Digital Object Identifiers for Tracking Datasets

Matthew Viljoen

Big Data Management Workshop

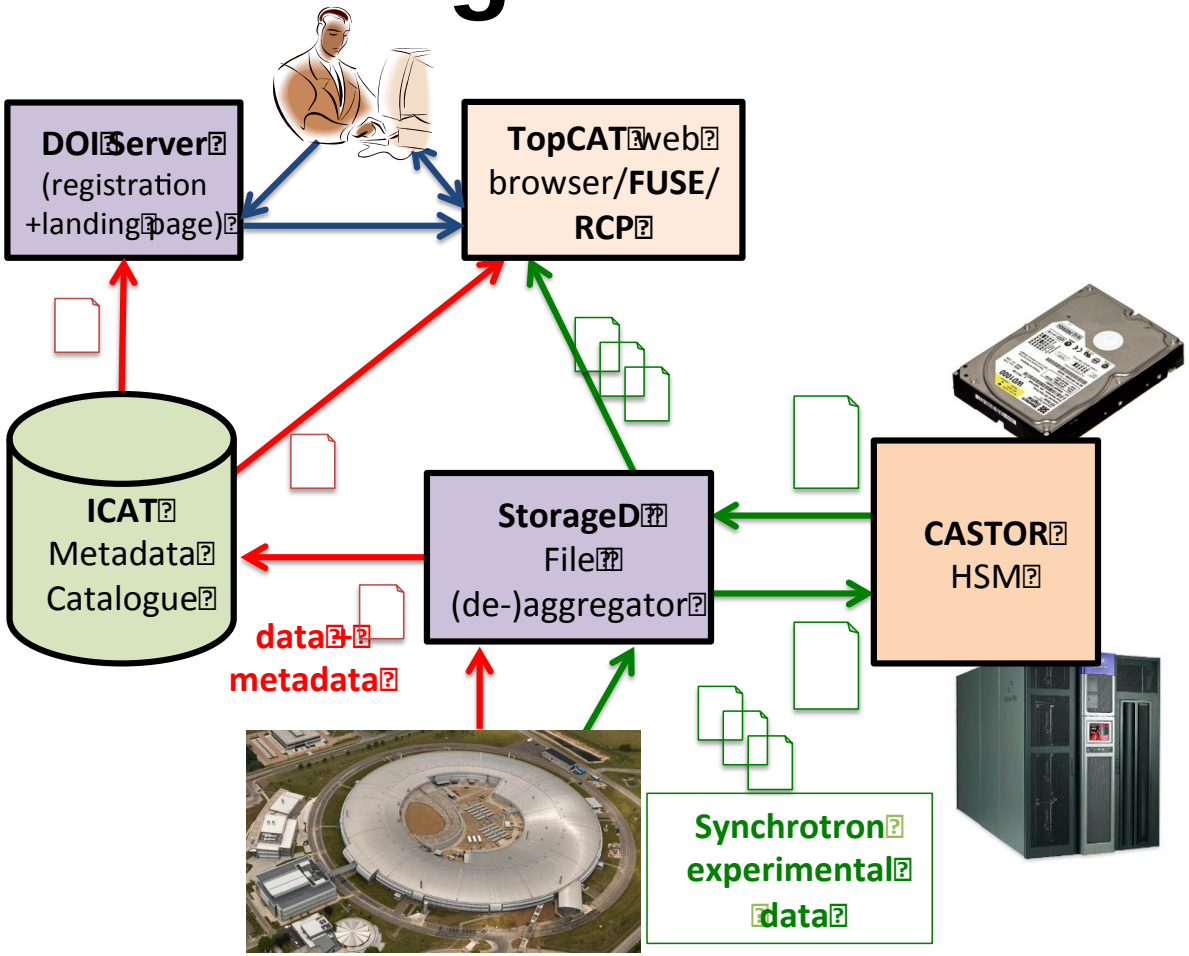
Imperial College, London
27-28 June 2013

