SPRINGER NATURE

Illustration inspired by the work of Chien Shiung Wu

Rewarding data sharing through data citation: from policy to practice

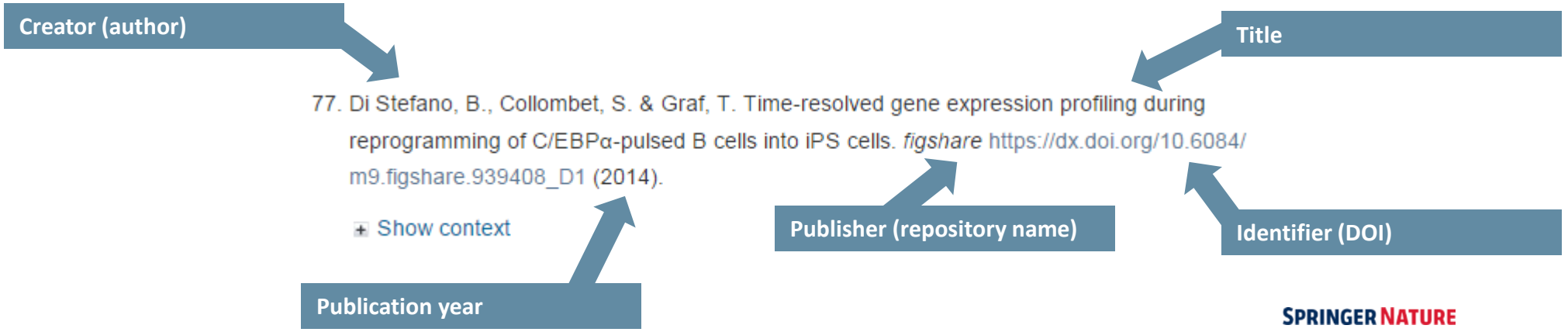
Iain Hrynaszkiewicz

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

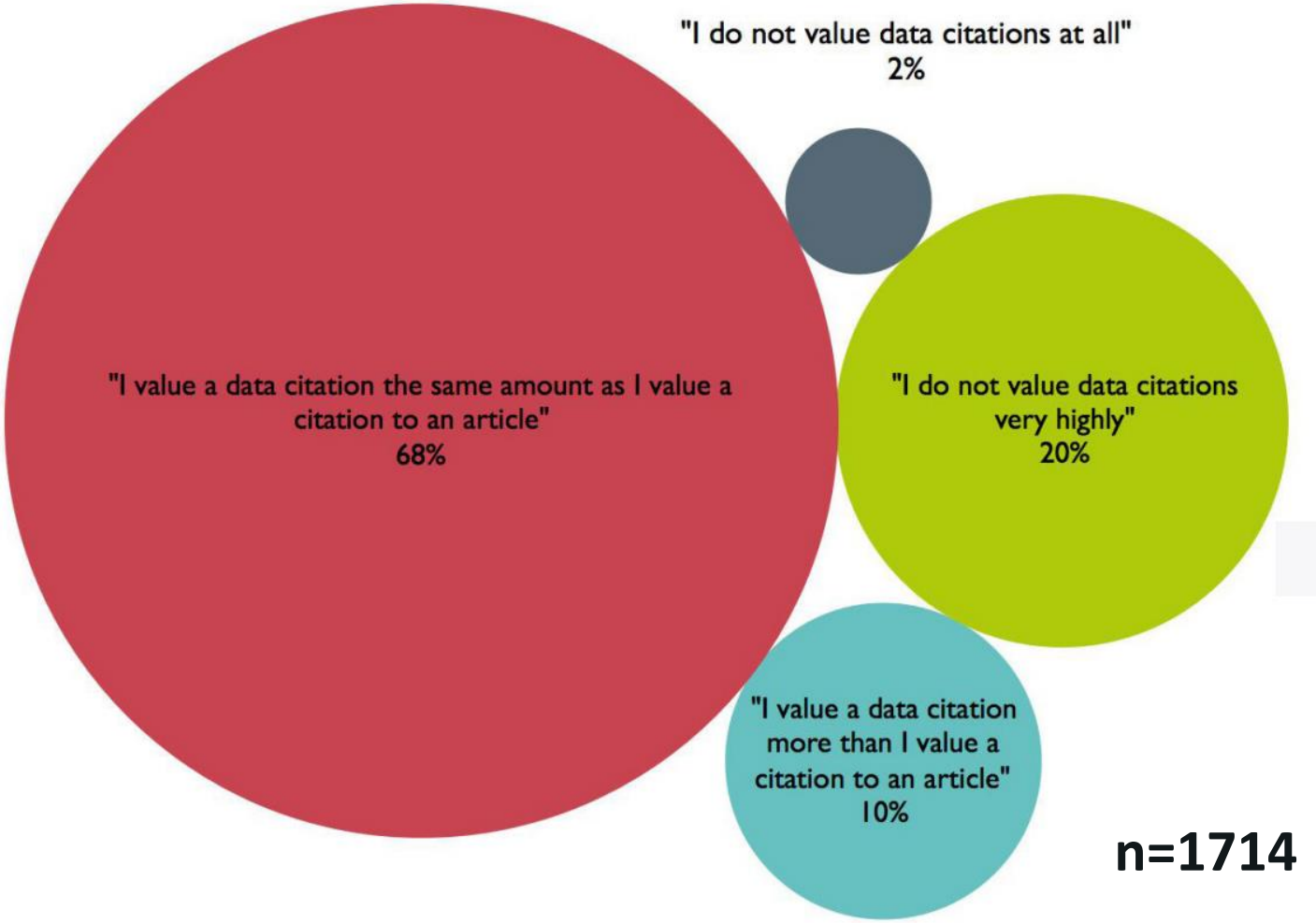


Researchers want to be credited for data reuse

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

Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, et