A publisher point of view on open research data

Wouter Haak – VP Research Data Management

Nov, 2016

New roles in Open Science and Data Stewardship
As a researcher, when I wake up in the morning the first thing I think of is

A) “I just love to write more papers about my research”

OR

B) “I just thought of a new way to filter my data that might actually lead to a result that makes more sense”
So let’s talk a bit about data
WWARN, the first malaria data sharing network, has used pooled analysis of shared data to provide evidence to **help improve dosing regimens** of malaria treatments

- **260 institutions globally** have worked with WWARN, and over 120,000 individual patient records have been contributed
- Based on the results, the World Health Organization has revised the recommended dose of DP, a commonly used antimalarial for young children
What are we really after: social sciences

Capital in the Twenty-First Century is a 2013 book by French economist Thomas Piketty.

- It focuses on wealth and income inequality in Europe and the United States since the 18th century.
- Central thesis is that when the rate of return on capital ($r$) is greater than the rate of economic growth ($g$) over the long term, the result is concentration of wealth, and this unequal distribution of wealth causes social and economic instability.
- All raw data, normalized data, the analysis, and methods have all been made publicly available on a dedicated website https://www.quandl.com/data/PIKETTY.

Piketty's Capital: An Economist's Inequality Ideas Are All the Rage" by Megan McArdle, Bloomberg Businessweek, May 29, 2014
https://en.wikipedia.org/wiki/Capital_in_the_Twenty-First_Century
When we talk about data, we really talk about the following:

- Machine & environment settings
- Raw data
- Processed data
- Scripts, analyses, algorithms
- Protocols, methods, workflows

Full provenance needed

Note: images for illustrative purpose only
The 10 components of effective research data

1. Stored (existing in some form)
2. Preserved (long-term & format-independent)
3. Accessible
4. Discoverable (data is indexed or data is linked from article)
5. Citable
6. Comprehensible (description / method is available)
7. Trusted (e.g. reviewed)
8. Reproducible
9. Re-usable (allow tools to run on it; interoperable)
10. Integrate upstream and downstream – make metadata to serve use.

https://www.elsevier.com/connect/10-aspects-of-highly-effective-research-data
De Waard, Cousijn, Aalbersberg
Just another way of saying: “Data needs to be FAIR”

1. Stored (existing in some form)
2. Preserved
3. Accessible
4. Discoverable
5. Citable
6. Comprehensible
7. Trusted
8. Reproducible
9. Re-usable (allow tools to run on it)
10. Integrate upstream and downstream – make metadata to serve use.

Findable
Accessible
Interoperable
Reusable

https://www.force11.org/group/fairgroup/fairprinciples
Compliance increasingly important for research data

Some funder examples

<table>
<thead>
<tr>
<th>RCUK</th>
<th>EPSRC</th>
<th>Wellcome T</th>
<th>NSF</th>
<th>NIH</th>
<th>Horizon 2020</th>
<th>NWO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Based on analysis of policies; August 2016

https://docs.google.com/document/d/1ODK1D4gs5Kl59o-ADo-Ps5_zhzdOlc_E0ur8EEB1hY/
Opportunity to move beyond compliance!

1. Stored
2. Preserved
3. Accessible
4. Discoverable
5. Citable
6. Comprehensible
7. Trusted
8. Reproducible
9. Re-usable
10. Integrate upstream and downstream - make metadata to serve use.

RCUK | EPSRC | Wellcome T | NSF | NIH | Horizon 2020 | NWO
--- | --- | --- | --- | --- | --- | ---
Integrated

Opportunity
The 10 components for effective research data

1. Stored
2. Preserved
3. Accessible
4. Discoverable
5. Citable
6. Comprehensible
7. Trusted
8. Reproducible
9. Re-usable
10. Integrate upstream and downstream – make metadata to serve use.

Most researchers already work with several data tools

- Home built ELN
- Other ELN
- Data curation / repository
- Data Center
- Data management plan
What can Elsevier do?
Linking papers to data: phase 1
Elsevier Database linking program – example Pangaea

Calcium carbonate corrosiveness in the South Atlantic during the Last Glacial Maximum as inferred from changes in the preservation of *Globigerina bulloides*: A proxy to determine deep-water circulation patterns?
A.N.A. Vollers & R. Henrich
University of Bremen, Faculty of Geosciences, Department of Paleoceanography and Sedimentology, P.O.Box 330440, D-28334 Bremen, Germany

