

Digital Object Identifiers for Tracking Datasets

Matthew Viljoen

Big Data Management Workshop

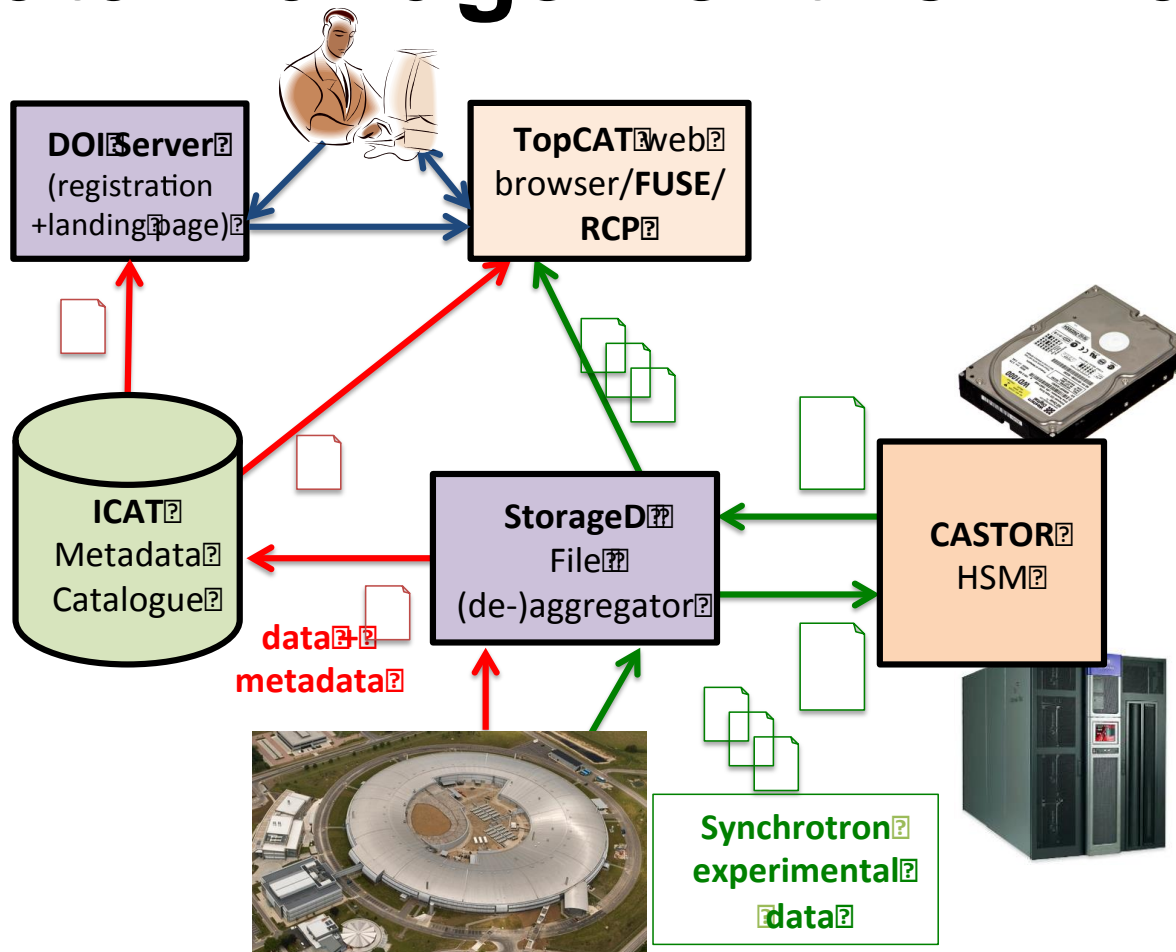
Imperial College, London
27-28 June 2013



