Presentations

Research Data Management

11.20 Break

11.50 Presentations: Research Data Management

Session Chair: Dr Fiona Murphy

Embedding Good RDM Practice in RDM Workflows
Effective research data management is one of the biggest challenges facing researchers in sharing their research data
Rebecca Grant
- Research Data Manager at Springer Nature

Towards a Librarian-led RDM Academy
Understanding the role librarians can play in training researchers about research data management
Jean P Shipman
- Vice President, Global Library Relations at Elsevier

1.00 Lunch
Sponsors

Supporting R2R

WILEY
Presentations
Research Data Management

11.20 Break

11.50 Presentations: Research Data Management

Session Chair: Dr Fiona Murphy

Embedding Good RDM Practice in RDM Workflows
Effective research data management is one of the biggest challenges facing researchers in sharing their research data
Rebecca Grant
• Research Data Manager at Springer Nature

Towards a Librarian-led RDM Academy
Understanding the role librarians can play in training researchers about research data management
Jean P Shipman
• Vice President, Global Library Relations at Elsevier

1.00 Lunch
Embedding good RDM practice in publishing workflows

Rebecca Grant,
Research Data Manager
February 2019
Springer Nature is a leading research, educational and professional publisher, providing quality content to our communities through a range of innovative platforms, products and services.


As the leading open access publisher, we see the rise of open research in all its manifestations as one of the major forces reshaping the way that researchers communicate and collaborate to advance the pace and quality of discovery.

Our focus is on investing in and creating tools, services or training that help the research community to understand and utilise new ideas and concepts.

Embedding good RDM practice in publishing workflows
Research data policies: implementation and impacts

Embedding good RDM practice in publishing workflows
Why data sharing?

Evidence is mounting on costs & scale of the issue

- Irreproducible biology research costs US $28 billion per year\(^1\)
- Pharma companies report 75%+ failure rates replicating conclusions of peer-reviewed papers \(^2,3\)

A recent *Nature* survey\(^4\) highlights concern in the research community

- >50% of researchers couldn’t reproduce their own experiments
- >70% couldn’t reproduce the work of others

There is evidence that data availability increases reproducibility

A study\(^5\) of eighteen *Nature Genetics* papers found:

- Two could be reproduced fully
- Six were reproduced partially
- Ten could not be reproduced

“The main reason for failure to reproduce was data unavailability, and discrepancies were mostly due to incomplete data annotation or specification of data processing and analysis.” — *Nature Genetics* **41**, 149–155 (2009)

---

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)
What are our authors asked to do?

- Springer Nature launched a data policy standardisation initiative in 2016.\(^1\)
- More than 1,500 (~60%) Springer Nature journals have adopted a standard research data policy.
- Approach is practical and pragmatic, enabling all journals to adopt a policy even if they are new to data sharing.
- All policies support community specific policies, mandates and repositories.
- All policies and journals promote data citation in Information for authors.
- Similar initiatives since introduced at Elsevier\(^2\), Wiley\(^3\), Taylor & Francis\(^4\).

---

1. *Standardising and harmonising research data policy in scholarly publishing*
   Iain Hrynashkiewicz, Aliaksandr Birukou, Mathias Astell, Sowmya Swaminathan, Amye Kenall, Varsha Khodiyar
   International Journal of Digital Curation; doi: https://doi.org/10.2218/ijdc.v12i1.531
3. https://authorservices.wiley.com/author-resources/Journal-Authors/licensing-open-access/open-access/data-sharing.html
Embedding good RDM practice in publishing workflows

Policy Types

<table>
<thead>
<tr>
<th>Type 1</th>
<th>Type 2</th>
<th>Type 3</th>
<th>Type 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data sharing and data citation is encouraged but not required</td>
<td>Data sharing and data citation is encouraged and evidence of data sharing required</td>
<td>Data sharing encouraged and statements of data availability required</td>
<td>Data sharing, evidence of data sharing and peer review of data required</td>
</tr>
</tbody>
</table>

