



**DataCite meeting**  
Describe, disseminate, discover:  
metadata for effective data citation  
British Library, 6 July 2 2012



## DataCite meets Linked Data

- DataCite metadata mapping to RDF – ‘DataCite2RDF’
- A Web entry form for DataCite metadata
- Data citations in the Open Citations Corpus

This presentation available at <http://bit.ly/LChj1H>

David Shotton, Silvio Peroni and Tanya Gray

Research Data Management and  
Semantic Publishing Research Group  
Department of Zoology  
University of Oxford, UK

***"It is a very sad thing  
that nowadays there is  
so little useless information"***

Oscar Wilde

e-mail: [david.shotton@zoo.ox.ac.uk](mailto:david.shotton@zoo.ox.ac.uk)



# An introduction to RDF and linked data

- The principles are quite simple
  - All entities (classes) and their relationships (properties) are identified and defined by unique URIs
  - URIs reference publicly available and commonly accepted structured vocabularies (ontologies)
  - Each relationship is expressed as a subject – predicate – object ‘triple’
  - The syntax defined by W3C’s Resource Description Framework (RDF)
- Examples:
  - `:my-dataset rdf:type fabio:Dataset .`
  - `:my-dataset dc:creator "Shotton, David" .`
  - `:my-dataset dc:title "Data citations, 2012" .`
- Such statements can be combined into interconnected information networks (RDF graphs) – forming ‘linked data’
  - the truth content of each original statement is maintained
  - thereby creating a web of knowledge, the Semantic Web



# Inadequacies of Dublin Core for DataCite metadata

## Dublin Core Mapping

The table below provides a mapping of the DataCite properties to the Dublin Core Simple elements and Qualified terms.<sup>7</sup>

ID	DataCite-Property	Dublin Core Simple Mapping (elements namespace)	Dublin Core Qualified Mapping (terms namespace)
1	Identifier	dc:identifier	dcterms:identifier
1.1	identifierType	dc:identifier	dcterms:identifier
2	Creator	dc:creator	dcterms:creator
2.1	creatorType	dc:creator	dcterms:creator
2.2	nameIdentifier	Not present in Dublin Core	Not present in Dublin Core
2.2.1	nameIdentifierScheme	Not present in Dublin Core	Not present in Dublin Core

