Current practices in data sharing and challenges ahead!

Timon Oefelein
Senior Manager Account Development, Strategic Partnerships and Outreach

Tromsø, November 2018
The importance of verification

“The strongest arguments provide nothing so long as conclusions are not verified by experience.” *Opus Tertium*, c. 1267

Roger Bacon, 1214-1291
The problem!

Scale and cost of issue

- Pharma reports 75%+ failure rates
  - leave no room for error
  - Irreproducible biology research costs US $28b per year

Community perspective

- 52% said “Yes, a significant crises”
- 50% couldn’t reproduce own work
- 70% couldn’t reproduce work of others

---

5. Ioannidis et al (2009) [https://www.nature.com/ng/journal/v41/n2/full/ng.295.html](https://www.nature.com/ng/journal/v41/n2/full/ng.295.html)
One of the main causes: missing data!

John Ioannidis:
“The main reason for failure to reproduce was data unavailability, and ... incomplete data annotation or specification of data processing and analysis.”

Additional benefits of sharing data =

**x2** the publication output of a study

**Study**: 7000 studies in social & behavioral science, funded by NSF and NIH:

**Papers with data**:

- Repository/archived = **10** (median) publications*
- **Not shared** = **4** (median) publications*¹

* Resulting publications by PI or a member of Research Team. Excludes secondary publications by non-team members.

**Meta-analysis**: Articles with data available are cited 9-50% more, depending on the field:

1. Pienta et al (2010) [https://deepblue.lib.umich.edu/handle/2027.42/78307](https://deepblue.lib.umich.edu/handle/2027.42/78307)
2. Piwowar & Vision (2013) [https://doi.org/10.7717/peerj.175](https://doi.org/10.7717/peerj.175)
5. Sears et al (2011) [https://figshare.com/articles/Data_Sharing_Effect_on_Article_Citation_Rate_in_Paleoceanography/1222998/1](https://figshare.com/articles/Data_Sharing_Effect_on_Article_Citation_Rate_in_Paleoceanography/1222998/1)
Current data sharing practices amongst researchers
Survey on data sharing practices *(at point of publication)*

- Focus: data sharing at “point of publication”
- Defined sharing: long term (repository or ESM)
- **We covered following aspects:**
  - Methodology: survey based
  - Participants: 7700 responses, global reach
  - White Paper

The importance that data are discoverable in different subject areas (1 is the least important, n=7626)
Depositing of data in subject areas (n=7664) (at point of publication..)

DS: State of open data in 2018 *(before/after publication)*

---

**Fig 1. How often researchers have made their data openly available**

<table>
<thead>
<tr>
<th>Year</th>
<th>Frequently &amp; Sometimes</th>
<th>Never &amp; Rarely</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>57%</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td>60%</td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td>64%</td>
<td></td>
</tr>
</tbody>
</table>

---

“Familiar with FAIR”

“Heard but not familiar”

“I’ve never heard of the FAIR principles before now”
Data sharing behaviour by size of dataset (n=6513)
What problems do authors have in sharing datasets?

- Lack of time to deposit data: 26%
- Organising data in a presentable and useful way: 46%
- Not knowing which repository to use: 33%
- Unsure about copyright and licensing: 37%
- Costs of sharing data: 19%
- Other: 6%
What can publishers do to help researchers overcome these problems?
How are publishers responding to these challenges?

