Sharing data alongside publications with Taylor & Francis

Presented by:
Caroline Sutton, Head of Open Scholarship Development

Scholarly Summit, Washington DC, Nov 8, 2017
What is Research Data?

"Research data is defined as recorded factual material commonly retained by and accepted in the scientific community as necessary to validate research findings; although the majority of such data is created in digital format, all research data is included irrespective of the format in which it is created."

The Engineering and Physical Research Council

- Research data varies by discipline and subject area.
- Research data may consist of primary data, secondary data, raw data or manipulated data/sub-set of data.
Considerations for Taylor & Francis

Readiness:

- Some subject areas and regions are more “ready” to share data than others; e.g. within Earth Sciences there is a push for data policies, in Asia our teams are being asked about data polices, while in Engineering, for instance, little discussion about this.

- Some sub-disciplines within a subject area are more ready than others: e.g. within Political Science quantitative scientists ready to share while qualitative methodologists are not.
Considerations for Taylor & Francis

Data variability

- Lots of variation in degree of structure in data and type of data; from genomics to literary analysis.

- Some researchers enter a lab and generate data via an experiment while others engage in re-interpretations or analyses of existing texts, artifacts, and similar.

- The minimal data set necessary to interpret or replicate findings can vary.
Image Credit: “DSC_4598w” by IAEA Image Bank, Flickr.com, distributed under a CCBY-SA 2.0 generic license.

Image Credit: “D 20 Hz 100 fs lasers room 121” by UCL Mathematical and Physical Sciences, Flickr.com, distributed under a CCBY 2.0 generic license.

Image Credit: “Riverside Park WRF Laboratory” by Eric Shea, Flickr.com, distributed under a CCBY 2.0 generic license.
Considerations for Taylor & Francis

Flexible:

• Must allow one to dip their toes into the waters of data sharing rather than taking a deep dive, yet we need to allow those who are ready to jump into the deep end to do so with our full support.
The Data Spectrum: Banking sector

Closed
- Internal access
  - Employment contract + policies
  - Sales reports

Shared
- Named access
  - Explicitly assigned by contract
  - Personal transaction history
- Group-based access
  - Via authentication
  - Regulatory information

Open
- Public access
  - Licence that limits use
  - Postcode lending data
- Anyone
  - Open licence
  - Bank products

The Data Spectrum helps you understand the language of data.
...highly site specific, potentially limiting their wider value. However, applying the approach as conducted in this paper to data such as that presented by Barnett et al (2013) to derive relative values for different organisms should provide a more generic set of 'reference data'. In taking the REML approach forward it will be beneficial to target...

References

Barnett et al., 2013 C.L. Barnett, N.A. Beresford, L.A. Walker, M. Baxter, C. Wells, D. Copplestone
Element and radionuclide concentrations in representative species of the ICRP's reference animals and plants and associated soils from a forest in North-west England.
NERC - Environmental Information Data Centre (2013) http://doi.org/10.5285/e40b53d4-6699-4557-bd55-10d196e9e9ea

Data in Reference Lists

**Numbered style:**


**Harvard style:**


**Vancouver style:**

A Data Citation Roadmap for Scientific Publishers

Helena Cousijn, Amye Kenall, Emma Ganley, Melissa Harrison, David Kerrohan, Fiona Murphy, Patrick Polschuk, Maryann Martone, Timothy Clark

doi: https://doi.org/10.1101/100784

This article is a preprint and has not been peer-reviewed [what does this mean?]

Abstract

This article presents a practical roadmap for scholarly publishers to implement data citation in accordance with the Joint Declaration of Data Citation Principles (JDDCP), a synopsis and harmonization of the recommendations of major science policy bodies. It was developed by the Publishers Early Adopters Expert Group as part of the Data Citation Implementation Pilot (DCIP) project, an initiative of FORCE11.org and the NIH BioCADDIE program. The structure of the roadmap presented here follows the 'life of a paper' workflow and includes the categories Pre-submission, Submission, Production, and Publication. The roadmap is intended to be publisher-agnostic so that all publishers can use this as a starting point when implementing JDDCP-compliant data citation.