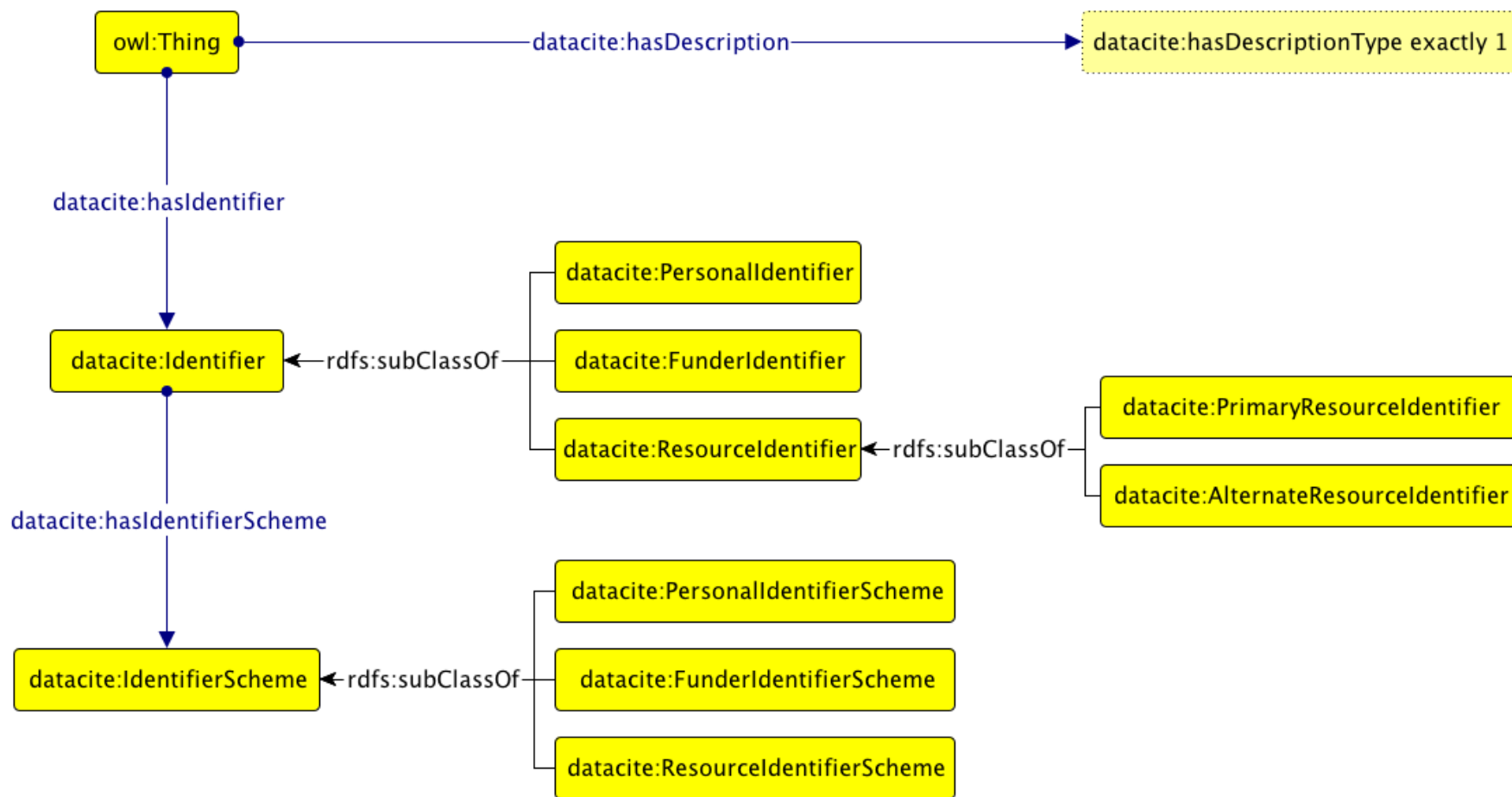
# The classes and properties of the DataCite Ontology

- The DataCite Ontology is designed to cover those aspects not well covered by other ontologies – particularly to permit the specification of various types of identifier required by the DataCite Metadata Kernel Scheme

<b>DataCite Classes</b>	<b>DataCite Object Properties</b>
<code>datacite:AlternateResourceIdentifier</code>	<code>datacite:hasDescription</code>
<code>datacite:DescriptionType</code>	<code>datacite:hasDescriptionType</code>
<code>datacite:FunderIdentifier</code>	<code>datacite:hasGeneralResourceType</code>
<code>datacite:FunderIdentifierScheme</code>	<code>datacite:hasIdentifier</code>
<code>datacite:Identifier</code>	<code>datacite:usesIdentifierScheme</code>
<code>datacite:IdentifierScheme</code>	
<code>datacite:PersonalIdentifier</code>	
<code>datacite:PersonalIdentifierScheme</code>	
<code>datacite:PrimaryResourceIdentifier</code>	
<code>datacite:ResourceIdentifier</code>	
<code>datacite:ResourceIdentifierScheme</code>	

- Available from <http://purl.org/spar/datacite/>, visualized as a human-readable web page using LODE, the Live OWL Documentation Environment (<http://www.essepuntato.it/lode>)

# Relationships between the DataCite Identifier and Identifier Scheme classes

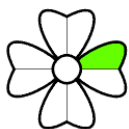


# SPAR (Semantic Publishing and Referencing) Ontologies

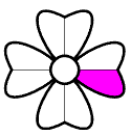
- The SPAR ontologies are described at <http://purl.org/spar/> and in my blog [Open Citations and Semantic Publishing](http://opencitations.wordpress.com) at <http://opencitations.wordpress.com>
  - Of these, six are relevant to what I will say today:



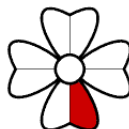
**The DataCite Ontology** <http://purl.org/spar/datacite/>



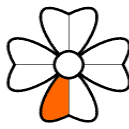
**CiTO, the Citation Typing Ontology** <http://purl.org/spar/cito/>, that enable characterization of the existence and the nature of citations



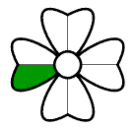
**FaBiO, the FRBR-aligned Bibliographic Ontology** <http://purl.org/spar/fabio/>, an ontology for describing bibliographic entities (books, articles, etc.)