Scientific Computing Department
Science & Technology Facilities Council

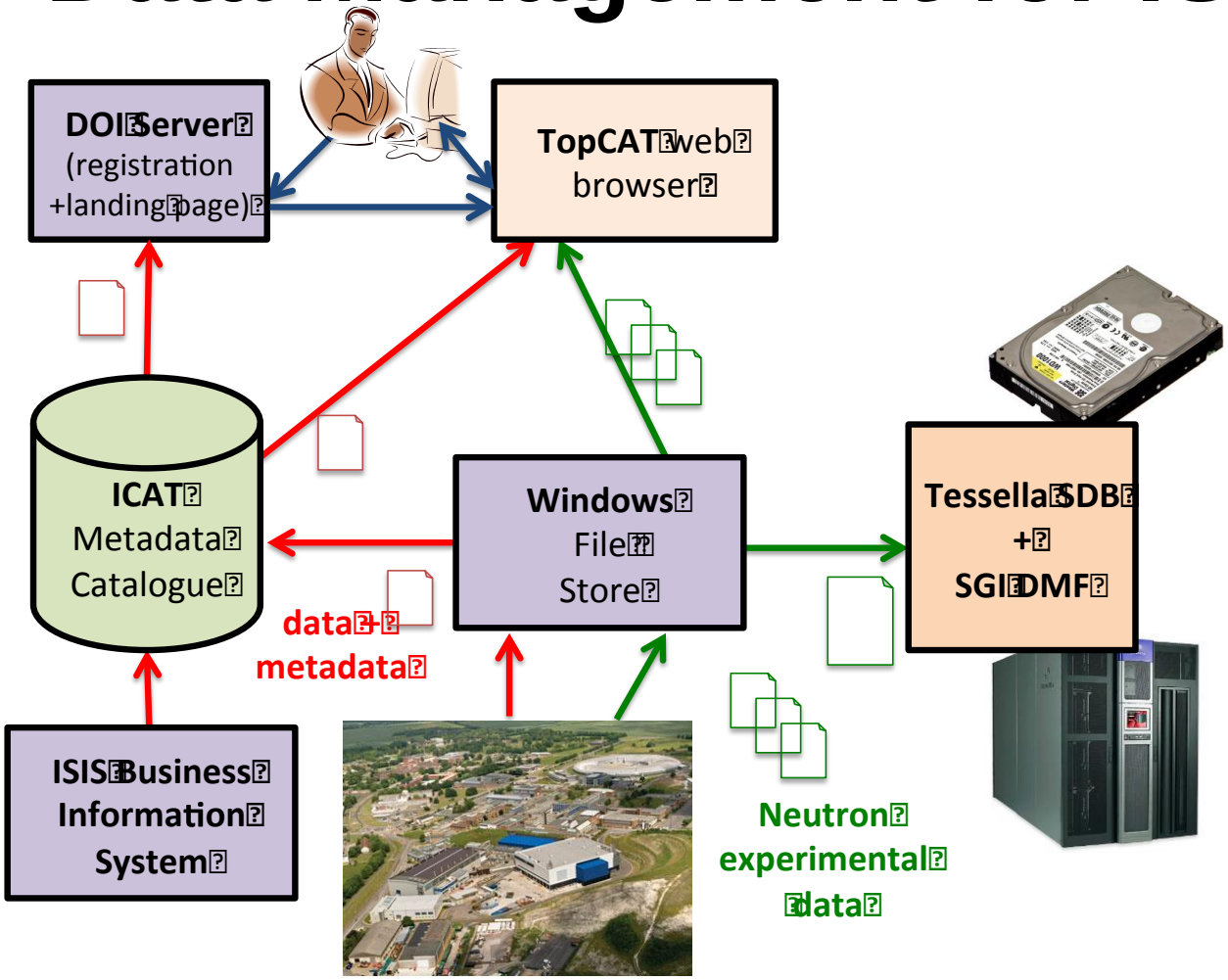
Big Data at RAL

- 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

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