Copyright  The copyright holder for this preprint is the author/funder. It is made
Data Availability Statements

Provides information on where and under what conditions the data directly supporting the publication can be accessed.

Examples:
The data that support the findings of this study are openly available in [repository name e.g “figshare”] at [http://doi.org/[doi]], reference number [reference number].

The data that support the findings of this study are available in [repository name] at [URL/DOI], reference number [reference number]. These data were derived from the following resources available in the public domain: [list resources and URLs]

The data that support the findings of this study are available [from] [third party]. Restrictions apply to the availability of these data, which were used under licence for this study. Data are available [from the authors / at URL] with the permission of [third party].

Data sharing is not applicable to this article as no new data were created or analysed in this study.
<table>
<thead>
<tr>
<th>Level of data sharing</th>
<th>Basic Data Sharing Policy</th>
<th>Share upon Reasonable Request Data Policy</th>
<th>Publicly Available Data Sharing Policy</th>
<th>Open Data Policy</th>
<th>Open and Fully FAIR Data Sharing Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors are encouraged to share or make open the data supporting the results or analyses presented in their paper where this does not violate the protection of human subjects or other valid privacy or security concerns.</td>
<td>Authors publishing with the Journal agree to make their data available upon reasonable request. It is up to the author to determine whether a request is reasonable.</td>
<td>Authors make their data freely available to the public, but under a license that limits re-use, or under unclear re-use conditions.</td>
<td>Authors must make their data freely available to the public, under a license allowing re-use by any third party for any lawful purpose. Data shall be findable and fully accessible.</td>
<td>Authors must make their data freely available to the public, under a license allowing re-use by any third party for any lawful purpose. Additionally, data shall meet with FAIR standards as established in the relevant subject area.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Availability Statement</th>
<th>Highly encouraged</th>
<th>Mandatory</th>
<th>Mandatory</th>
<th>Mandatory</th>
<th>Mandatory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistent Identifier for Data</td>
<td>Highly encouraged</td>
<td>Highly encouraged</td>
<td>Highly encouraged</td>
<td>Mandatory</td>
<td>Mandatory</td>
</tr>
<tr>
<td>License applied to dataset</td>
<td>Author’s choice</td>
<td>Author’s choice</td>
<td>Author’s choice</td>
<td>CCO, CCBY or equivalent</td>
<td>CCBY, CC0 or equivalent</td>
</tr>
<tr>
<td>Data Citation</td>
<td>Highly encouraged</td>
<td>Highly encouraged</td>
<td>Highly encouraged</td>
<td>Mandatory</td>
<td>Mandatory</td>
</tr>
</tbody>
</table>
Visualizing the default policy

Author decides to submit article to a T&F title with default policy

Is there a dataset associated with this paper? Yes/No

Has the data been deposited in a repository? Yes/No

Provide the URL, reserved URL or other permanent identifier when asked to do so.

Include a Data Availability Statement with Submission

Check that the dataset has been cited in the paper and in the reference list.

Templates available on Taylor & Francis Online Author Services

Answer "no" when asked this question during submission and complete submission as normal.

Pause submission and submit data to appropriate repository

Information on policies, recommended repositories, and other relevant information available on Taylor & Francis Online Author Services

Submit paper
Where can researchers deposit their data?

- Generalist repositories
  - Figshare
  - Dryad
  - Zenodo
  - Dataverse
  - Github (code)

- Subject specific repositories
  - National repositories
  - Institutional repositories
  - Niche repositories
FAIRsharing is here! From our first incarnation, BioSharing.org, which focussed on the life sciences, we are growing into FAIRsharing.org, to serve users across all disciplines and support Findable, Accessible, Interoperable and Reusable (FAIR) data.

A curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies.

Find

👍 Recommendations
Standards and/or databases recommended by journal or funder data policies.

Discover

😊 Collections
Standards and/or databases grouped by domain, species or organization.

Learn

🎓 Educational
About standards, their use in databases and policies, and how we can help you.
re3data.org Reaches a Milestone and Begins Offering Badges

re3data.org has reached a milestone of identifying and listing 1,500 research data repositories, making it the largest and most comprehensive registry of data repositories available on the web. It has grown steadily since its launch four years ago to cover a wide range of disciplines from around the world.