**PRO, the Publishing Roles Ontology** <http://purl.org/spar/pro/>, an ontology for the roles of agents (e.g., author, editor, publisher, librarian) in the publication process, and the times during which those roles are held



**SCORO, the Scholarly Contributions and Roles Ontology** <http://purl.org/spar/scoro/>



**FRAPO, the Funders, Research Administration and Projects Ontology**, a CERIF-compliant ontology <http://purl.org/cerif/frapo/>

# An example of citation metadata in RDF

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

**dcterms:bibliographicCitation** "Reis RB, Ribeiro GS, Felzemburgh RDM, Santana FS, Mohr S, et al. (2008) Impact of environment and social gradient on Leptospira infection in urban slums. PLoS Negl Trop Dis 2(4): e228."

rdf:type **fabio:JournalArticle** ; # expression

frbr:realizationOf [ a **fabio:ResearchPaper** ] ; # work

**cito:cites** [http://dx.doi.org/10.1016/S0140-6736\(99\)80012-9](http://dx.doi.org/10.1016/S0140-6736(99)80012-9) # Reference [6];

**cito:obtainsBackgroundFrom**

[http://dx.doi.org/10.1016/S0140-6736\(99\)80012-9](http://dx.doi.org/10.1016/S0140-6736(99)80012-9) ;

**cito:sharesAuthorsWith** [http://dx.doi.org/10.1016/S0140-6736\(99\)80012-9](http://dx.doi.org/10.1016/S0140-6736(99)80012-9) .



# Mapping the DataCite Metadata Schema to RDF

- Starting data: DataCite Metadata Scheme Terms (v2.2)  
<http://test.datacite.org/schema/meta/kernel-2.2/index.html>
- Using the DataCite Ontology, other specialist SPAR ontologies, and standard vocabularies - Dublin Core, FOAF, and PRISM (Publishing Requirements for Industry Standard Metadata)
- Includes exemplar RDF usages
- Currently available as a Word document from

[http://sempublishing.svn.sourceforge.net/viewvc/sempublishing/DataCite/DataCiteMetadata2.2\\_mapping\\_to\\_RDF\\_09-07-2012.docx](http://sempublishing.svn.sourceforge.net/viewvc/sempublishing/DataCite/DataCiteMetadata2.2_mapping_to_RDF_09-07-2012.docx)



# A sample from the DataCite2RDF document

ID	DataCite property	Equivalent ontology class or property
1	Identifier	<p>datacite:Primary ResourceIdentifier (A sub-class of datacite:ResourceIdentifier that uses a datacite:IdentifierScheme that is restricted to datacite:doi, an individual in the datacite:ResourceIdentifierScheme)</p> <p><i>Exemplar usage:</i></p> <pre>:my-dataset rdf:type fabio:Dataset ; datacite:hasIdentifier [ rdf:type datacite:PrimaryResourceIdentifier ;   literal:hasLiteralValue "doi:10.1371/journal.pntd.0000228.g002.x001" ] .</pre>
1.1	IdentifierType	<p>Restricted to datacite:doi, an individual in the datacite:ResourceIdentifierScheme</p> <p><i>Exemplar usage:</i></p> <pre>:my-dataset rdf:type fabio:Dataset ; datacite:hasIdentifier [ rdf:type datacite:PrimaryResourceIdentifier ;   literal:hasLiteralValue "doi:10.1371/journal.pntd.0000228.g002.x001" ;   datacite:usesIdentifierScheme datacite:doi ] .</pre>



# Mapping the DataCite Metadata Schema to RDF

- Currently available as a Word document from

[http://sempublishing.svn.sourceforge.net/viewvc/sempublishing/DataCite/DataCiteMetadata2.2\\_mapping\\_to\\_RDF\\_09-07-2012.docx](http://sempublishing.svn.sourceforge.net/viewvc/sempublishing/DataCite/DataCiteMetadata2.2_mapping_to_RDF_09-07-2012.docx)

- Feedback and comments welcome !

