

Rewarding data sharing through data citation: from policy to practice

Iain Hrynaszkiewicz

JOSS, Tokyo 18 June 2018

ADVANCING
DISCOVERY



What is a data citation?

- A reference made to data in the same way as researchers routinely provide a bibliographic reference to journal articles and books
- When public datasets have Digital Object Identifiers (DOIs), or equivalent identifiers, **it is the same as citing a journal article – and this should be our simple message to researchers (authors and editors)**
- Include the minimum information recommended by DataCite and follow journal style e.g. for *Nature*: authors, title, publisher (repository name), identifier, year

Creator (author)

Title

77. Di Stefano, B., Collombet, S. & Graf, T. Time-resolved gene expression profiling during reprogramming of C/EBP α -pulsed B cells into iPS cells. *figshare* https://dx.doi.org/10.6084/m9.figshare.939408_D1 (2014).

[+ Show context](#)

Publisher (repository name)

Identifier (DOI)

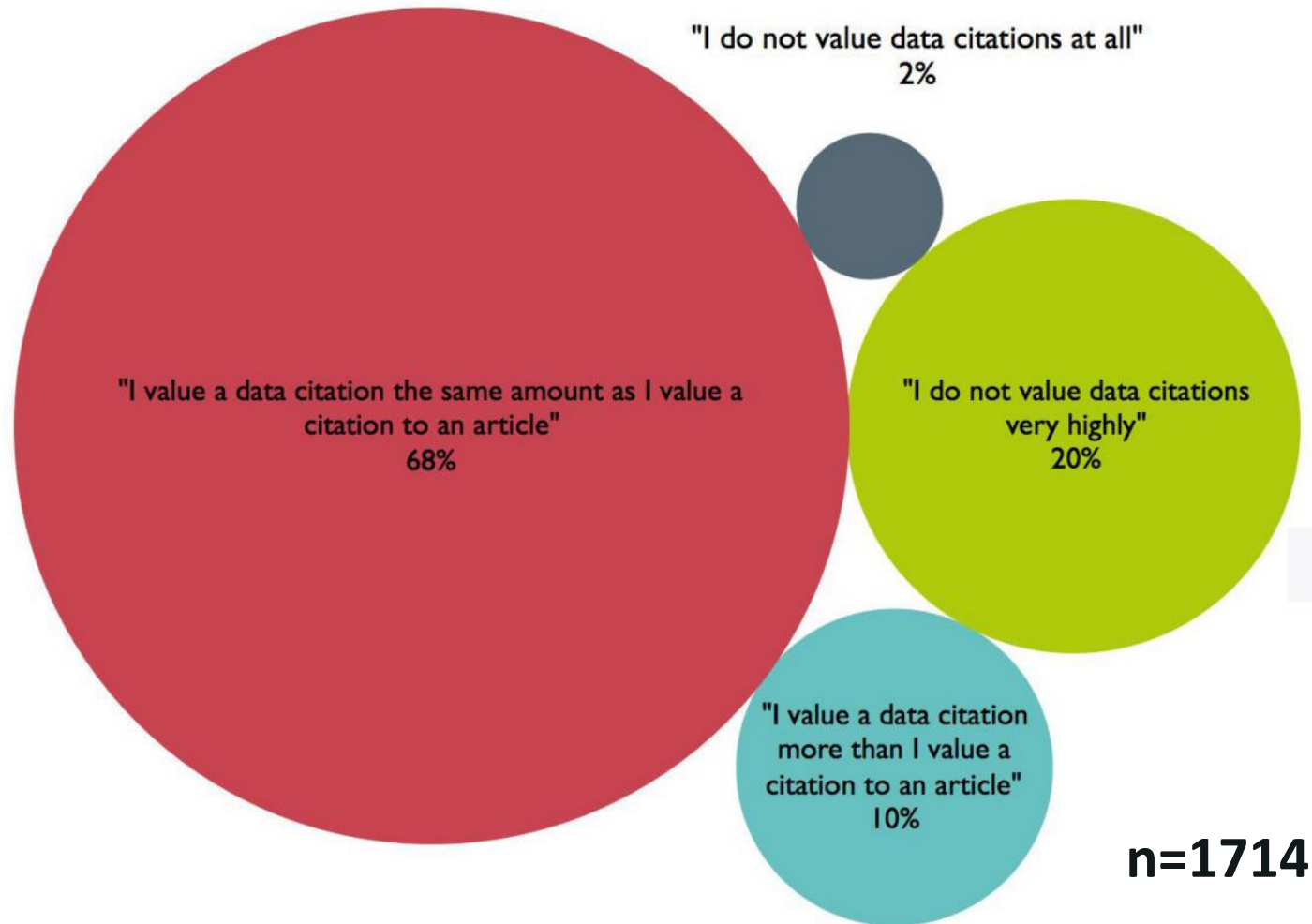
Publication year

SPRINGER NATURE

Researchers want to be credited for data reuse