Read more

Enhancements to creating and updating re3data

We are happy to announce a new feature that enables users to more easily suggest corrections and enhancements of information about research data repositories registered in re3data.org.

Read more

New re3data.org Schema and Search Functionality

We are pleased to announce the publication of version 3.0 of the "Metadata Schema for the Description of Research Data Repositories" (Rieckmann et al., 2015). Rieckmann, J., Vinkant, P., Ulrich, P., Klütsch, G., Schremp, E., Fichtmüller, O., ... Kleschnaft, R. (2013). Metadata schema for the description of research data repositories....

Read more
How can you start preparing?

- Begin familiarizing yourself with the data sharing discussions in your subject area and/or region.
- Identify relevant repositories for data deposition.
- Identify which general repositories are most likely to be used in your subject area or region.
Further Questions to Explore

- Are there standard practices in your subject area regarding data management and availability?
- Are there funders within your subject area that are mandating or encouraging specific practices with regard to data management and availability?
- Is data sharing a concern for authors, funders and/or institutions in this area?
- Why would we want to introduce a specific policy for this title?
- What are the likely drivers and barriers with respect to authors sharing data?
Further Questions to Explore

- What data should be shared? Raw vs. sub-set underlying findings described in article?
- What exceptions should be made to the policy?
- What repercussions will there be for authors who do not comply?
- Where an open data policy is adopted, this will have consequences for a double-blind peer review process
Transparency, open sharing, and reproducibility are core values of science, but not always part of daily practice. Journals, funders, and scholarly societies can increase reproducibility of research by adopting the Transparency and Openness Promotion (TOP) Guidelines and helping them evolve to meet the needs of researchers and publishers while pursuing the most transparent practices.

**Over 5,000* journals and organizations have become signatories to the TOP Guidelines.**

Journal signatories are expressing their support of the principles of openness, transparency, and reproducibility, expressing interest in the guidelines and commit to conducting a review within a year of the standards and levels of adoption.

Organization signatories are expressing their support of the principles of openness, transparency, and reproducibility and encouraging associated journals to conduct a review of the standards and levels of adoption.

See Springer Nature's recent statement about the TOP Guidelines.

*Includes the journals included in Elsevier's recent statement about the TOP Guidelines.
News

Call for Nominations for CODATA Officers (President and Vice-President(s)) and Executive Committee Members: Deadline 9 April 2018
26 October 2017
Help CODATA deliver on its strategy to mobilise the data revolution for research!

Read more

Northeast Big Data Hub Announces CRUX: The Collaborative Resource and Understanding Exchange Program
24 October 2017
The Northeast Big Data Innovation Hub (NEDIH) has recently announced the CRUX program: a...

Read more

RDA Events

RDA Meets Estonian Researchers, 8 November 2017, Tartu, Estonia
08 Nov 2017 - 09:00 to 13:00
Date and time: Wednesday 8.11.2017, 09:00 – 13:00 Audience: Research communities, policy makers....

Read more

Managing Digital Research Objects in an Expanding Science Ecosystem
15 Nov 2017 - 15:00 to 23:00

Read more

Request for comments

❓ Recommendations for Implementing a Virtual Layer for Management of the Complete Life Cycle of Scientific Data
By Tobias Weigel

❓ CODATA/RDA Research Data Science Schools for Low and Middle Income Countries - Charter Statement
By Lynn Yarmey on 17 October 2017
Thank you!

@CarolineSutton
Caroline.Sutton@informa.com
Policies and Services to Make Sense of Data

Taylor & Francis Scholarly Summit

Patricia Cruse, Executive Director, DataCite

Institute of Peace
Washington DC
How many journals make data sharing a requirement of publication?