- Problem: Poor understanding of how DataCite wants to use

- 17 Description

- 17.1 Description type

since in the XML example it related not to the dataset that was the target of all the other metadata, but to the XML example itself.

# A fragment from the DataCite XML example

## XML Example

This XML example conforms to the XML schema. More examples for various object types can be found at <http://schema.datacite.org/meta/kernel-2.2/index.html>.

```
<resource xmlns="http://datacite.org/schema/kernel-2.2"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://datacite.org/schema/kernel-2.2
    http://schema.datacite.org/meta/kernel-2.2/metadata.xsd">
  <identifier identifierType="DOI">10.1594/WDCC/CCSRNIES_SRES_B2</identifier>
  <creators>
    <creator> <creatorName>Miller, John</creatorName>
    </creator> <creator>
    <creatorName>Smith, Jane</creatorName> <nameIdentifier
  nameIdentifierScheme="ISNI">1422 4586 3573
  0476</nameIdentifier> </creator>
  </creators> <titles>
    <title>National Institute for Environmental Studies and Center for Climate System Research
    Japan</title>
    <title titleType="Subtitle">A survey</title>
```

## The same bit of the XML example mapping to RDF

```
<http://dx.doi.org/10.1594/WDCC/CCSRNIES_SRES_B2>
  rdf:type fabio:Dataset ;
  datacite:hasIdentifier [ rdf:type datacite:PrimaryResourceIdentifier ;
    literal:hasLiteralValue "10.1594/WDCC/CCSRNIES_SRES_B2" ;
    datacite:usesIdentifierScheme datacite:doi ] ;
# Note: fictitious names. Real author: Nozawa, Toru.
  dcterms:creator [rdf:type foaf:Person ; foaf:name "Miller, John" ] ;
  dcterms:creator [rdf:type foaf:Person ; foaf:name "Smith, Jane" ;
    datacite:hasIdentifier
      [rdf:type datacite:PersonalIdentifier ;
        literal:hasLiteralValue "1422 4586 3573 0476" ;
        datacite:usesIdentifierScheme datacite:isni ] ] ;
  dcterms:title "National Institute for Environmental Studies and Center
for Climate System Research Japan" ;
  fabio:hasSubtitle "A survey" ;
```

- Note now simple and compact the RDF representation is
- This RDF version of the DataCite XML example is available on-line in Turtle format at [http://sempublishing.svn.sourceforge.net/viewvc/sempublishing/DataCite/RDF\\_mapping\\_of\\_DataCitev2.2\\_XML\\_example.ttl](http://sempublishing.svn.sourceforge.net/viewvc/sempublishing/DataCite/RDF_mapping_of_DataCitev2.2_XML_example.ttl)


# A Web form for entry of DataCite metadata

<http://www.miidi.org:8080/datacite/>

## DataCite Mandatory Properties

[1] Identifier *A persistent identifier that identifies a resource. Currently, only DOI is allowed.*

[1.1] Identifier type

1. [2] **Creator of Data Collection**  *Name the creator(s) of the dataset being annotated, in priority order, or the corporate/institutional name or a personal name. Use + to add additional names if there are multiple authors.*

[2.1] Creator name *Format for personal names: FamilyName, GivenName.*

1. [2.2] **Personal identifier**  *(text string, e.g. 0137-1963-7688-2319)*

[2.2.1] Personal identifier scheme

1. [3] **Title**  *A name or title by which a resource is known.*

[3.1] Title type

*Select from the drop down list*

[4] **Publisher** *(including archives as appropriate) or institution which submitted the work. Any others in the citation, so called "secondary publishers", should be listed in the optional field. Examples: World Data Center for Climate (WDCC); GeoForschungsZentrum Potsdam. The publisher is required to mean making the data available to the community of researchers.*

[5] **Publication year** *Year when the data is made publicly available. If an embargo period has been in effect, use the date*