	Agree strongly	Agree somewhat
<i>It is important that my data are cited when used by other researchers (n=1291)</i>	885 (69%)	293 (23%)

Researchers also value credit received from data citations



Funding agencies value research data and data citation

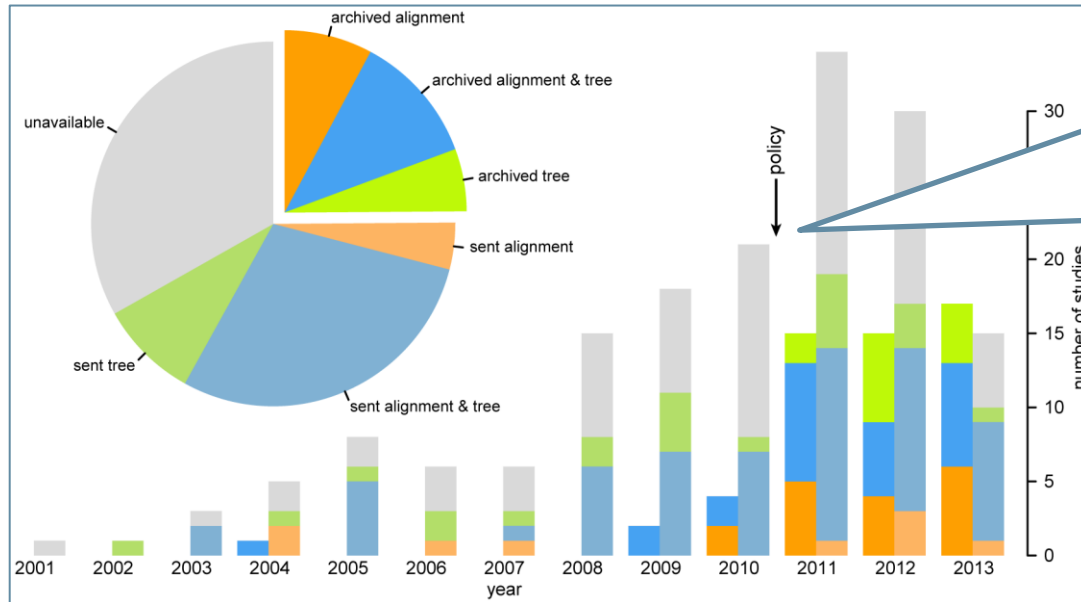
UK concordat on open research data, endorsed by funding agencies including Wellcome Trust and the UK research councils: ***“The obligation to recognise through citation and acknowledgement the original creators of the data must be respected”***

<https://www.ukri.org/files/legacy/documents/concordatonopenresearchdata-pdf>

National Science Foundation (US) *“For all new grant applications from 14 January [2013], the US National Science Foundation (NSF) asks a principal investigator to list his or her research “products” rather than “publications” in the biographical sketch section. This means that, according to the NSF, a scientist’s worth is not dependent solely on publications. Data sets, software and other non-traditional research products will count too.”*