al. (2011) Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6(6): e21101. https://doi.org/10.1371/journal.pone.0021101

Researchers also value credit received from data citations



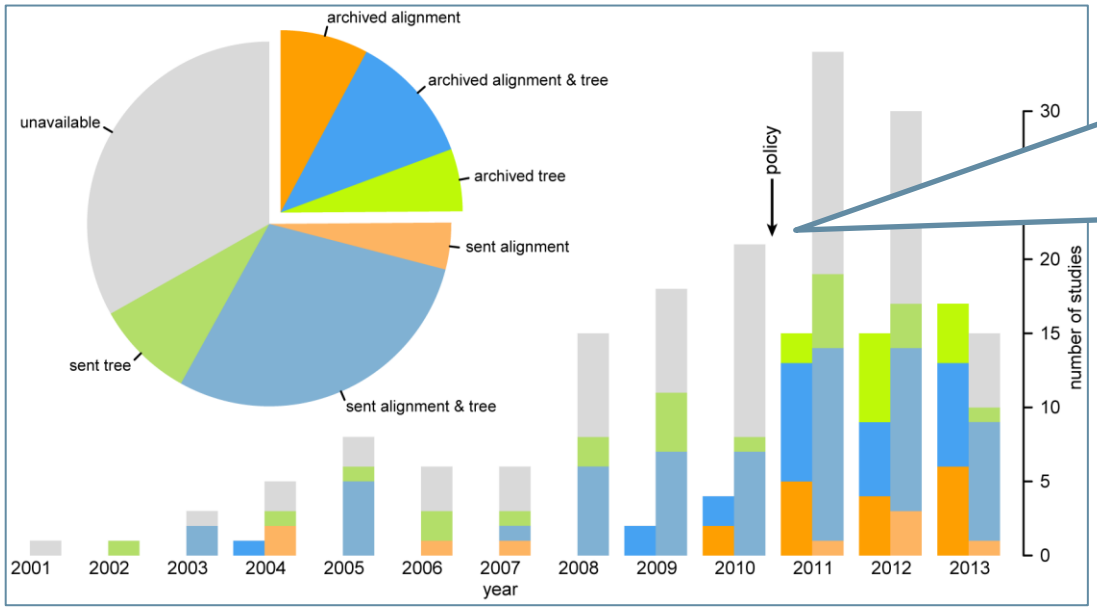
Treadway, Jon; Hahnel, Mark; Leonelli, Sabina; Penny, Dan; Groenewegen, David; Miyairi, Nobuko; Hayashi, Kazuhiro; O'Donnell, Daniel; Digital Science; Hook, Daniel (2016): The State of Open Data Report. figshare. Paper. https://doi.org/10.6084/m9.figshare.4036398.v1

