Research Data meets Research Information:
Pure CERIF-CRIS & Eprints with CERIF extensions
“… that there had been "a consistent pattern of failing to display the proper degree of openness". Scientists also failed to appreciate the risk their lack of transparency posed to the university and "indeed to the credibility of UK climate science".

The Guardian, Thursday 8 July 2010
Call to arms on data integrity

18 July 2013 Patrick Walter

Data integrity in the chemical sciences is under the spotlight. In a broadside aimed at shoddy data practices, the editor of Organic Letters, Amos Smith, has called on chemists to put their house in order and root out poor practice. But with the focus now on this murky issue the question remains just how widespread manipulation of raw data actually is in the chemical sciences?

Smith declined to be interviewed by Chemistry World but in his editorial he explains that his journal has hired an analyst to scrutinise supplementary data more closely. As a result a number of instances have been discovered where raw data, such as spectra, have been 'cleaned up', in some cases to remove evidence of impurities.

The number of cases of image manipulation brought before the US Office of Research Integrity has been growing steadily.
“Unquestionably, a significant contributor to failure in oncology trials is the quality of published preclinical data.” …
“scientific findings were confirmed in only 6 (11%) cases. Even knowing the limitations of preclinical research, this was a shocking result.”
‘… three-quarters of published scientific papers in the field of machine learning are bunk …’, says Sandy Pentland, a computer scientist at MIT’

Economist, Oct 19th, 2013

“Publishing data in a reusable form to support findings must be mandatory “
President Barack Obama signed an order Thursday requiring federal agencies to make their publicly accessible data open and machine-readable. Obama's order, which enforces a new Open Data Policy, was issued in hopes entrepreneurs will use government data...
UK Open Access policies

• RCUK policy from 01/04/2013
  Papers should also include a statement on how the underlying research materials - such as data, samples or models – can be accessed. It is not necessary that the data itself be openly accessible if there are compelling reasons against this

• REF2020
  To be eligible for the next Research Excellence Framework, peer-reviewed manuscripts must be deposited in an institutional or subject repository on acceptance for publication. These requirements apply to all journal articles and conference proceedings accepted for publication after 1st April 2016.

Didn’t go as far as saying the data underlying these publications should be open.
RCUK Common Principles on Data Policy

Making research data available to users is a core part of the Research Councils’ remit and is undertaken in a variety of ways. We are committed to transparency and to a coherent approach across the research base. These RCUK common principles on data policy provide an overarching framework for individual Research Council policies on data policy.

Principles

- Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property.
- Institutional and project specific data management policies and plans should be in accordance with relevant standards and community best practice. Data with acknowledged long-term value should be preserved and remain accessible and usable for future research.
- To enable research data to be discoverable and effectively re-used by others, sufficient metadata should be recorded and made openly available to enable other researchers to understand the research and re-use potential of the data. Published results should always include information on how to access the supporting data.
- RCUK recognises that there are legal, ethical and commercial constraints on release of research data. To ensure that the research process is not damaged by inappropriate release of data, research organisation policies and practices should ensure that these are considered at all stages in the research process.
- To ensure that research teams get appropriate recognition for the effort involved in collecting and analysing data, those who undertake Research Council funded work may be entitled to a limited period of privileged use of the data they have collected to enable them to publish the results of their research. The length of this period varies by research discipline and, where appropriate, is discussed further in the published policies of individual Research Councils.
- In order to recognise the intellectual contributions of researchers who generate, preserve and share key research datasets, all users of research data should acknowledge the sources of their data and abide by the terms and conditions under which they are accessed.
- It is appropriate to use public funds to support the management and sharing of publicly-funded research data. To maximise the research benefit which can be gained from limited budgets, the mechanisms for these activities should be both efficient and cost-effective in the use of public funds.
<table>
<thead>
<tr>
<th>Research Funders</th>
<th>Published outputs</th>
<th>Data</th>
<th>Time limits</th>
<th>Data plan</th>
<th>Access/sharing</th>
<th>Long-term curation</th>
<th>Monitoring</th>
<th>Guidance</th>
<th>Repository</th>
<th>Data centre</th>
<th>Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>AHRC</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>BBSRC</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>CRUK</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>EPSRC</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>ESRC</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>MRC</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>NERC</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>STFC</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Wellcome Trust</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>