Process

1. Identify and agree the most relevant policy type
2. Implement standardized test and processes into the workflow
3. Provide a consistent and easy-to-follow journal data management plan

All policy types:
- Preference sharing of data via repositories (rather than ESM/SI)
- Allow citation of public datasets in reference lists/bibliographies
- Encourage use of publisher helpdesk to ensure compliance with funder mandates
Data policy implementations December 2018

- Type 1: 608
- Type 2: 528
- Type 3: 405
- Type 4: 6

n = 1547

Embedding good RDM practice in publishing workflows
The data publishing team began a short researcher project working with Nature journals, aiming to:

Assess how researchers choose to share their data

Measure the time required by editors and production staff to add data availability statements to manuscripts (which gives an indication of cost to the publisher).
What is a data availability statement (DAS)?

Required by many journals/publishers e.g. PLOS, Royal Society, BMJ, Hindawi and funders including RCUK

“The datasets generated during and/or analysed during the current study are available in the [NAME] repository, [PERSISTENT WEB LINK TO DATASETS].”

“The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.”

“All data generated or analysed during this study are included in this published article (and its supplementary information files).”
Results of the analysis

An average of 10 minutes’ editorial time per paper
An average of 5 minutes additional copy edit time

![Bar chart showing median and average minutes for Nature Neuroscience, Nature Physics, Nature Cell Biology, and Nature Geoscience.]
Coding the Data Availability Statements

**Type 1** stated that the data is available from the author on request.

**Type 2** stated that the data had been included in the manuscript or its supplementary material.

**Type 3** stated that some or all of the data is publicly available, for example in a repository.

**Type 4** stated that figure source data was included with the manuscript. This is a method of data sharing used by some authors in a subset of Nature journals that publish life sciences research. Some journals encourage authors to provide the source data behind their figures/plots as spreadsheets.

*Data availability*

The data that support the plots within this paper and other findings of this study are available from the corresponding author upon reasonable request.
Results of the analysis

Average time by statement type

- Statement Type 1: 5.9
- Statement Type 2: 14.5
- Statement Type 3: 18.2
- Statement Type 4: 12.6
Findings

Making Data Availability Statements mandatory increases manuscript processing time.

But the benefits of including them outweighs the issue of increased processing times.

Processing time is impacted most when data are shared in repositories.

But these types of availability statements have the most positive impact for authors and other researchers, e.g. through increased accessibility and increased citations.

Blogpost on this analysis: https://tinyurl.com/y5saccff
Researcher challenges and data sharing support

Embedding good RDM practice in publishing workflows
Practical Challenges for Researchers in Data Sharing white paper

A global survey of nearly 8000 researchers

Stuart, David; Baynes, Grace; Hrynaszkiwicz, Iain; Allin, Katie; Penny, Dan; Lucraft, Mithu; Astell, Mathias (2018): Whitepaper: Practical challenges for researchers in data sharing

https://doi.org/10.6084/m9.figshare.5975011.v1

Embedding good KUM practice in publishing workflows
Private sharing of data is more common than public sharing of data

62% have shared data both privately and publicly
36% have only shared data privately
2% of respondents have only shared their data publicly

Global levels of data sharing:
- Poland – 76% (highest)
- Germany – 75%
- UK – 58%
- USA – 55%

The three most common methods of private sharing were:
- Email (65%)
- USB or flash drives (41%)
- File sharing services (39%)

The three most common ways of public sharing were: (n=569)
- Supplementary information to journal articles (51%)
- Lab or personal website (27%)
- Subject specific repository or data archive (25%)
76% of respondents highly rate the importance of their data being discoverable: most popular ranking was 10/10.
Embedding good RDM practice in publishing workflows

