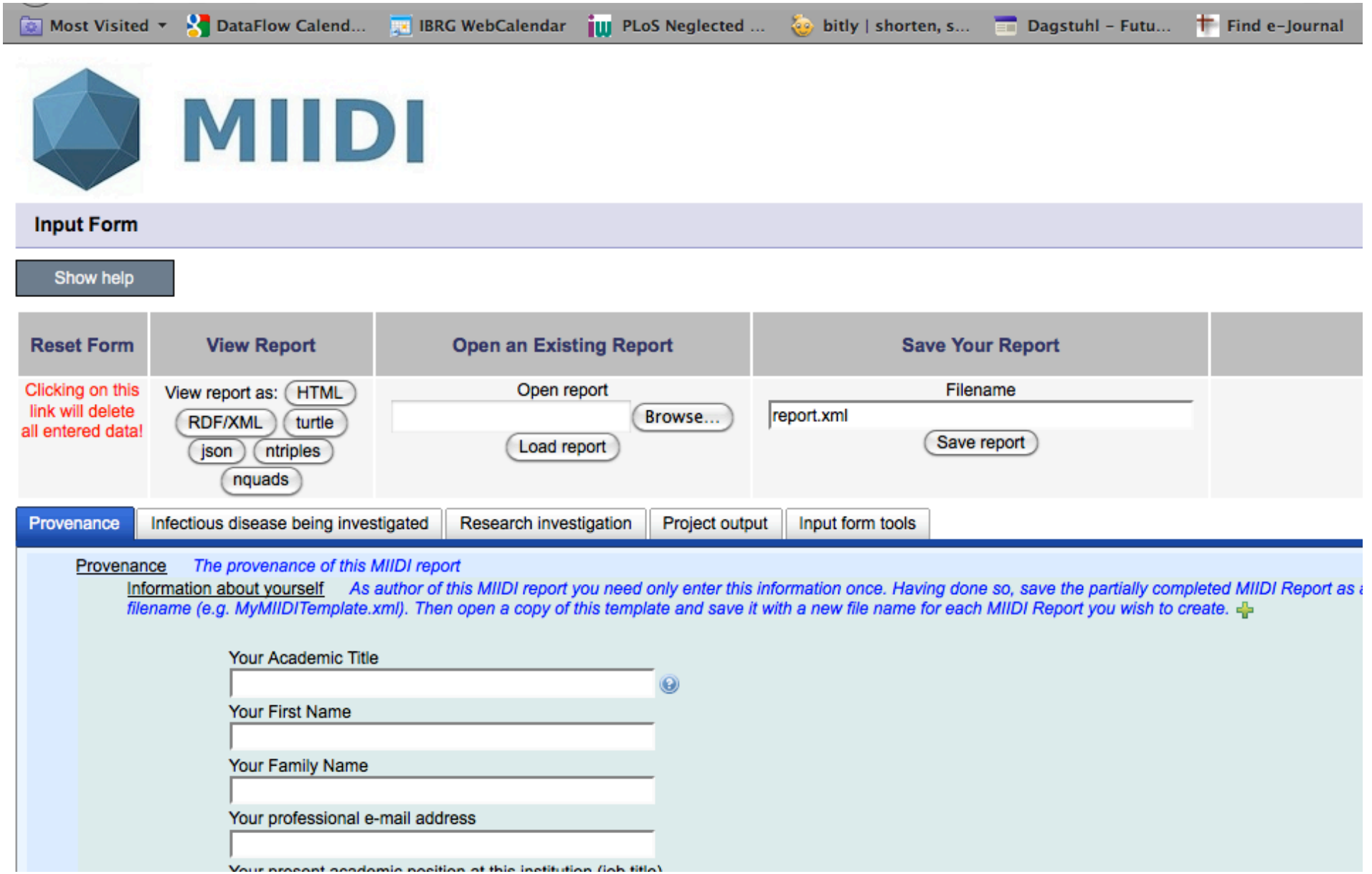
Developing MIIDI

- The September 2009 MIIDI draft was just a simple textual description
- On 31 January 2011, Tanya Gray started work with me, funded by the JISC Dryad-UK Project
- Her first activity was to develop MIIDI into a validated XML data model
- She then used Orbeon Forms to create a MIIDI Form based on the model, that permits easy Web metadata entry conforming to the MIIDI standard