http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies
• By **May 2015**, EPSRC has 9 clear *expectations of organisations* receiving EPSRC funding, including:
• ‘.. that *appropriately structured metadata* describing the research data they hold is published (normally within **12 months** of the data being generated) and made **freely accessible on the internet**; in each case the metadata *must be sufficient to allow others to understand what research data exists, why, when and how it was generated, and how to access it.*’
• Where access to the *data is restricted* the published metadata should also give the *reason* and summarise the conditions which must be satisfied for access to be granted. For example ‘commercially confidential’ data, in which a business organisation has a legitimate interest, might be made available to others subject to a suitable legally enforceable non-disclosure agreement.
• ‘.. that EPSRC-funded research data is *securely preserved for a minimum of 10-years from the date that any researcher ‘privileged access’ period expires or, if others have accessed the data, from last date on which access* to the data was requested by a third party; all reasonable steps will be take to ensure that publicly-funded data is not held in any jurisdiction where the available legal safeguards provide lower levels of protection than are available in the UK.’
So ... managing research data is important for ...

- Future funding
- Increased impact/citation of research
- Scientific integrity (reproducibility)
- Re-use and sharing of data -> innovation
- Aligned with University strategy
- Future national assessment exercises – beyond REF2020
So what have we been doing?

Components of a data management service: Digital Curation Centre:
http://www.dcc.ac.uk/resources/how-guides
The Discipline of Creativity: Exploring the Paradox

Lead Research Organisation: University of St Andrews
Department Name: Management

Publications


Barbara Townley (2009) Managing in the creative industries: Managing the motley crew in Human Relations

Barbara Townley (2010) What is a creative field?
<table>
<thead>
<tr>
<th>Project Award Code</th>
<th>Project No</th>
<th>Project Name</th>
<th>Principal Investigator</th>
<th>Funder's Name</th>
<th>Funder Ref</th>
<th>Lead Dept</th>
</tr>
</thead>
<tbody>
<tr>
<td>55242</td>
<td>1</td>
<td>Monolithic Resonant TeraHertz Detectors</td>
<td>David Cumming</td>
<td>Engineering &amp; Physical Sciences Research Council (EPSRC)</td>
<td>EP/I017461/1</td>
<td>ENG - ENGINEERING ELECTRONICS &amp; NANO ENG</td>
</tr>
<tr>
<td>58833</td>
<td>1</td>
<td>Triple wavelength superspectral camera focal-plane array (SUPERCAMERA)</td>
<td>David Cumming</td>
<td>Engineering &amp; Physical Sciences Research Council (EPSRC)</td>
<td>EP/J018678/1</td>
<td>ENG - ENGINEERING ELECTRONICS &amp; NANO ENG</td>
</tr>
<tr>
<td>58833</td>
<td>2</td>
<td>Triple wavelength superspectral camera focal-plane array (SUPERCAMERA)</td>
<td>David Cumming</td>
<td>Engineering &amp; Physical Sciences Research Council (EPSRC)</td>
<td>EP/J018678/1</td>
<td>ENG - ENGINEERING ELECTRONICS &amp; NANO ENG</td>
</tr>
</tbody>
</table>
Common approach

• Pragmatic
  – Joining up data silos and reusing data
  – Using standards and authority sources
  – Using carrots as well as sticks
  – Demonstrating added value to researchers, managers and funders

• Communication is key
Adding datasets: extending our systems

- Institutional dataset register/catalogue
- Metadata harmonisation
  - Cerif for Datasets * Datacite * Pure & EPrints UG
- UK National data registry pilot
  - JISC * DCC * Institutions * Funders * Data Centres
  - CASRAI-UK working group
User groups working together