<http://www.nature.com/nature/journal/v493/n7431/full/493159a.html>

Journal and publisher policies can increase data sharing



Magee *et al* (2014) show increase in sharing and deposition of data from evolution research after 2010, after this research community and its journals adopted stronger research data policies

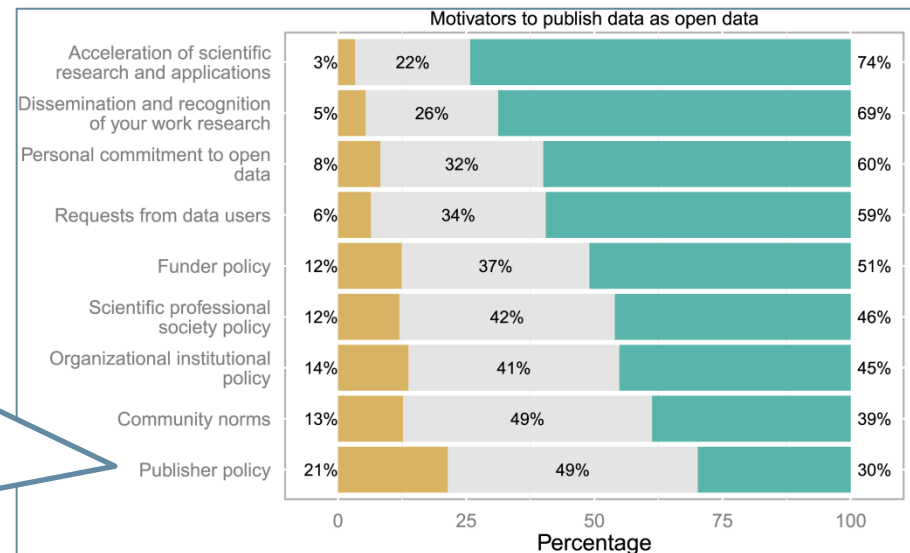
Citation and image credit (CC BY)

<https://doi.org/10.1371/journal.pone.0110268>

Survey data from Schmidt *et al* (2016) (n=843) identifies publisher policy as a motivator to publish data as open data

Citation and image credit (CC BY)

<https://doi.org/10.1371/journal.pone.0146695>



Journal data policies are becoming more consistent

Policy Types



- Springer Nature launched a data policy standardisation initiative in 2016¹
- More than 1,400 (~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 Elsevier², Wiley³, Taylor & Francis⁴

1. *Standardising and harmonising research data policy in scholarly publishing*

Iain Hrynaszkiewicz, Aliaksandr Birukou, Mathias Astell, Sowmya Swaminathan, Amye Kenall, Varsha Khodiyar
International Journal of Digital Curation; doi: <https://doi.org/10.2218/ijdc.v12i1.531>

2. <https://www.elsevier.com/authors/author-services/research-data/data-guidelines>

3. <https://authorservices.wiley.com/author-resources/Journal-Authors/licensing-open-access/open-access/data-sharing.html>

4. <https://authorservices.taylorandfrancis.com/understanding-our-data-sharing-policies/>


There is motivation to standardise data policies across publishers and other stakeholders

Co-chairs:



Natasha Simons (ANDS), Simon Goudie (Wiley), Azhar Hussain (Jisc), Iain Hrynaszkiewicz (Springer Nature)

- **Primary objective is to define common frameworks for research data policy – starting with journals and publishers and latterly funding agencies**



[ABOUT RDA](#)
[GET INVOLVED](#)
[GROUPS](#)
[RECOMMENDATIONS & OUTPUTS](#)
[RDA FOR D](#)

Data policy standardisation and implementation

Home » Working And Interest Groups » I

IG


Group details

Status: Recognised & Endorsed

Chair (s): Iain Hrynaszkiewicz, Natasha Simons, Simon Goudie, Azhar Hussain

Secretariat Liaison: Kathy Fontaine

TAB Liaison: Devika Madalli

 IG Established

[History](#)