<table>
<thead>
<tr>
<th>Results of Journal Survey</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total no. of Journals surveyed</td>
<td>371</td>
</tr>
<tr>
<td>Total no. of Journals with data sharing policies</td>
<td>162</td>
</tr>
<tr>
<td>Total no. of Journals that make sharing a requirement of publication</td>
<td>31</td>
</tr>
<tr>
<td>Total no. of Journals that enforce the policies</td>
<td>27</td>
</tr>
<tr>
<td>Total no. of Journals that state consequences for non compliance</td>
<td>7</td>
</tr>
</tbody>
</table>

https://jordproject.wordpress.com/2013/07/05/going-back-to-basics-reusing-data/

JoRD (Journal Research Data Policy Bank) sheds light the policies devised by academic publishers to promote linkage between journal articles and underlying research… (Jisc funded, ended in 2014)
Back in the day...

- Figure 1. Two pages (scan) from Galilei's Sidereus Nuncius (“The Starry Messenger” or “The Herald of the Stars”), Venice, 1610.

**data** = drawings of Jupiter and its moons

**metadata** = timing of each observation

**text** = descriptions of methods, analysis, conclusions

---


http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1003542
Summing up the problem

We need reliable and unambiguous access to data!

- attribution
- collaboration and reuse
- reproducibility
- faster (and efficient) progress
- feed future researchers
But wait, there’s more

- Get credit and attribution (the village)

- Comply with publishers’ data sharing policies

- Meet funder mandates

- Meet institutional requirements

- Respond to community norms and practices
And still more…

- Researchers & Contributors
- Data and Software
- Grants and Projects
- Publications
- Funders
- Institutions
DataCite’s approach

Provide technical infrastructure:
- Create DOIs for research data
- Build and adopt services that promote data sharing
- Integrate with other community services

Provide community infrastructure:
- Advocate & communicate about the importance of data sharing
Our Mission

NOT-FOR-PROFIT GLOBAL INITIATIVE

MEMBER ORGANIZATION

COMMUNITY-DRIVEN

OVER 1300 DATA CENTERS

OVER 9 MILLION DOIS

Not-for-profit global initiative – Member organization – Community driven – over 1300 data centers – over 9 million DOIs
The centrality of a DOI

DOI = Digital Object Identifier

an alphanumeric string created to:

- uniquely *identify/name* digital content
- serve as a *stable, persistent link* to that content’s location on the web

**Syntax**

```
+----------------+-----------------+
| DOI Name       | Name            |
| Prefix         | URL             |
| /              | Metadata        |
+----------------+-----------------+
Example: 10 . 1234 / data567
```
Down in the weeds

1. Take a dataset

2. Describe it
   - Title
   - Authors
   - Year
   - Description
   - And others...

3. Assign a DOI
   - 10.1234/exampledata

4. Reuse and reference!
   
   ATLAS Collaboration, “Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC: $H \rightarrow \gamma\gamma$,” http://doi.org/10.7484/INSPIREHEP.DATA.A78C.HK44

5. Enjoy the benefits
   - Findability
   - Track citations
   - Reusability
   - Measure impact

✓ Unique  ✓ Persistent
1. **Importance**: legitimate, citable products of research

2. **Credit and Attribution**: scholarly credit and normative and legal attribution to all contributors to the data

3. **Evidence**: In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited

4. **Unique Identification**: include a persistent method for identification

5. **Access**: data citations should facilitate access to the data themselves

6. **Persistence**: identifiers, and metadata should persist

7. **Specificity and Verifiability**: data citations should facilitate identification of, access to, and verification of data

8. **Interoperability and Flexibility**: data citation methods should be sufficiently flexible

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [https://www.force11.org/group/joint-declaration-data-citation-principles-final].
DataCite’s Service integration = building bridges

- make it easier
- integrate with related content
- provide impact
- give credit
Putting the pieces together

- Researchers & Contributors
- Data and Software
- Grants and Projects
- Publications
- Funders
- Institutions
Linking Data with Data

Why it matters

- provide a complete picture of the data environment
- multiple versions of the same dataset
- subsets of larger datasets or heterogeneous collections
- dynamic data
- software
- workflows
Linking data to data

Dataset

Biomineralization control related to population density under ocean acidification
Stefano Goffredo, Fiorella Prada, Erik Caroselli, B Capaccioni, Francesco Zaccanti, Luca Pasquini, Paola Fantazzini, Simona Fermani, Michela Reggi, Oren Levy, Katharina Elisabeth Fabricius, Zvy Dubinsky & Giuseppe Falini
Dataset published 2014 via PANGAEA - Data Publisher for Earth & Environmental Science