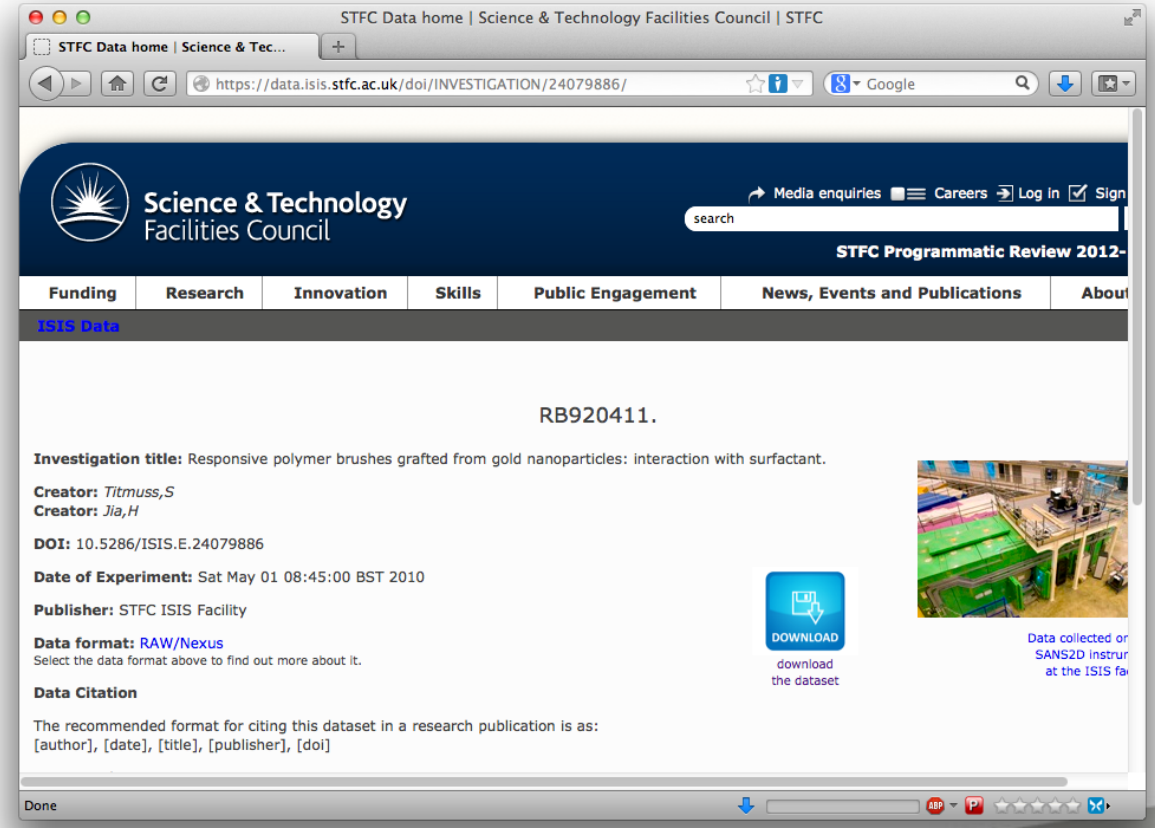
- Current Version 2.2, July 2011
- Available at: doi:10.5438/0005

ID	Mandatory Property
1	Identifier
2	Creator
3	Title
4	Publisher
5	PublicationYear