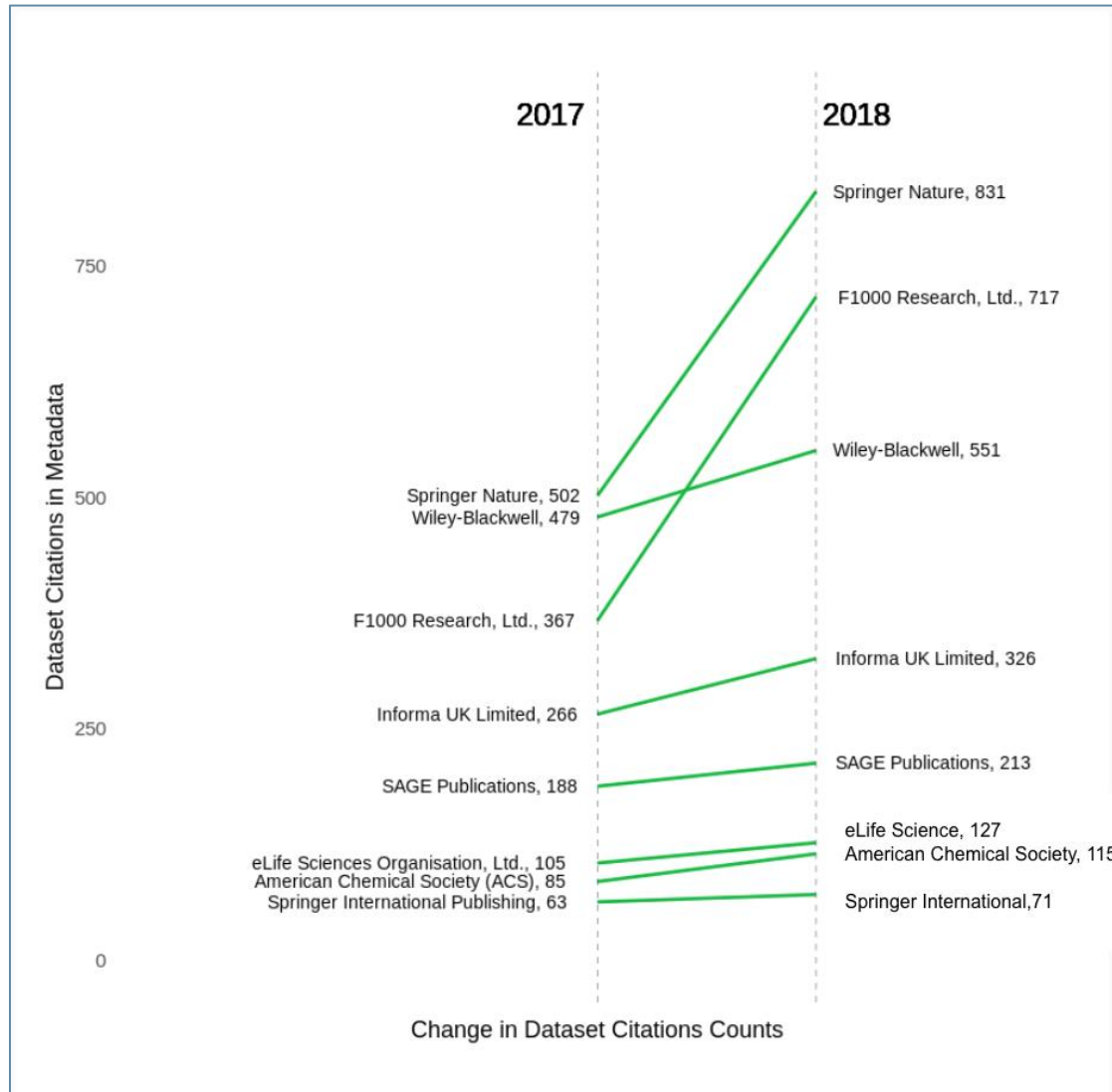
Introduction

Increasing the availability of research data for reuse is in part being driven by research data policies and the number of funders and journals and institutions with some form of research data policy is growing. The research data policy landscape of funders, institutions and publishers is however too complex (Ref: <http://insights.uksg.org/articles/10.1629/uksg.284/>) and the implementation and implications of policies for researchers can be unclear. While around half of researchers share data, their primary motivations are often to carry out and publish good research, and to receive renewed funding, rather than making data available. Data policies that support publication of research need to be practical and seen in this context to be effective beyond specialist data communities and publications.