References, supplements

1 Related Work

Raw Data - “Biomineralization control related to population density under ocean acidification”
Stefano Goffredo, Fiorella Prada, Erik Caroselli, Bruno Capaccioni, Francesco Zaccanti, Luca Pasquini, Paola Fantazzini, Simona Fermani, Michela Reggi, Oren Levy, Katharina Fabricius, Zvy Dubinsky & Giuseppe Falini
Fileset published 2014 via Figshare

Linking Data with Researchers & Contributors

Why it matters
- Credit and attribution
- Answers who, what, when, where
- Link one or more contributors to research output
Seamless integration with ORCID

**Researchers:** (1) use ORCID iD when submitting dataset (2) authorize DataCite to update your ORCID record.

**Data centers:** (1) collect ORCID identifiers during submission (2) embed iD in the work and include the iD when submitting to DataCite.

**DataCite:** Upon receipt of data from a data center with a valid identifier, DataCite automatically pushes information to the researcher’s ORCID record.
Linking Data with Articles

Why it matters

- Increase visibility and discovery of research data and articles
- Place research data in the right context to enable reuse
- Support credit attribution

Challenges:

- Data underlying findings are not always fully available
- Data underlying findings described in a are made available, but hidden in supplementary information
- Data underlying the findings are available, but not properly linked to/from article
Linking Data with Articles: Follow FAIR Data Principles

Force11: Data Sharing Principles

FAIR data

- **Findable**
  - Describe your data in a data repository
  - Apply a persistent identifier

- **Accessible**
  - Consider what will be shared
  - Obtain participant consent & perform risk management

- **Interoperable**
  - Use open formats
  - Consistent vocabulary
  - Common metadata standards

- **Reusable**
  - Consider permitted use
  - Apply appropriate licence

http://slideshare.net/lshtm/preparing-data-for-sharing-the-fair-principles
Example 1: One article links to five datasets

Journal article

Temperature-Induced Syntheses, Iodine Elimination, Enantiomers Resolution, and Single-Crystal-to-Single-Crystal Transformation of Imidazole-Co(II) Coordination Polymers with Amino-isophthalic Acid as Co-Ligand

Journal article published June 9, 2016

Related data

CCDC 1414538: Experimental Crystal Structure Determination
Hui-Fang Zhou, Tian He, Ke-Fen Yue, Yong-Liang Liu, Chun-Sheng Zhou, Ni Yan & Yao-Yu Wang
Work published 2016 via Cambridge Crystallographic Data Centre

Is supplement to http://doi.org/10.1021/ACS.CGD.6B00527

CCDC 1059747: Experimental Crystal Structure Determination
Hui-Fang Zhou, Tian He, Ke-Fen Yue, Yong-Liang Liu, Chun-Sheng Zhou, Ni Yan & Yao-Yu Wang
Work published 2016 via Cambridge Crystallographic Data Centre

Is supplement to http://doi.org/10.1021/ACS.CGD.6B00527

CCDC 1484151: Experimental Crystal Structure Determination
Hui-Fang Zhou, Tian He, Ke-Fen Yue, Yong-Liang Liu, Chun-Sheng Zhou, Ni Yan & Yao-Yu Wang
Work published 2016 via Cambridge Crystallographic Data Centre

Is supplement to http://doi.org/10.1021/ACS.CGD.6B00527
Example 2: Software described in Journal of Open Source Software

Armadillo: a template-based C++ library for linear algebra
Conrad Sanderson & Ryan Curtin
Journal article published June 10, 2016 via JOSS

Armadillo C++ Linear Algebra Library
Conrad Sanderson
Work published 2016 via Zenodo

Sources
- DataCite (Crossref)

Relation Types
- Is cited by

Publisher
The Open Journal

Share on
- Twitter
- Facebook
Example 3: PLOS articles linked with at least one DataCite DOI