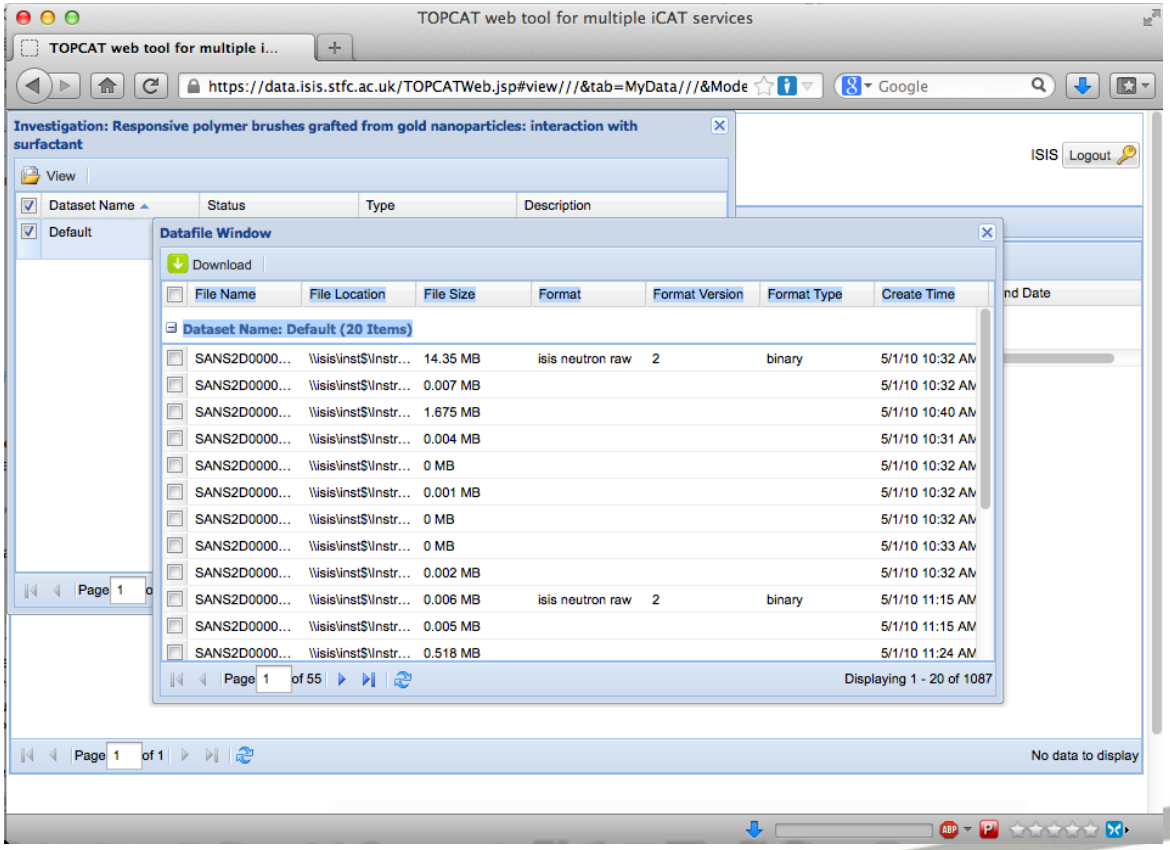
ID	Optional Property
6	Subject
7	Contributor
8	Date
9	Language
10	ResourceType
11	AlternateIdentifier
12	RelatedIdentifier
13	Size
14	Format
15	Version
16	Rights
17	Description

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

Accessing Data via DOI – Landing Page



Accessing Data via DOI – Data Portal



Scientific Computing Department

Science & Technology Facilities Council

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

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.

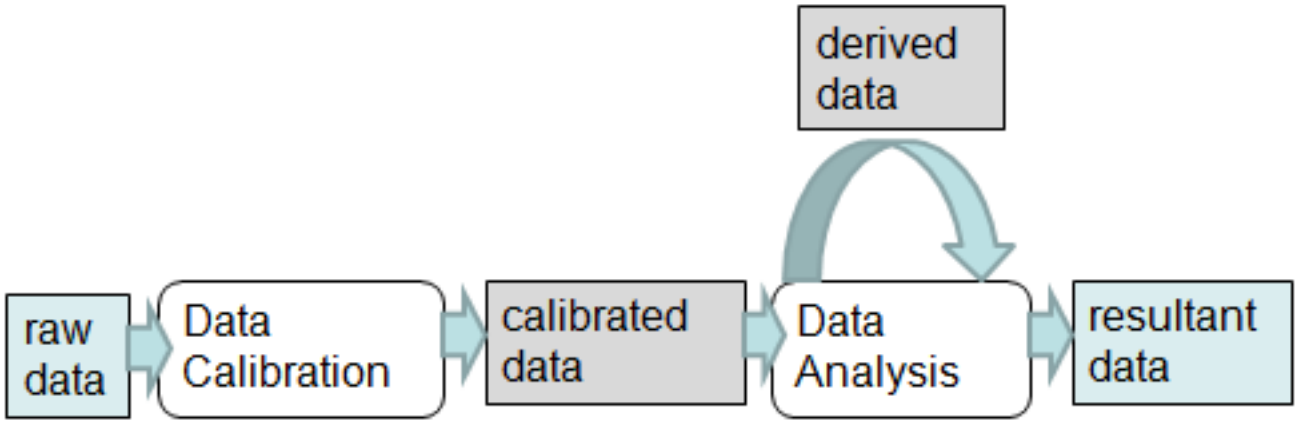
Not that the data is actually there … Or unchanged.

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