Use cases and user scenarios

<https://www.rd-alliance.org/groups/data-policy-standardisation-and-implementation>

But are journal and publisher policies increasing data citation?



Yes, but progress has been slow

Scientific Data – at the cutting edge of data citation?

nature.com > scientific data

a natureresearch journal

MENU

SCIENTIFIC DATA

Search E-alert Submit Login

Data Descriptor | 20 June 2017 | OPEN

High-resolution bioclimatic variables for ecological modelling

Editorial | 13 June 2017 | OPEN

Open for business

Data Descriptor | 20 June 2017 | OPEN

Single-cell transcriptome of early embryos and cultured embryonic stem cells of cynomolgus monkeys

Tomonori Nakamura, Yukihiko Yabuta [...] Mitinori Saitou

Announcement

Author's corner: Providing incentives and ensuring quality in citizen science

Steffen Fritz, Linda See & Ian McCallum share tips on running an effective citizen science campaign.

scientificdata updates

Search Scientific Data

All Subjects

Find out more about Scientific Data

Read our Aims & Scope

SCIENTIFIC DATA

Discover data associated with our content

isaexplorer

a data discovery tool

Find the right repository for your data

Access our Recommended Repositories list

SCIENTIFIC DATA

Part of the Nature Research Group, *Scientific Data* is an open-access journal for descriptions of scientifically valuable datasets.

- Broad scope covering physical, life and quantitative social sciences
- In-house metadata curation
- Data-focused peer-review process
- Supports community data repositories
- Integrated submission of data to general repositories
- Data deposition and citation are mandatory for submission

What can we learn from its first 4 years of publishing?

The Data Descriptor article

Peer-reviewed articles enabling authors to provide comprehensive methodological detail about their data. Does not contain tests of new scientific hypotheses.

Sections:

- Title
- Abstract
- Background & Summary
- **Methods**
- **Technical Validation**
- **Data Records** ←
- **Usage Notes**
- Figures & Tables
- References
- **Data Citations** ←

Data Records

All the samples used in this study are summarized in Table 1. Consistent identifiers are used in Tables 2 and 3 to allow mapping between the proteomic and transcriptomic data outputs.

Data Record 1

The raw data, peaklists (.mgf), ProteomeDiscoverer result files (.msf) and ProteomeDiscoverer workflow files (.xml) have been uploaded to ProteomeXchange (<http://www.proteomexchange.org/>) with the following accession number PXD000134 (ref. 67; Table 2).

Data Record 2

Microarray data are available at the NCBI Gene Expression Omnibus (GEO) database under the accession numbers GSE26451 (ref. 68) and GSE26453 (ref. 69; Table 3).

Data Record 3

The peptide and protein identification data sets have been annotated by The Global Proteome Machine at <http://gpmdb.thegpm.org/>

Data Record 4

The peptide and protein identification data sets have been annotated by the StemCellOmicsRepository (SCOR) at <http://scor.chem.wisc.edu/>

Data Citations

67. Low, T. Y. *et al.* ProteomeXchange: PXD000134 (2013).

68. Chin, A. *et al.* Gene Expression Omnibus: GSE26451 (2011).

69. Chin, A. *et al.* Gene Expression Omnibus: GSE26453 (2011).

In-article data citation

SCIENTIFIC DATA | DATA DESCRIPTOR OPEN



Plant traits, productivity, biomass and soil properties from forest sites in the Pacific Northwest, 1999–2014