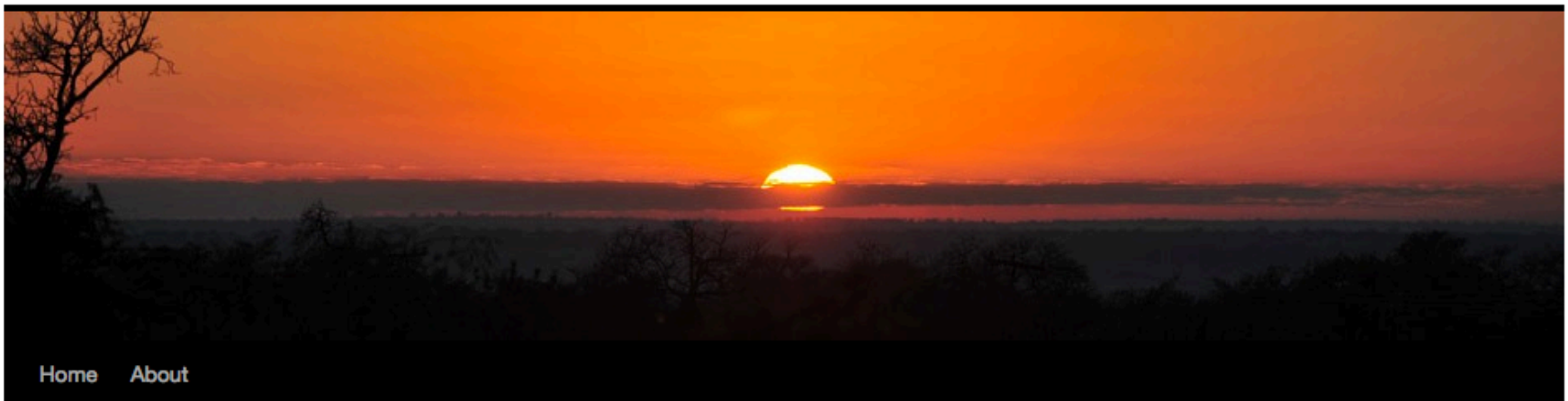


# How to cite data

## Open Citations and Semantic Publishing

<http://opencitations.wordpress.com/2011/06/30/how-to-cite-data/>

*Enhancing scholarly communication, publishing research data, and exposing bibliographic citations as Open Linked Data – tools, ontologies and recommendations.*



← Questions of granularity – Dryad's use of DataCite DOIs for data citation, and the Annotation Ontology

Pensoft Journals policy and author guidelines on data publication and citation →

### How to cite data

Posted on [June 30, 2011](#) by [davidshotton](#)

As an approach towards developing best practice for data citation, I recently wrote a [Data Citation Best Practice Discussion Document](#) that is available on Google Docs, and that I have now slightly revised to Version 2 [1].

#### Recent Posts

- [Oxford University Press to support Open Citations](#)
- [Open Citations and Semantic Publishing](#)
- [Science joins Nature in opening reference citations](#)
- [Access to Citation Data](#)
- [Nature to open its reference](#)



# How to cite data

- Proper data citations require *both* an **in-text citation and reference pointer**, and a **proper data reference in the reference list**

- **Example in-text citation and reference pointer:**

"The raw data underpinning this analysis are deposited in the Dryad Data Repository at <http://dx.doi.org/10.5061/dryad.8684> (Vijendravarma et al., 2011).

- **Example data reference in reference list:**

[32] 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. doi:10.5061/dryad.8684. <http://dx.doi.org/10.5061/dryad.8684>."