<table>
<thead>
<tr>
<th>Work Title</th>
<th>Authors</th>
<th>Journal publication details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seasonality and Locality Affect the Diversity of Anopheles gambiae and Anopheles coluzzii Midgut Microbiota from Ghana</td>
<td>Jeweina Akorli, Mathilde Gendrin, Nana Adjoe P. Pels, Dorothy Yeboah-Manu, George K. Christophides &amp; Michael D. Wilson</td>
<td>Journal article published June 20, 2016 via PLOS ONE</td>
</tr>
<tr>
<td>Genetic Diversification and Dispersal of Taro (Colocasia esculenta (L.) Schott)</td>
<td>H. Chaei, R. E. Tracey, M. F. Duval, R. Rivallan, A. Mukherjee, L. M. Aboagye … &amp; V. Lebot</td>
<td>Journal article published June 17, 2016 via PLOS ONE</td>
</tr>
</tbody>
</table>

- **Publisher:** Public Library of Science (PLoS)
- **No. of works:** 542
- **Related content:**
  - Seasonality and Locality Affect the Diversity of Anopheles gambiae and Anopheles coluzzii Midgut Microbiota from Ghana
  - Genetic Diversification and Dispersal of Taro (Colocasia esculenta (L.) Schott)
  - Mosquito Saliva Increases Endothelial Permeability in the Skin, Immune Cell Migration, and Dengue Pathogenesis during Antibody-Dependent Enhancement
Linking data to funders and organizations

Why it matters

• The research environment is complicated (everyone wants credit)
  • Bring together scholarly output with all of the stakeholders
    • Researchers
    • Funders
    • Organizations
    • Grants
    • Projects
Linking to organization identifiers

a documented need for a comprehensive, open, and accessible organization identifier infrastructure

Content identifiers: DataCite, Crossref

Contributor identifiers: ORCID

Organization Identifiers?
Organization Identifier Working Group

Summary

The Organization Identifier (OrgID) Working Group was established in January 2017 to refine the structure, principles, and technology specifications for an open, independent, non-profit organization identifier registry to facilitate the disambiguation of researcher affiliations. The scope of work includes three separate but interdependent areas: Governance, Registry Product Definition, and Business Model & Funding. The goal of the Working Group is to create an implementation plan by the end of 2017.
Data, a first-class research output

LEARN MORE!
Journal articles v/s data

Primacy of journal articles
- currency of research
- sophisticated methods to gauge impact (citations, page views, downloads)
- understand relative impact and identify relationships

What about data?
- 1st class scholarly object
- lagging infrastructure
- broader role in research process
- has its own use & reuse profile
- Community of best practices are lacking
For example - usage tracking: downloads

How do we count downloads? Sum, average, maximum, whole package?

A path forward

Make Data Count: data usage stats and data citations

1. provide formal recommendation for measuring data usage

2. further develop *Data Level Metrics* (DLM) Hub and services

3. expose exemplars to drive adoption

4. engage in community building
Thank you!

patricia.cruse@datacite.org
https://www.datacite.org
Twitter: @datacite
NIH Data Stewardship Policy

November 8, 2017

Dina N. Paltoo, Ph.D., M.P.H.
Director, Division of Scientific Data Sharing Policy
Office of Science Policy, Office of the Director, NIH
NIH’s View of the Benefits of Sharing Data

• **Preserves scientific record**
  — Sharing encourages better data management
  — Not all results are published

• **Facilitates research integrity**
  — Validation of experiments/results
  — Ethical obligation to human subjects
  — Transparency for greater trust

• **Advances science and application**
  — “Standing on the shoulders of giants”
  — Accelerate translation of results into practice
  — Suggest new hypotheses
  — Innovation of statistical methods, resources, and tools

• **Increases efficiency**
  — Large volumes of data generated more than enough for one team to analyze

• **Fosters rigor and reproducibility**
  — Enables data generated from one study to be used to explore additional research questions
  — Less money spent on duplicating (and revalidating) existing data, more budget dollars for funding advanced research

• **Facilitates portfolio planning**

• **Achieves synergies when combining data**
  — Increases statistical power and value
Data Management and Sharing: Growing Interest