- To permit encoding of MIIDI terms in RDF, we have mapped them to appropriate ontologies, including
  - the SPAR (Semantic Publishing and Referencing) Ontologies, and
  - IDO, the Infectious Disease Ontology

The MIIDI XML data model



Input Form for creating a MIIDI Report



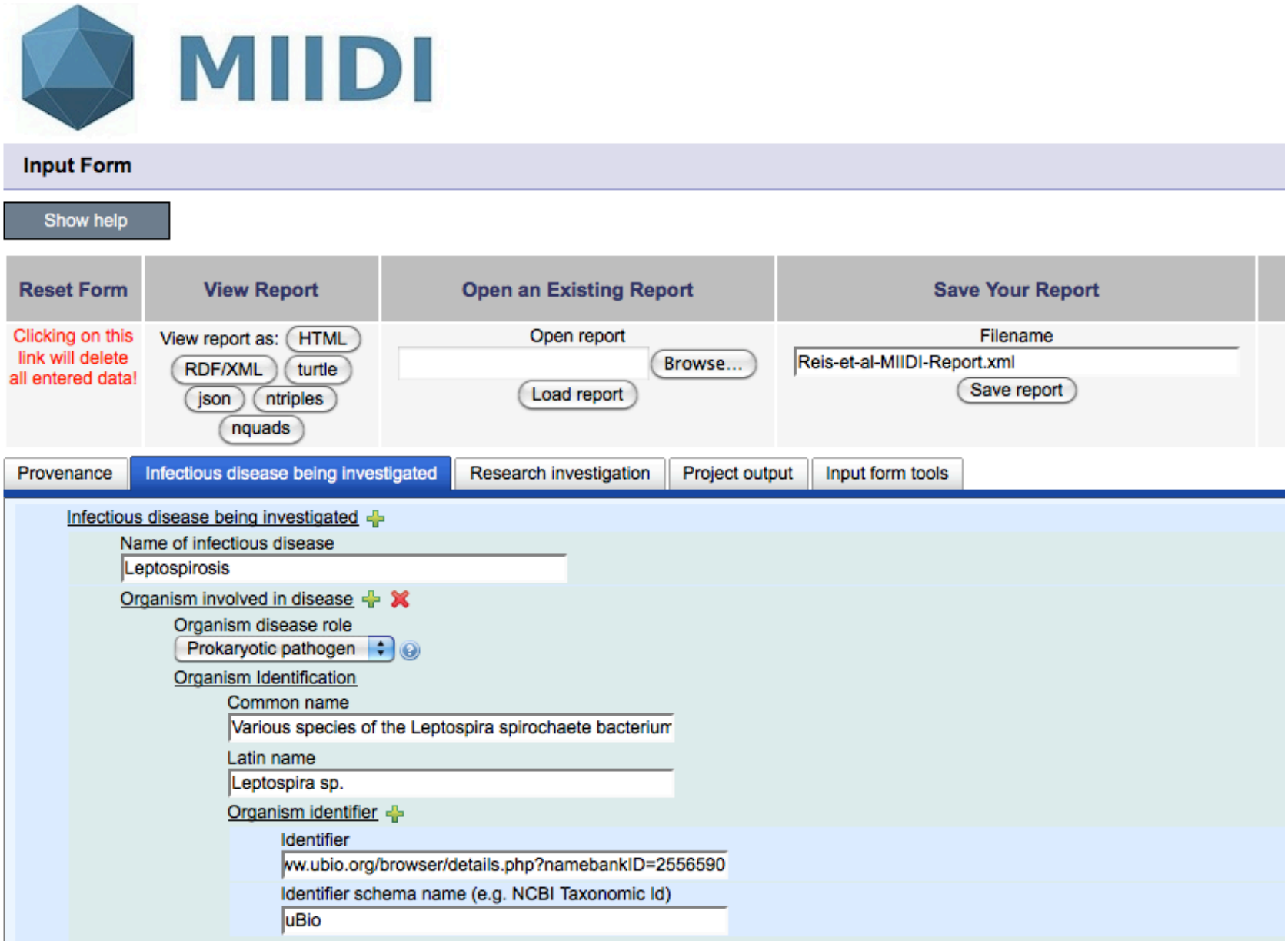
Advantages of the MIIDI metadata entry form