*The main challenge to data sharing is organising data in a presentable and useful way.*
<table>
<thead>
<tr>
<th><strong>Research Data Helpdesk</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Queries are answered within two business days</td>
</tr>
<tr>
<td>Run by members of the Springer Nature Data Publishing team</td>
</tr>
<tr>
<td>Expertise in data curation and management, archiving and digital preservation, copyright and licensing, Open Access publishing</td>
</tr>
<tr>
<td>Always encourage best practices, e.g. the use of community repositories for specific data types</td>
</tr>
<tr>
<td>Email: <a href="mailto:researchdata@springernature.com">researchdata@springernature.com</a></td>
</tr>
</tbody>
</table>

http://www.springernature.com/gp/group/data-policy/helpdesk
2018 queries to our Helpdesk

- Data types: 1
- Copyright and licensing: 2
- Policy implementation: 3
- Data citation: 3
- Other: 5
- Data availability and access: 24
- Research Data Support: 33
- Policy compliance: 33
- Repositories and depositing data: 59

n = 163
Springer Nature Research Data Support

To help authors and journals follow good practice in sharing and archiving of research data, we provide optional data deposition and curation services.

Researchers submit their data files securely

The Research Data team curates the data and metadata

The data are published and linked to the author’s paper

More information is available on our website here:
Before data curation: a researcher’s dataset in a desktop folder

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.
After Research Data Support

Working with the researcher’s manuscript or published paper, we draft a comprehensive metadata record for the dataset which is sent to the researcher for approval before being published. Embargoes can be applied if necessary.

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 reproductive 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 rates and HCC includes the results of capture sessions held in October 2012, 2013, 2014, April 2013, and March 2014, during which a total of 171 individuals (74 females, 97 males) were captured. The same dataset, excluding the October 2014 session, was used to assess the effect of SMI on semi-annual survival probability, for a total of 140 individuals (63 females, 77 males).

The curated dataset will be published with its own metadata record which includes rich descriptive information, reuse conditions, licence, DOI, metrics and keywords (this example is https://doi.org/10.6084/m9.figshare.5259415)
Example author feedback report

SPRINGER NATURE

Springer Nature Research Data Support feedback

Job number: RDS-XXXX-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 record:
- includes a useful, descriptive title
- includes keywords, categories and other metadata that will aid discovery
- is indexed and searchable online
- has a unique, persistent link
- contains descriptive details that can be easily viewed and retrieved within the repository
- describes the data files in sufficient detail to make them easily understood

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.
Open data badges pilot: BMC Microbiology

Does the application of badges to published papers which share their data openly:

• Affect levels of data sharing by authors?
• Affect reader engagement with articles and their supporting data, and perceptions of article quality?

And can we assess the benefits and costs of consistently assessing and awarding open data badges for a Springer Nature journal?
Open data badges pilot: BMC Microbiology

Cees van Nieuwkoop, Willize E. Reijden, Albert M. Vollaard, Nathalie Leyten, Ted Koster, Hans C. Ablij, Martina T. van der Beek, Mirjam J. Knoi and Jaap T. van Dissel

Read a blogpost on the project here: https://tinyurl.com/y8atpy8a
Summing up

Embedding good RDM practice in publishing workflows
Embedding good data management practice in publishing workflows

- Properly managing and sharing research data is a means to conduct and publish more reliable and reproducible research – and for researchers to gain more recognition for doing so.

- Large surveys show researchers value data sharing and reuse but can have practical problems that prevent them sharing data.

- Support and encouragement for data sharing can be provided by publishers in a number of ways, e.g. advice (helpdesk) hands-on assistance (data curation) or rewards (data badges).

- Policy enforcement requires intervention (and therefore time) on behalf of the publisher.

- Provision of advisory or practical support requires data and disciplinary expertise.
Thank you

Rebecca Grant, Research Data Manager

Researchdata@springernature.com / Rebecca.Grant@springernature.com

https://go.nature.com/ResearchDataServices

https://researchdata.springernature.com/

The story behind the image

John Maynard Keynes (1883–1946)

John Maynard Keynes was a British economist who revolutionised the theory and practice of macroeconomics, reformed economics and had a profound influence on economic policy. This illustration represents the Keynesian model which shows that in a monetary economy it is possible to have periods of high unemployment unless governments use active monetary and fiscal policy to stimulate aggregate demand.