- Science increasingly digital, generating vast amounts of data
- Non-research data also being digitized and available for research uses
  - E.g., research on data collected in clinical care and stored in EHRs
- U.S. Government initiatives
  - White House initiatives to increase access to publications and the results of federally-funded scientific research, and foster open data
  - NIH Big Data to Knowledge (BD2K) Initiative
  - Common Rule revisions to support consent to maximize utility of biospecimens and data
  - Regulations and NIH Policy for clinical trial registration and results reporting
  - 21st Century Cures Act - NIH authority to require data sharing
- OECD Science and Technology Ministerial Declaration
- G7 Science and Technology Ministers Communiqué
- G20 Leaders’ Communiqué
- Advances in Information Technology and bioinformatics
  - Easier to collect, organize, discover, access, and analyze data to advance science, stimulate innovation, improve health, environment, national security, and other public missions
- Changing scientific ethos and practice
  - More open and transparent
- International interest in open science
MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren
       Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

1. Policy Principles

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

Scientific research supported by the Federal Government catalyzes innovative breakthroughs that drive our economy. The results of that research become the grist for new insights and are assets for progress in areas such as health, energy, the environment, agriculture, and national security.

Access to digital data sets resulting from federally funded research allows companies to focus resources and efforts on understanding and exploiting discoveries. For example, open weather
Interagency Working Group on Open Science

- Co-Chaired by NIH and National Science Foundation
- Exchange information on implementation of public access policies and practices
- Facilitate interagency coordination and cooperation on open science
- Recommend additional objectives for Federal open science policies
- Outline effective strategies for improving preservation, discoverability, and accessibility of scientific data
- Identify effective approaches for data preservation & access; assess requirements for scaling up; and identify gaps
- Facilitate coordination of training, education, and workforce development
- Liaise with other National Science and Technology Council groups
- Identify opportunities for international communication and collaboration
Examples of Other Stakeholders Engaged in Open Science

- Publishers
- Association of American Universities/Association of Public & Land-Grant Universities (Public Access Working Group)
- Association of Research Libraries
- Private Foundations
  - Bill and Melinda Gates Foundation
  - Howard Hughes Medical Institute
  - Wellcome Trust
  - Arnold Foundation
- National Academy of Sciences
  - Board on Research Data and Information
- Patient-Centered Outcomes Research Institute
- Research Data Alliance
- CODATA (Committee on Data of the International Council for Science)
NIH’s Culture of Data Sharing

- NIH Data Sharing Policy
- Model Organism Policy
- NIH Public Access Policy (Publications)
- Genome-wide Association (GWAS) Policy
- Big Data to Knowledge (BD2K) Initiative
- Genomic Data Sharing (GDS) Policy
- Modernization of NIH Clinical Trials
- White House Initiative (2013 “Holdren Memo”)
- NIH Intramural Human Data Sharing Policy
- HHS Rule and NIH Policy on Clinical Trial Results Dissemination
- All of Us Program
- Cancer Moonshot
- NIH Data Commons Pilot

Culture Change and Challenges

- **Time and effort**
  - Determine which data to preserve (not necessarily all data)
  - Clean data, put in accessible format (consistency; standardized elements)
  - Provide metadata
  - Limited training in data management and sharing

- **Requires infrastructure**
  - Repositories for long-term archiving
  - Procedures for providing data access

- **Lack of rewards/incentives**
  - Citations/publications used for academic credit
  - How to cite/credit data collection and sharing

- **Considerations for ethical, legal, and social implications, human participant protections, privacy and trust**

- **Proprietary interests**
  - Researchers want to analyze & publish first
  - Institutions/Individuals want to protect competitive advantage

- **Human resources**
Increasing Access to Publications and Digital Scientific Data

February 2015: “NIH Plan” released

- Publications: NIH Public Access Policy
- Digital Scientific Data: Plan for Public Access to Digital Scientific Data

Plan ≠ Policy; NIH to establish priorities for data sharing
Establishing NIH Priorities for Data Stewardship

RFI on Strategies for NIH Data Management, Sharing, and Citation (NOT-OD-17-015)

- **Section I: Data Sharing Strategy Development**
  - What, when, and how data should be managed and shared
  - Value in sharing different types of data
  - Barriers and how to overcome them