- The form is created directly from the XML data model
- A user fills in the MIIDI Form in a web browser at
  http://www.miidi.org:8080/input-form/
- The form can be customized for that user, so that personal metadata need only be entered once
- The user can save a MIIDI Report at any stage and complete it later

- The MIIDI Form eases the task of metadata creation by using a number of web services that provide metadata look-up, returning
  - bibliographic metadata in response to an input DOI or PubMed ID
  - geo-coordinates from a location on Google Maps, and
  - formal biomedical ontology terms via the BioPortal API

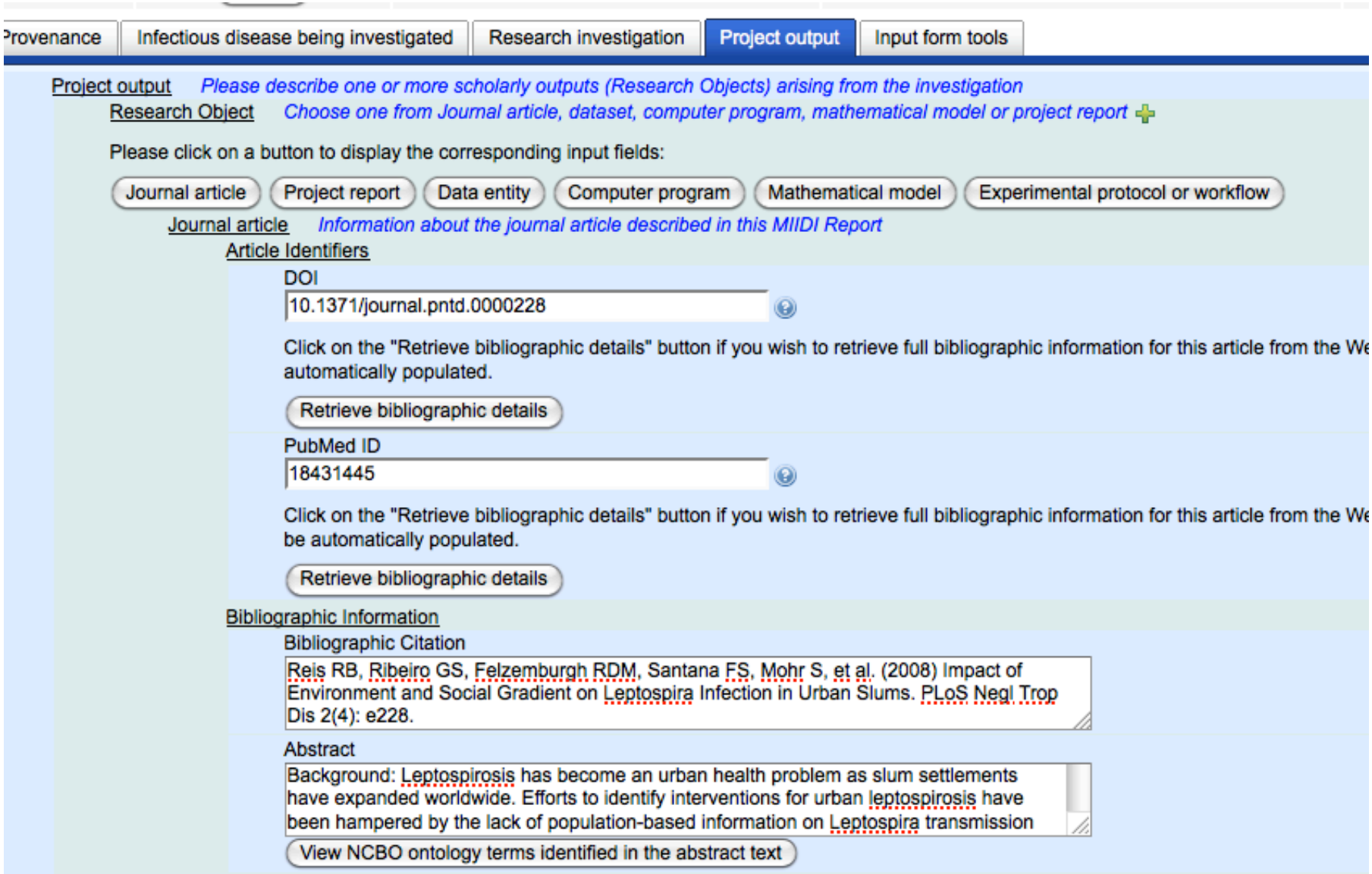
- The completed MIIDI Report is saved in XML, and can also be output in HTML, RDF/XML, JSON, Turtle and other RDF serializations