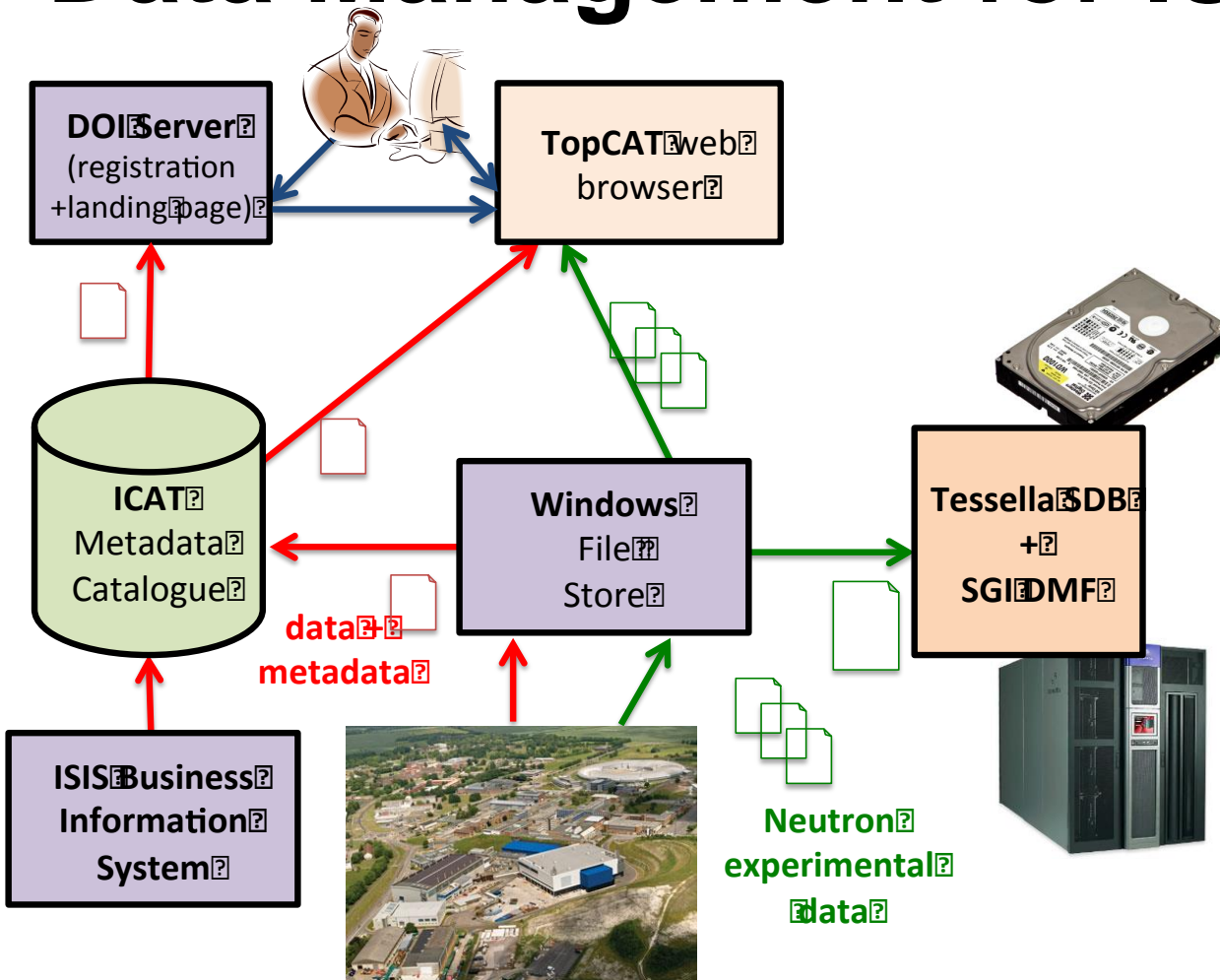
Big Data at RAL

- Solutions using CASTOR, DMF, SDB, Panasas and home grown
- Primarily Linux based. ORACLE SL8500 robot with T10K(A,B,C)
- 18PB on tape and 9PB on disk (CASTOR) 6PB on disk (Panasas)
- Users:
 - High Energy Particle Physics (CERN users)
 - STFC Facilities (Diamond Synchrotron, ISIS Neutron Source, ...)
 - Systems backup
- Complete end-to-end data solution offered for Facilities:
 - Data ingest, data archival, metadata, portal for data retrieval and DOI services

Data Management for Diamond



Data Management for ISIS



DOIs and Citing Data

- Digital Object Identifiers for Data. Citation of data just like papers/journals and articles.