- **Section 2: Inclusion of Data and Software Citation in NIH Research Performance Progress Reports (RPPR) and Grant Applications**
  - Impact of citations on reporting and the need for technical guidance
  - Strengthen and incentivize data and software sharing

- **General feedback on relevant topics**

- **Released November 14, 2016, comment period closed on January 19, 2017**
95 submissions received from both national and international stakeholders

http://osp.od.nih.gov/sites/default/files/resources/Public_Comments_Data_Management_Sharing_Citation.pdf
Summary of Public Comments

• **Section I: Data Sharing Strategy Development**
  – Data underlying or supporting a publication, dissertation, or supplemental materials, pre-registration/preliminary data, metadata and any data needed to replicate a study, would be the most valuable to share
  – Data should be made available for secondary research purposes for a minimum of 10 years
  – Establishing a culture of sharing that would incentivize and encourage data sharing

• **Section 2: Inclusion of Data and Software Citation in NIH Research Performance Progress Reports (RPPR) and Grant Applications**
  – Increased reporting of data and software through citations (e.g., via use of global unique persistent identifier), as long as it was conducted as a means to incentivize researchers to share data, thus enabling them to get appropriate credit or attribution for their work
  – Promoting the importance of versioning datasets and software when generating citations

• **General feedback on relevant topics**
  – NIH should discourage the use of proprietary software for uploaded/shared data
  – NIH should consider the inclusion of individuals with appropriate expertise in data management and sharing (e.g., bioinformaticians) in the peer review process
Considerations for NIH Policy Development

- Definitions, Scope and Applicability, and existing NIH data sharing policies
- Requirements
  - Data Management and Sharing Plans
    - Require in funding applications
    - Evaluate during peer review
    - Structured format, updateable, publicly available (e.g., RePORTER)
    - Contingency plans
    - Data sharing exceptions
  - Data Sharing (e.g., 21st Century Cures Act)
- Use of existing repositories and citation and research products
- Request budget in funding application
- Mechanisms for compliance and enforcement
- Implementation
- Next Steps: release draft Policy for public comment, outreach, and communication
October 13, 2017

Purpose: Multi-stakeholder group to address the issue of identifying prospectively those data that would be valuable if shared in order to inform decisions and priorities regarding data management and sharing policies, as well as investments in infrastructure supporting data sharing.

- Identified metrics, data sources, methods, and analytical approaches that could be used to assess the value of particular data for sharing.

- Overall workshop themes that contribute to the value of data sharing:
  - Outcomes and products resulting from sharing
  - Data reuse, culture, and scientific discipline impact on sharing
  - Costs, economic factors, and impact of sharing
  - Return on investment of sharing different types of data

Next Steps: Development of a white paper and determine any NIH-NSF funding opportunities.
Reporting Preprints and Other Interim Research Products (NOT-OD-17-050)

• Effective: March 25, 2017
  — Encourages investigators to use interim research products, such as preprints, to speed the dissemination and enhance the rigor of their work
  — Allows citing of interim research products in applications, proposals, and reports
  — To ensure the integrity and impact of interim research products, authors and repositories must:
    • Declare competing interests, track versioning
    • Ensure interim products are findable through DOIs, open metadata
    • Adopt a number of license and technical processes (e.g. Creative Common Attribution licenses (CC-BY), application programming interfaces, archival plans, etc.), and to use repositories that support these practices

• These products may be cited in progress reports
  — Research Performance Progress Report (RPPR), section C – Products
Effective Data Sharing...

- Relies upon appropriate identification, adoption, and crediting of good data management and sharing practices (stewardship) – consistent with the FAIR (Findable, Accessible, Interoperable, Reusable) principles

- Relies upon compliance
  - Culture change and incentives
  - Consistent/complementary policies among various stakeholders, such as funders and publishers/journals
Additional Resources

• For General Inquiries:
  SciencePolicy@od.nih.gov (OSP)
  Dina.Paltoo@nih.gov (OSP)

• Subscribe to the OSP LISTSERV
  Send and email to: LISTSERV@list.nih.gov
  with the message: Subscribe OSP_News

Learn more about the Office of Science Policy from our blog “Under the Poliscope”
http://osp.od.nih.gov/under-the-poliscope