‘Disease’ section of the MIIDI Report for Reis *et al*. 2008



‘Investigation’ section of the MIIDI Report for Reis *et al*. 2008

‘Output’ section of the MIIDI Report for Reis *et al*. 2008



MIIDI Reports and Dryad

- Dryad now has a small but increasing number of affiliated infectious disease journals
- MIIDI Reports can now be used to create rich metadata for Dryad data submissions of infectious disease datasets
- Indeed, a MIIDI Report can describe the research investigation that has led *both* to the journal article *and* its associated Dryad datasets

- Outstanding questions:
  - Is the metadata content rich enough?
  - Is the form too complex and time-consuming?
  - Can we persuade authors to complete MIIDI Reports?
  - Do we need paid curators to do this work?
  - What added value will such metadata bring?

end

. . . with thanks again to the JISC for funding

JISC

The problem of access to the biomedical literature

- The free access to biomedical journals in developing countries offered by the HINARI Programme, set up in 2002 by WHO together with major publishers, is at risk

The Lancet Editorial, 22 January 2011:

DOI:10.1016/S0140-6736(11)60066-4

- "When news came last week that several large publishers—including Elsevier (our publisher), Lippincott Williams & Wilkins, and Springer—had withdrawn journals from HINARI's Bangladesh programme (and other countries too, such as Kenya and Nigeria), there was a collective cry of betrayal."
- "Elsevier says that Bangladesh is a country that could move to a 'discounted commercial agreement', and that there will be other countries too."
- "Our view is that any country designated as "low human development" by the UN justifies a clear and unambiguous commitment by all publishers to full and free access to research results through HINARI."

The vision: Open Research Reports

- The pre-existing ideas
  - of a structured digital abstract to encapsulate the basic facts in an infectious disease article,
  - of the MIIDI metadata standard to guide its encoding
- led to the first articulation of the vision for Open Research Reports
  - in January 2011, following the 'Beyond the PDF' meeting in San Diego, while discussing the *Lancet* editorial over dinner with Leslie Chan, Cameron Neylon and Peter Murray Rust

1 To get experts to create Open Research Reports for papers they read

- employing a tool they find easy to use, in a way that creates annotations that are also useful for their own personal use

2 To publish these reports in a subscription-free open access journal

- bringing the authors academic credit for a citable mini-publication

3 To tackle first the most cited papers for major infectious diseases