**Springer Nature:**

- Research Data Help Desk
- Recommended repositories
- Data policies
- Helping set standards:
- Implement citation and linking practices
- Research Data Service
- **New: Data Availability Reporting**
- Credit for data sharing via badges
Springer Nature research data policy initiative

- **Type 1**
  - Encourages:
    - Data sharing
    - Data citation

- **Type 2**
  - Encourages:
    - Data sharing
    - Data citation
    - Data-Availability-Statement

- **Type 3**
  - Assumes:
    - Data-sharing
    - Data Avail-Statement
    - Data citation
  - **Mandatory:**
    - Data sharing (reader + reviewer)
    - Data-Avail-Statement
    - Data citation

- **Type 4**
  - **Mandatory:**
    - Data sharing (reader + reviewer)
    - Data-Avail-Statement
    - Data citation

- SRG journals: **over 1,400 (>50%)** adopted a policy
- Nature journals: most are Type 3
- Preference for **repositories over Electronic Supplementary Material**
Proportion of articles with data sharing statement

Helpdesk and Data Curation Service

Researchers submit their data files securely

The Research Data team checks the data and curates metadata

The data are published and linked to the author’s paper
Before data curation: a researcher’s dataset in a desktop folder

The dataset is stored as an Excel file in a desktop folder.

*No one other than the creator can access the data, or even knows that it exists.*

The file title is not comprehensible to anyone but the creator.

No description or keywords available.

No one other than the creator can access the data, or even knows that it exists.
Before curation begins

Once received, we check to make sure that the dataset is suitable for our curation services. Multiple files in any format are accepted.

Pre-curation data checks:

- The data aren’t sensitive
- The data don’t include direct or indirect human identifiers
- The data shouldn’t be in a community repository
- The data are associated with a trusted publication

After making these checks, we begin the curation process. If necessary we may recommend that the dataset is split into smaller groups or collections.
Example of output of Research Data Support

Paper published in Nature

Data published in Figshare

Data availability statement included with the paper
Open Badges Pilot – BMC Microbiology

Before badge: 7% of articles reported open data
After 6 months: 23% articles ""
After 12 months: 39% articles ""

Psychological Science

Badge Criteria (with Center for Open Science):

- Data availability statement
- Dataset in a public repository
- Persistent identifier (DOI, Accession number)
- Dataset checked/confirmed as relevant to paper
Some funders require sharing of research data

393 Funders surveyed (157 from Europe):

- 46 (22%) funders “mandate” data sharing
- 31 (8%) encourage data sharing
- **303 (77%)** have “no mandate”
Additional key developments:

1. Researchers citing data in References/DAS, via DataCite standard
2. Interconnectivity of repositories, e.g. DANS, dataMED

- Data search engines, e.g. Google Dataset Search
- Metadata standards, e.g. DataCite, Schema.org, Scholix
- Best practices for data archiving/sharing, e.g. FAIR
- Standardization of data policies, e.g. RDA
- Growing community of advocates, e.g. FORCE11
- Indexes of data repositories, e.g. re3data.org
- Data Citation Index e.g. Clarivate’s DCI, Scholix Framework
- Growth of institutional repositories

- EU New data infrastructure pilots..
  - European Open Science Cloud (€30m)
  - OpenAIRE-Advance (€10m)
Thank You!

Nature, Robust research: Institutions must do their part for reproducibility, 1 Sept 2015. Illustration by David Parkins
Researchers at Pennsylvania State University have been considering how quickly a glacial ice melt in Antarctica would raise sea levels. By updating models with new discoveries and comparing them with past sea-level rise events they predict that a melting Antarctica could raise oceans by more than 3 feet by the end of the century if greenhouse gas emissions continued unabated, roughly doubling previous total sea-level rise estimates. Rising seas could put many of the world’s coastlines underwater or at risk of flooding and storm surges.
Two success stories!

**Human Genome Project**

Cost of sequencing human genome fallen from $1 Billion in 1990s to $3-500 today.

Nr. of diseases with an identified genomic cause risen from 61 in 1990 to 5000 today.

More than 100 drugs currently on the market have a pharmacogenomics label which facilitates personal medicines.