***Citable* data brings together publications and their data**



Encourages easier scrutiny & new research



better value for money for taxpayers!

DataCite Service

- Not for profit organization
- DataCite provides means to:
 - mint, sustain, discover DOIs
- We:
 - associate DOI with data
 - provide metadata and landing page
- www.datacite.org



DataCite Service

Australia

- Australian National Data Service (Member)

Canada

- Canada Institute for Scientific and Technical Information (Member)

Denmark

- Technical Information Center of Denmark (Member)

France

- Institute for Scientific and Technical Information - INIST (Member)

Germany

- German National Library of Science and Technology - TIB (Member)
- German National Library of Medicine - ZB MED (Member)
- GESIS - Leibniz Institute for Social Science (Member)
- German National Library of Economics - ZBW (Member)

Netherlands

- TU Delft Library (Member)

Republic of Korea

- Korea Institute of Science and Technology Information - KISTI (Associate Member)

Sweden

- The Swedish National Data Service - SNDS (Member)

Switzerland

- ETH Zurich (Member)

United Kingdom

- The British Library (Member)
- Digital Curation Centre (Associate Member)

United States

- California Digital Library (Member)
- Office of Scientific and Technical Information, US Department of Energy (Member)
- Purdue University Libraries (Member)
- Interuniversity Consortium for Political and Social Research - ICPSR (Associate Member)
- Microsoft Research (Associate Member)

<http://www.datacite.org/members>

Scientific Computing
Department



Science & Technology
Facilities Council

Using the DataCite Metadata Store

1. Register for an account with a DataCite member
 - In the UK - the British Library
2. Receive: user name, password to store service
 - For STFC - DOI:10.5286/
3. Write software to call the API
4. Create landing page for each DOI
5. Mint DOI for each dataset through the API
 - e.g. DOI:10.5286/ISIS.E.24079627
6. Assign some suitable metadata to the DOI

Metadata Store API 1/2

- **DOI API-RESTful**

- **GET**

- *URI: <https://mds.datacite.org/doi/{doi}> where {doi} is a specific DOI.*

- This request returns an URL associated with a given DOI.

- **POST**

- *URI: <https://mds.datacite.org/doi>*

- POST will mint new DOI if specified DOI doesn't exist.
 - This method will attempt to update URL if you specify existing DOI.
 - A Data centre's doiQuotaUsed will be increased by 1.
 - A new record in Datasets will be created.

See <https://mds.datacite.org/static/apidoc>

Metadata Store API 2/2

- **DOI API-RESTful**

- GET**

- *URI: <https://mds.datacite.org/metadata/{doi}> where {doi} is a specific DOI.*

- This request returns the most recent version of metadata associated with a given DOI.

- POST**

- *URI: <https://mds.datacite.org/metadata>*

- This request stores new version of metadata. The request body must contain valid XML.

- DELETE**

- *URI: <https://mds.datacite.org/metadata/{doi}>*

- This request marks a dataset as 'inactive'.
 - To activate it again, POST new metadata.