- The reference lists from all **204,637 articles** in the Open Access Subset of PMC on 24 January 2011, encoded in RDF using the SPAR ontologies
- These lists contain **6,325,178 individual references**, some unique, but many from different citing articles to highly cited papers
- These references cite **3,373,961 papers** outside the Open Access Subset
  - ~ **20% of all PubMed Central papers** (approx. 3,200,000 papers)
  - includes **ALL** the highly cited papers in every biomedical field
- Each reference list is maintained as a distinct unit, by encoding it as a Named RDF Graph with a unique URI
- Encoded these bibliographic records and the citations between them in RDF, creating **236,499,781 quads** occupying 2.1 gigabytes of compressed storage
- Freely available under a CC0 waiver from <http://opencitations.net/data/>
- The complete corpus can be downloaded, or can be queried via a SPARQL endpoint



# Viewing citation networks at <http://opencitations.net>

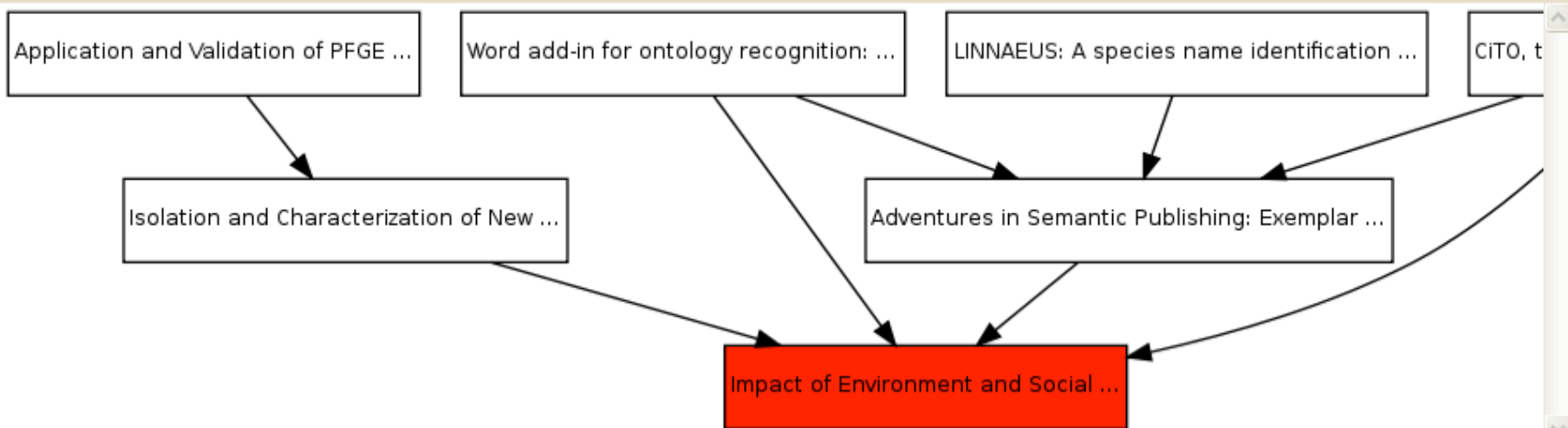
## JISC Open Citations

[Home](#) [About](#) [Journals](#) [Articles](#) [Search](#) [SPARQL](#) [Source code](#) [Source data](#) [Contact](#)

Citation network for the article:

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

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





## Expanding the OCC to include data citations

- With new funding from the **JISC**, and in partnership with **CrossRef**, we now wish to
  - expand the Open Citations Corpus to include references from articles in subscription access journals, in addition to open access journals
  - harvest these on an on-going basis
  - *Nature*, *Science* and Oxford University Press are already signed up
- As part of this expansion, **we would also like to partner with DataCite**, to include within the corpus all DataCite citation metadata of datasets citing journal articles, and of journal articles citing datasets
  - We would like to harvest these as XML on a monthly basis,
  - transform the citations to RDF using the DataCite2RDF mapping
  - and include the DataCite citations in the Open Citations Corpus

# DataFlow data management services



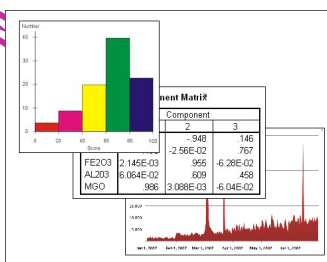
Researchers

<http://www.dataflow.ox.ac.uk/>



DataStage file system

Zipped BagIt Data Package with  
RDF metadata manifest

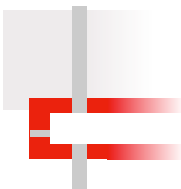


SWORD deposit protocol

Researchers, other users



DataBank repository



end