The dataset (*NACP TERRA-PNW: Forest Plant Traits, NPP, Biomass, and Soil Properties, 1999–2014*) is hosted with other contributions from the North American Carbon Program (NACP) by the Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics
 (Data Citation 1) Oak Ridge National Laboratory Distributed Active Archive Center

2016



PDF



ISA tab



Data Citations

[Abstract](#) • [Background & Summary](#) • [Methods](#) • [Data Records](#) • [Technical Validation](#) • [Additional Information](#) • [References](#) • [Data Citations](#) • [Acknowledgements](#) • [Author information](#)

1. Law, B. E., & Berner, L. T. *Oak Ridge National Laboratory Distributed Active Archive Center*
<http://dx.doi.org/10.3334/ORNLDAAC/1292> (2015).

Abstract

[Abstract](#) • [Background & Summary](#)
 • [References](#) • [Data Citations](#)

Plant trait measurements are needed for evaluating ecological responses to environmental conditions and for ecosystem process model development, parameterization, and testing. We

How does *Scientific Data* do it?

Manuscript Information | **Files** | Validate | Submit

Upload Files | Replace Files | File Type | File Description | File Order | Dataset Deposition | Dataset Upload

Dataset Deposition * Have you already uploaded ALL of your primary data files to a publicly accessible repository? *

Yes No

Please enter all details for datasets uploaded to publicly accessible repositories in the fields below

Repository Name *	Title *	Accession Number *	Url *	Password
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

Additional Data Files

Were the data collected and reported according to any specific community reporting standards, such as those described in the [EQUATOR Network](#) for health-related research, the [ARRIVE guidelines](#) for animal research, or the [MIBBI](#) checklists, available via the BioSharing portal? *

Yes No

- Authors provide text based on journal guidelines
https://www.nature.com/sdata/publish/submission-guidelines#data_citations
 - Reference within the manuscript
 - And in the designated reference list for data citations
- Reinforced during manuscript submission, which prompts provision of information (metadata) needed to create a functional data citation
- Also reinforced throughout the editorial and production process

Making it easier to deposit data during manuscript submission



Scientific Data maintains (and shares) a list of recommended data repositories and requires deposition into a community (discipline specific) repository, where these repositories are available for a researcher's dataset

Upload Files | Replace Files | File Type | File Description | File Order | Dataset Deposition | Dataset Upload

FigShare Upload *

i) Upload File | ii) File Description

These files will be stored privately on **figshare** by default.

By default, all authors associated with the submission will be listed as potential authors for each data file, although the names of authors who did not contribute to the preparation of a file should be removed for that file.

Accessing this tab multiple times will enable you to upload multiple files to the private space in **figshare** that will be accessible by editors and reviewers.