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>ReCollect field</td>
<td>Description</td>
<td>Description of the dataset.</td>
<td>There is not consensus whether this is optional or mandatory</td>
<td>Description of the dataset - most likely an abstract</td>
<td>DataCite Metadata 3.0</td>
</tr>
<tr>
<td></td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>ReCollect field</td>
<td>Keywords</td>
<td>Keyword or key phrase describing the resource.</td>
<td>Delimited list of keywords or key phrases describing the resource.</td>
<td>Delimited list of keywords or key phrases describing the resource.</td>
<td>Based on DataCite Metadata 3.0</td>
</tr>
<tr>
<td></td>
<td>Keywords</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>ReCollect field</td>
<td>Subject</td>
<td>Subject or classification code describing the resource.</td>
<td>Should we recommend standard ones? Does DC and national catalogue have anything to say about this?</td>
<td>Should we recommend standard ones? Does DC and national catalogue have anything to say about this?</td>
<td>Based on DataCite Metadata 3.0</td>
</tr>
</tbody>
</table>

SN: 6044
Persistent identifier: 10.5255/UKDA-SN-6044-1
Pure as a data repository?

Tim Banks, University of Leeds
http://prezi.com/3xkzp8x8stfl/university-of-leeds/#
Search for datasets

Items where Author is "Cresswell, Alan"

Group by: Dataset Type | No Grouping

Jump to: Dataset

Number of items: 1.

Dataset

Validated radiometric mapping in 2012 of areas in Japan affected by the Fukushima-Daiichi nuclear accident


Archive
ValidatedRadiometricMappingData.zip
Download (27Mb)

Description

On March 11 2011 the north-eastern region of Japan was hit by a magnitude 9 earthquake, which occurred underneath the sea-bed 70 km east of the Oshika peninsula in Tohoku. The north-eastern shore of Honshu was hit by a tsunami resulting from this earthquake.
Title

The title of the item. The title should not end with a full stop, but may end with a question mark. There is no way to make italic text, please enter it normally. If you have a subtitle, it should be preceded with a colon [:]. Use capitals only for the first word and for proper nouns.
Example: A brief history of time
Example: Life: an unauthorised biography
Example: Mathematics for engineers and scientists. 5th edition

Description

Creators

<table>
<thead>
<tr>
<th>Family Name</th>
<th>Given Name / Initials</th>
<th>Email</th>
<th>GUID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Digital object identifier DOI

Datacite DOI

DOI from Datacite.org
If you do not know what a DOI is or what Datacite is, please refer to the Datacite DOI section of the About page.

Format to use is doi:xx.xxxx/xxxxxxx
Relate dataset to awards

![Image of a form for editing dataset information with tabs for Details, DOI, Subjects, Funding, and Deposit, along with a table for Funding Details with columns for Project Code, Award No, Project Name, Principal Investigator, and Funded.]

Anna Clements & Valerie McCutcheon  
CRIS2014, 12th May 2014
So what have we been doing?

Components of a data management service: Digital Curation Centre:
http://www.dcc.ac.uk/resources/how-guides
Challenges

• Data is much more heterogeneous than publications
  – How it is manifested
  – How it is created and reused
  – How embedded within the research process

• Not just picking up a pdf & standard bibliographic metadata at the end of the process
Academic-led review

“We need to avoid the idea of a data repository as “a place where data goes to die”.

“.. should be designed to act as a hub to encourage and facilitate cross-School interactions .... data science … ‘

“.. broadening accessibility means that datasets must also be suitably annotated so that they can be found easily .. .. from simple keyword associations (“tagging”), through to structured metadata such as CERIF, up to full linked open metadata.”

“Many subject areas already have multiple options for data storage, which need to be integrated rather than competed with.”
Conventional research life cycle

Where and how should institutions tap into this life cycle to improve research data management

David Shotton, University of Oxford, 2011, JISC ADMIRAL project
Enhanced research life cycle

Dissemination

- Open data on Web
  - Scholarly publications: conference papers and journal articles

Hypothesis formulation and project design

Research plan

Experimentation and data creation

Raw data in research notebooks and live PC files

Research results and conclusions

Data selection and interpretation

Publication activities

Institutional repositories – papers and datasets

CRIS

External repositories

Preservation

Conditioning

David Shotton, University of Oxford, 2011, JISC ADMIRAL project
Conclusion

Researchers should be in the driving seat & some already are.

We can provide tools, infrastructure & standards … & carrots (or sticks) along the way …

BUT

We can’t do the research!

“We need to avoid the idea of a data repository as “a place where data goes to die”. 
Tim Banks, University of Leeds
http://prezi.com/jlrw_nt5igdw/research-data-management-storage/
Research Data meets Research Information: Thank you for listening