Abstract
The modern Atlantic Ocean, dominated by the interactions of North Atlantic Deep Water (NADW) and Antarctic Bottom Water (AABW), plays a key role in redistributing heat from the Southern to the Northern Hemisphere. During the Last Glacial Maximum, the

- Supplementary data at PANGAEA
- Bidirectional links between PANGAEA & ScienceDirect
- Data visualized next to the article

http://www.elsevier.com/databaselinking
Linking papers to data: phase 2

www.Scholix.org

- ICSU/WDS/RDA Publishing Data Service Working group
- Creating linked-data model for exposing DOI to DOI links outside publisher’s firewall
- Collaboration between CrossRef, DataCite, Europe PubMed Central, ANDS, Thompson Reuters, Elsevier, OpenAire

**Objective:** move from a plethora of (mostly) bilateral arrangements between the different players…
.. to ..

.. a one-for-all cross-referencing service for articles and data
Phase 3: we need a research dataset search engine!

- Many (broad) dataset search examples already available

- Some common themes:
  - search of metadata only (i.e. ranking based on metadata only)
  - And/or federated search (i.e. no ranking)
  - And/or focused on giving credit (citation) rather than on discoverability

- Uncommon (because difficult):
  - Deep indexing of datasets (so real ranking and filtering)
  - Search engine really focused on data discovery
Phase 3: Elsevier Data Search

E.g. search for “Temperature viscosity ionic liquids”

DataSearch.Elsevier.com

1. Across repositories
2. (Deep) indexing of data, so not just metadata
3. Data preview
Hivebench Electronic Lab Notebook

Lab notebook

Experiments

Schizophrenia risk gene

ZNF804A expression

eexpression of tcf4

expression of znf804a

fetal brain expression

Protocols and results

schizophrenia risk gene mir137

helena created on April 15th 2016

- Wash in PBS with 0.1%Tween20, 3 X, and 10 min of each time
- Incubate with secondary antibody (peroxidase-conjugated goat anti-mouse IgG, etc.) following the manufacturer’s instruction, usually 1:200 to 1:2000; 1 hour at room temperature.
- Wash in PBS with 0.1%Tween20, 3 X, and 10 min of each time

5. Developing: (Use of chromomeric substrate - 3,3’-diaminobenzidine, DAB)
- Transfer the membrane to a shallow tray.
- Add 10ul H2O2 (30%) to 10 ml of 0.05%DAB in PBS, mix well immediately;
- Pour the DAB to the membrane, incubate at room temperature with gentle shaking in the dark if possible;
- Monitor the progress of the reaction carefully. When the bands are of the desired intensity (2-5 min), wash the filter briefly in water, and in PBS.
- Dry the membrane and photograph it to provide a permanent record of the experiment.

www.hivebench.com
Manage, Store: Mendeley Data repository

Reproducible experiments on dynamic resource allocation in cloud data centers

Description of this data
In Wolke et al., we compare the efficiency of different resource allocation strategies experimentally. We focused on dynamic environments where virtual machines need to be allocated and deallocated to servers over time. In this companion paper, we describe the simulation framework and how to run simulations to replicate experiments or run new experiments within the framework.

Experiment data files
- Results.zip
- CSV files with simulation and experimentation results.
- github.paper.IS2015-master.zip
- github.workload-master.zip
- dockerfile
- IS2015.tar.gz
- repzip.zip
- ReproZip package of the simulation executed in the Docker container.

Linked to published papers – or not

Versioning and provenance

Researcher in control: how/when to share

https://data.mendeley.com/datasets/xz6gv65m6d/6
Publish data and software: e.g. SoftwareX and DataInBrief have top 25% most read OA articles on ScienceDirect
The 10 components for effective research data

1. Stored
2. Preserved
3. Accessible
4. Discoverable
5. Citable
6. Comprehensible
7. Trusted
8. Reproducible
9. Re-usable
10. Integrate upstream and downstream - make metadata to serve use.

Summary of initiatives discussed

Hivebench (Electronic lab notebook)
Mendeley data repository

Integrated with: Data journals / Research Elements
Integrated with: Data Linking
Integrated with: Data Search
The 10 components for effective research data

1. Stored
2. Preserved
3. Accessible
4. Discoverable
5. Citable
6. Comprehensible
7. Trusted
8. Reproducible
9. Re-usable
10. Integrate upstream and downstream – make metadata to serve use.

Moving towards an integrated ecosystem to help the researcher and the institution

Measure data uptake on a structural basis, followed by a pro-active researcher adoption program
Grazie mille!

Questions?