Data is openly accessible

---

**European Bioinformatics Institute**

£47 million: Annual operating cost

£1 billion: Annual efficiency savings to researchers worldwide\(^2\)

£920 million: Estimated annual estimate of future research impacts

Data is openly accessible

---

Towards transparency in research reporting

2013: Advocate/Encourage:

- Raise standards of reporting on **statistics**:
- Ensure reporting on: **sample size estimation, randomization, blinding, handling of data**
- Information on **materials** used
- Eliminate length restrictions for **Methods** sections
- Encourage linked **step-by-step protocols** on Protocol Exchange

In 2017: Formal reporting:

- 1. Reporting Summary and 2. Editorial Policy Checklist
- The Reporting Summary is published with the manuscript
- Authors fill them out prior to peer review and update them on resubmission
- Available to referees during peer review
Example author feedback report

Springer Nature Research Data Support feedback

Job number: RDS-XXX-XXXX

To improve the discoverability and accessibility of your data, we have made the enhancements your data record in springernature.figshare.com:

- Copy edited the title
- Checked/corrected author name and order
- Copy edited/enhanced your description
- Added keywords and categories to improve discoverability
- Added links to your publication
- Added funding information
- Unpackaged your data files and uploaded as individually-previewable file

Following deposition and curation, your data now features the following benefits:

Your data:

- have been assigned a unique identifier by the host repository https://doi.org/10.6084/m9.figshare.XXXXX
- can be previewed in the repository using a web browser
- will be clearly and appropriately linked to the associated publication
- have been assigned a licence that will permit wide reuse
- are easily citable via the figshare repository

Areas that could still be improved:

- Your dataset appears to be incomplete based on the description and other data files; it appears that ... are missing. Please check that all relevant data files are present.
- Your data record would benefit from a more detailed description of the underlying methods used to generate these data

Specific guidance from our Research Data Editors:

In the description are you referring to file XXX.xls when you refer to ‘data file 1’, or to file XXY.xls? I have edited the description to include a more explicit reference to this.
Capture-mark-recapture data modelling survival rates of Microcebus murinus in relation to glucocorticoid level, parasite infection and body condition

This dataset consists of an Excel spreadsheet containing capture-mark-recapture data, which were used to model survival of Microcebus murinus in different contexts.

These were:
- A Multistate modelling approach to model semi-annual survival relative to hair cortisol concentration (HCC) and scaled mass index (SMI). Median or third quartile were used as categorization cut-off.
- A Cormack-Jolly-Seber (CJS) modelling approach to model survival over the productive season relative to hair cortisol concentration, scaled mass index, and pattern of parasitism which was measured as the parasite species richness (number of distinct parasite morphotypes found per individual), the multiple infection (presence of more than one parasite morphotype), and the overall parasite prevalence (presence of at least one parasite morphotype)

The data used to assess the link between semi-annual survival and HCC includes results of capture sessions held in October 2012, 2013, 2014, April 2013, and March 2014, during which a total of 371 individuals (74 females, 97 males) were captured. The time dataset, excluding the October 2014 session, was used to assess the effect of HCC on semi-annual survival probabilities, for a total of 149 individuals (63 females, 86 males). The dataset used for the CJS models includes data collected during monthly trapping sessions between September 2012 and April 2013, for a total of 48 individuals (18 females, 32 males).

All research activities conducted in Madagascar received official approval from the Ministère de l'Environnement, de l'Ecologie, and de la Mor et des Forêts, and comply with national animal care legislation of Madagascar.

FUNDING
Deutscher Akademischer Austauschdienst (A/12/00420) and Deutsche Forschungsgemeinschaft (KR 3634/4-1)

RESEARCH DATA SUPPORT
Research data support provided by Springer Nature.

Source: https://doi.org/10.6084/m9.figshare.5259415
Number of funders citing this issue, n= 13

- Inadequate metadata/ lack of metadata
- ID of data to deposit/ preserve
- Lack of incentives
- Disciplinary variation
- Concern over misuse of data by 3rd parties
- Compliance with Federal regulations
- Data protection/ treatment of sensitive data
- Identification of suitable repositories
- Meeting the costs of RDM
- Cultural resistance to sharing

Springer Nature research 2017; n = 13