Book1.xlsx ✓ uploaded

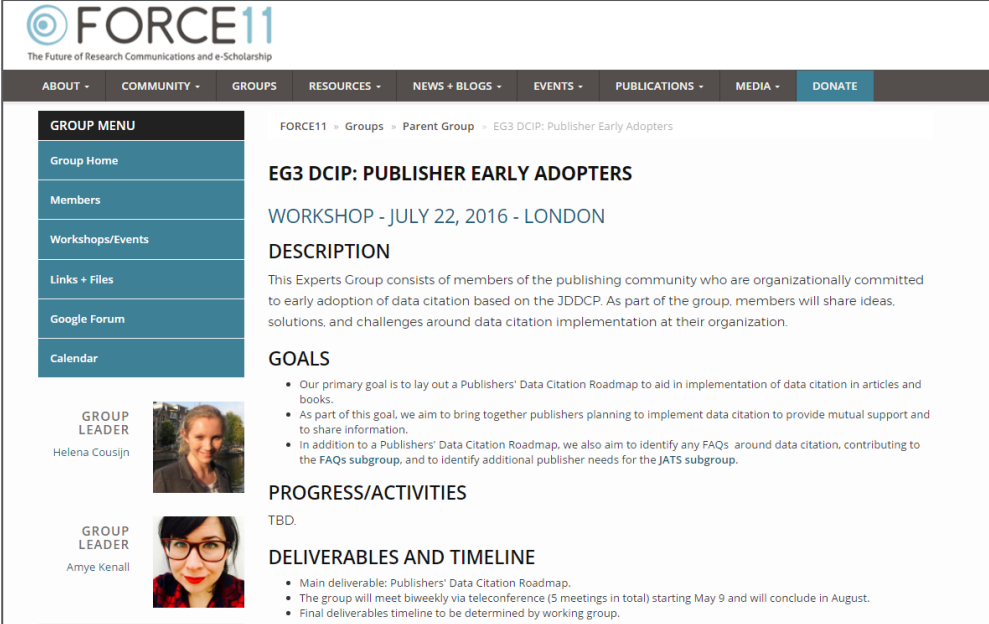
Choose file(s) to upload

Save and continue

For datasets that do not have a suitable community repository (about 50% of datasets), *Scientific Data* offers integrated upload to the figshare and dryad data repositories

Persistent identifiers and data citations can be created during submission

Making it work at scale: a Data Citation Roadmap Publishers



FORCE11
The Future of Research Communications and e-Scholarship

ABOUT - COMMUNITY - GROUPS - RESOURCES - NEWS + BLOGS - EVENTS - PUBLICATIONS - MEDIA - DONATE

GROUP MENU

- Group Home
- Members
- Workshops/Events
- Links + Files
- Google Forum
- Calendar

FORCE11 > Groups > Parent Group > EG3 DCIP: Publisher Early Adopters

EG3 DCIP: PUBLISHER EARLY ADOPTERS

WORKSHOP - JULY 22, 2016 - LONDON

DESCRIPTION

This Experts Group consists of members of the publishing community who are organizationally committed to early adoption of data citation based on the JDDCP. As part of the group, members will share ideas, solutions, and challenges around data citation implementation at their organization.

GOALS

- Our primary goal is to lay out a Publishers' Data Citation Roadmap to aid in implementation of data citation in articles and books.
- As part of this goal, we aim to bring together publishers planning to implement data citation to provide mutual support and to share information.
- In addition to a Publishers' Data Citation Roadmap, we also aim to identify any FAQs around data citation, contributing to the FAQs subgroup, and to identify additional publisher needs for the JATS subgroup.

PROGRESS/ACTIVITIES

TBD.

DELIVERABLES AND TIMELINE

- Main deliverable: Publishers' Data Citation Roadmap.
- The group will meet biweekly via teleconference (5 meetings in total) starting May 9 and will conclude in August.
- Final deliverables timeline to be determined by working group.

GROUP LEADER
Helena Cousijn

GROUP LEADER
Amye Kenall

<https://www.force11.org/group/dcip/eg3publisherearlyadopters>

- Working group to define a Publishers' Data Citation Roadmap to aid in implementation of data citation in articles and books
- Covers policy, author guidelines and content production and technical issues
- Many publishers collaborating with infrastructure providers and policy makers to take a common approach
- Should keep process simple for researchers while delivering benefits of better measurement of credit and reuse

A Data Citation Roadmap for Scientific Publishers. Helena Cousijn, Amye Kenall, Emma Ganley, Melissa Harrison, David Kernohan, Fiona Murphy, Patrick Polischuk, Maryann Martone, Timothy Clark
 bioRxiv 100784; doi: <https://doi.org/10.1101/100784>

Giving researchers more support to share cite data

To help Springer Nature authors and journals follow good practice in sharing and archiving of research data, we provide a free-to-use Research Data Support helpdesk and an optional data deposition and data curation service, Research Data Support. Advise and provide support on all aspects of sharing data associated with publications including data citation.