See <https://mds.datacite.org/static/apidoc>

DataCite Metadata Schema

- Current Version 2.2, July 2011
- Available at: doi:10.5438/0005

ID Mandatory Property

- 1 Identifier
- 2 Creator
- 3 Title
- 4 Publisher
- 5 PublicationYear

ID Optional Property

- 6 Subject
- 7 Contributor
- 8 Date
- 9 Language
- 10 ResourceType
- 11 AlternateIdentifier
- 12 RelatedIdentifier
- 13 Size
- 14 Format
- 15 Version
- 16 Rights
- 17 Description

<http://schema.datacite.org/meta/kernel-2.2/index.html>



Accessing Data via DOI – Landing Page

The screenshot shows a web browser window displaying the STFC Data home landing page. The browser's address bar shows the URL: <https://data.isis.stfc.ac.uk/doi/INVESTIGATION/24079886/>. The page header features the Science & Technology Facilities Council logo and navigation links for Media enquiries, Careers, Log in, and Sign up. A search bar is also present. Below the header is a navigation menu with categories: Funding, Research, Innovation, Skills, Public Engagement, News, Events and Publications, and About. The main content area displays the DOI: RB920411. The investigation title is "Responsive polymer brushes grafted from gold nanoparticles: interaction with surfactant." The creator is listed as Titmuss, S and Jia, H. The DOI is 10.5286/ISIS.E.24079886. The date of experiment is Sat May 01 08:45:00 BST 2010. The publisher is STFC ISIS Facility. The data format is RAW/Nexus. A download button is available. A data citation is provided: [author], [date], [title], [publisher], [doi].

STFC Data home | Science & Technology Facilities Council | STFC

STFC Data home | Science & Tec...

<https://data.isis.stfc.ac.uk/doi/INVESTIGATION/24079886/>

Media enquiries Careers Log in Sign up

search

STFC Programmatic Review 2012-

Funding Research Innovation Skills Public Engagement News, Events and Publications About

[ISIS Data](#)

RB920411.

Investigation title: Responsive polymer brushes grafted from gold nanoparticles: interaction with surfactant.

Creator: *Titmuss, S*
Jia, H

DOI: 10.5286/ISIS.E.24079886

Date of Experiment: Sat May 01 08:45:00 BST 2010

Publisher: STFC ISIS Facility

Data format: [RAW/Nexus](#)
Select the data format above to find out more about it.

Data Citation
The recommended format for citing this dataset in a research publication is as:
[author], [date], [title], [publisher], [doi]

DOWNLOAD
download the dataset

Data collected on SANS2D instrument at the ISIS facility

Accessing Data via DOI – Data Portal

The screenshot displays the TOPCAT web tool interface. The browser address bar shows the URL: <https://data.isis.stfc.ac.uk/TOPCATWeb.jsp#view///&tab=MyData///&Mode>. The main content area is titled "Investigation: Responsive polymer brushes grafted from gold nanoparticles: interaction with surfactant". A "Datafile Window" is open, showing a table of datasets. The table has columns for File Name, File Location, File Size, Format, Format Version, Format Type, and Create Time. The data is paginated, showing Page 1 of 55, with 20 items displayed.

File Name	File Location	File Size	Format	Format Version	Format Type	Create Time
Dataset Name: Default (20 Items)						
<input type="checkbox"/> SANS2D0000...	Wisis\instS\Instr...	14.35 MB	isis neutron raw	2	binary	5/1/10 10:32 AM
<input type="checkbox"/> SANS2D0000...	Wisis\instS\Instr...	0.007 MB				5/1/10 10:32 AM
<input type="checkbox"/> SANS2D0000...	Wisis\instS\Instr...	1.675 MB				5/1/10 10:40 AM
<input type="checkbox"/> SANS2D0000...	Wisis\instS\Instr...	0.004 MB				5/1/10 10:31 AM
<input type="checkbox"/> SANS2D0000...	Wisis\instS\Instr...	0 MB				5/1/10 10:32 AM
<input type="checkbox"/> SANS2D0000...	Wisis\instS\Instr...	0.001 MB				5/1/10 10:32 AM
<input type="checkbox"/> SANS2D0000...	Wisis\instS\Instr...	0 MB				5/1/10 10:32 AM
<input type="checkbox"/> SANS2D0000...	Wisis\instS\Instr...	0 MB				5/1/10 10:33 AM
<input type="checkbox"/> SANS2D0000...	Wisis\instS\Instr...	0.002 MB				5/1/10 10:32 AM
<input type="checkbox"/> SANS2D0000...	Wisis\instS\Instr...	0.006 MB	isis neutron raw	2	binary	5/1/10 11:15 AM
<input type="checkbox"/> SANS2D0000...	Wisis\instS\Instr...	0.005 MB				5/1/10 11:15 AM
<input type="checkbox"/> SANS2D0000...	Wisis\instS\Instr...	0.518 MB				5/1/10 11:24 AM