Researchers submit their data files securely

The Research Data team checks data and curates the metadata

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

<https://go.nature.com/ResearchDataServices>

Example output of Research Data Support

Paper published in Nature
(<https://doi.org/10.1038/nature23654>)

nature
International journal of science

Altmetric: 180 Citations: 9 [More detail >>](#)

Letter | Published: 30 August 2017

Early members of 'living fossil' lineage imply later origin of modern ray-finned fishes

Sam Giles , Guang-Hui Xu, Thomas J. Near & Matt Friedman

Nature **549**, 265–268 (14 September 2017) | [Download Citations](#)

Data availability

The CT data that support the findings of this study, as well as 3D surface files of described material, are available in figshare⁴³ with the identifier <https://doi.org/10.6084/m9.figshare.c.3814360>. All other data files are included in the paper.

Data availability statement included with the paper

43. Giles, S. Xu, G.-H., Near, T. J. & Friedman, M. *Fukangichthys*: CT scan data and surface files from Middle Triassic fossil scanilepiform fish.

<https://doi.org/10.6084/m9.figshare.c.3814360> (2017)

[Show context](#)

Google Scholar

Data citation included as Reference 43

Dataset published in the Springer Nature figshare repository
(<https://doi.org/10.6084/m9.figshare.c.3814360>)

figshare

Fukangichthys: CT scan data and surface files from middle Triassic fossil scanilepiform fish

77 views | 1 citations

7

Published on 30 Aug 2017 - 17:30

This collection includes: CT scan data (.vol files) and associated metadata (.xtekt) files for reconstructing the specimens *Fukangichthys* IVPP V4096.6 and *Fukangichthys* IVPP V4096.13; a reconstructed Mimics file (.mcs file) for *Fukangichthys* IVPP V4096.6 and 3D surface files (.ply) for each specimen.

X-ray computed microtomography scanning for these specimens was performed at IVPP, Chinese Academy of Sciences (CAS), Beijing, China, using a 225 kV microCT. After scanning, data were segmented in Mimics (biomedical.materialise.com/mimics; Materialise, Leuven, Belgium). Surface meshes were then exported into and imaged in Blender (blender.org; Stitching Blender Foundation, Amsterdam, the Netherlands).

Most scanilepiform fossils are heavily compressed, limiting investigations to external anatomy. The Middle Triassic *Fukangichthys* represents an important exception. High-resolution micro computed tomography (μCT) of three-dimensionally preserved skulls

[Read more](#)

CITE THIS COLLECTION
Giles, Sam; Xu, Guang-Hui; Near, Thomas J.; Friedman, Matt (2017): *Fukangichthys*: CT scan data and surface files from middle Triassic fossil scanilepiform fish. figshare.
<https://doi.org/10.6084/m9.figshare.c.3814360>
Retrieved: 09:28, Sep 07, 2017 (GMT)

KEYWORD(S)
living fossil ray-finned fishes
Ray-finned fish Actinopterygii
Paleozoic taxa taxonomy evolution
paleontology paleontology
polyplend Paleozoic ray-fins
actinopterygian vertebrate diversity

How do we increase adoption of data citation faster?

Learning from elements of data focused journals such as Scientific Data, publishers need to work in three key areas:

1. Adoption of clearer and stronger policies on data citation in reference lists and promote the benefits to researchers
2. Implementation of these policies with checks, processes and supporting tools in the manuscript submission and peer-review process – that are appropriate for the research communities served by the journal(s) or publisher
3. Changes in the content production workflow to capture data citations and make them available to readers, and other content consumers in a meaningful and machine-readable way

Thank you

Iain Hrynaszkiewicz

iain.hrynaszkiewicz@nature.com

@iainh_z

For more information on Research Data Support and other data-related activities at Springer Nature:

Email: researchdata@springernature.com

Website: <http://go.nature.com/ResearchDataServices>

Slide acknowledgements:

Rebecca Grant, Varsha Khodiyar

The story behind the image



Chien Shiung Wu (1912–1997)

Chien Shiung Wu was a Chinese American experimental physicist best known for conducting The Wu experiment that bears her name. This experiment showed that the conservation of parity was violated by a weak interaction and it was possible to distinguish between a mirrored variation of the world and the mirror image of the current world. This discovery earned Wu the Wolf Prize in Physics in 1978.

This presentation is licensed as CC-BY-ND