Recommended Way of Citing Data

Creator (PublicationYear): Title. Publisher.
Identifier

Example:

Easton,S; Barnes,C H W; Ionescu,A; (2011):
RB820232: Magnetic moment of EuO in spin
filtering magnetic tunnel structures.; STFC ISIS
Facility. [doi:10.5286/ISIS.E.24066298](https://doi.org/10.5286/ISIS.E.24066298)

Implementation Issues 1

The data granularity question

- DOIs currently issued by experiment.
- What about finer granularity?
 - Per visit?
 - file?
 - arbitrary?



Implementation Issues 2

When to publish DOIs?.

The minimum metadata are available from the registry when data are collected:

- before publication of results
- before expiration of embargo period.

If this is too soon, should DOI be issued later ?

- When works is published ?
- Later ?

Implementation Issues 3

Just how persistent is persistent?

Guarantees of a DOI. Identifier guarantees its own persistence.

Not that the data is actually there ... Or unchanged.

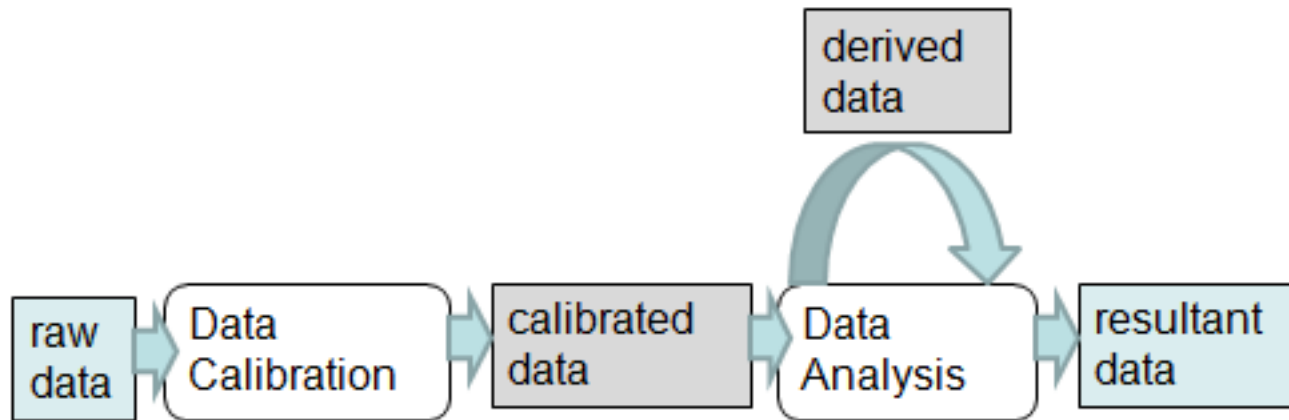


Implementation Issues 4

Content of the accessed data

What data should be released?

Raw/calibrated/derived data? Some combination of the above?



Current Status

- ISIS Neutron Source:
 - 1690 DOIs issued to date
 - Issued manually by staff. Next step is user generation
- Diamond Synchrotron:
 - Software ready
 - Pending managerial and policy decisions



Acknowledgements

Implementation:

- Sri Nagella, Antony Wilson

Borrowed slides and supplementary information:

- Michael Wilson, Brian Matthews, Tom Griffin



Time for a Demo?

Questions

matthew.viljoen@stfc.ac.uk