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

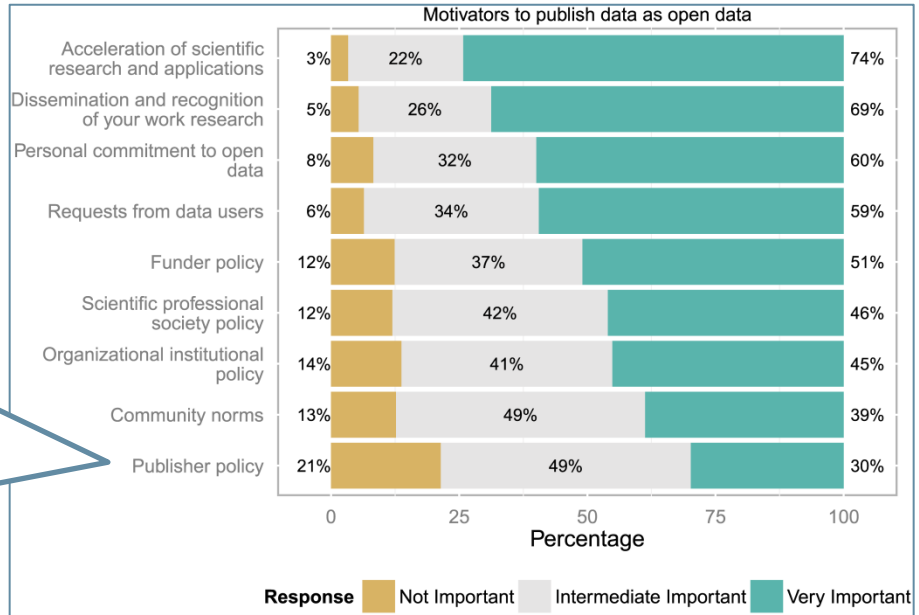
National Science Foundation (US) "*For all new grant applications from 14 January [2013], the US National Science Foundation (NSF) asks a principal investigator to* ***list his or her research "products" rather than "publications" in the biographical sketch section.*** *This means that, according to 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

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



Journal data policies are becoming more consistent

Policy Types

- Springer Nature launched a data policy standardisation initiative in 2016[1]
- 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 Elsevier[2], Wiley[3], Taylor & Francis[4]

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/

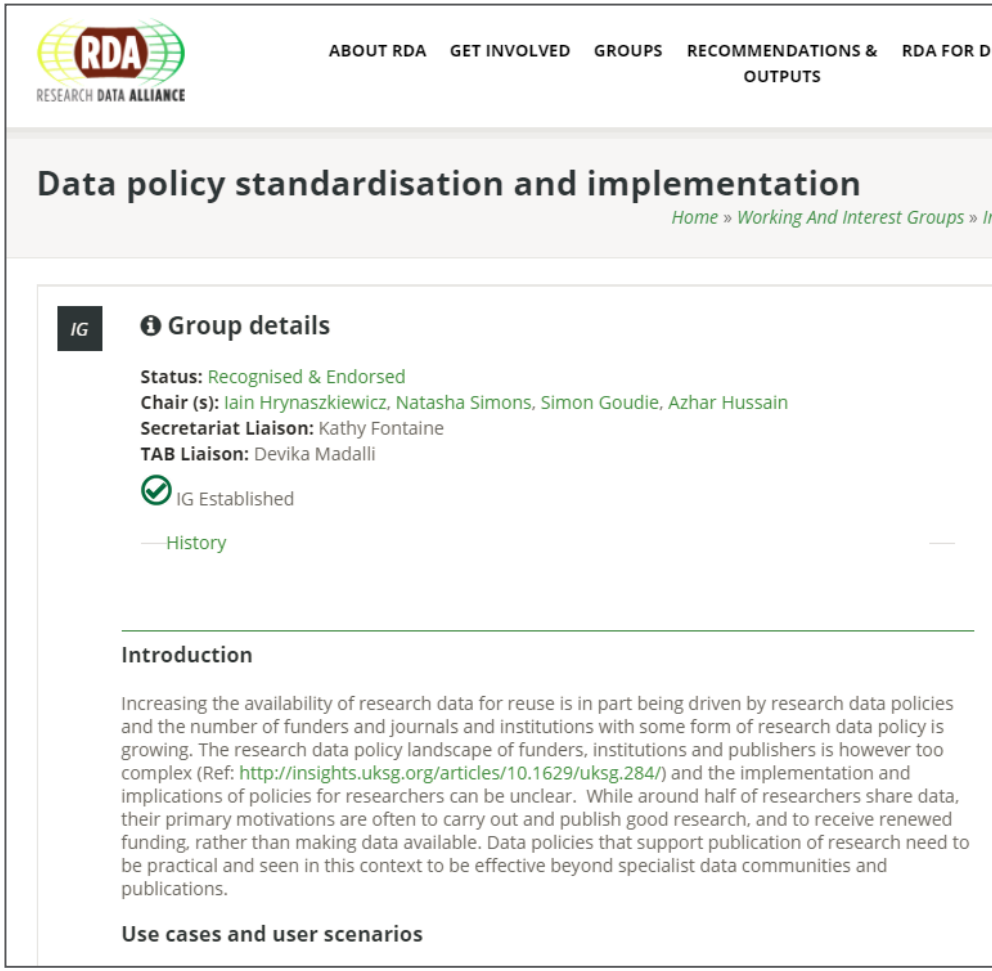
There is motivation to standardise data policies across publishers and other stakeholders

Co-chairs:



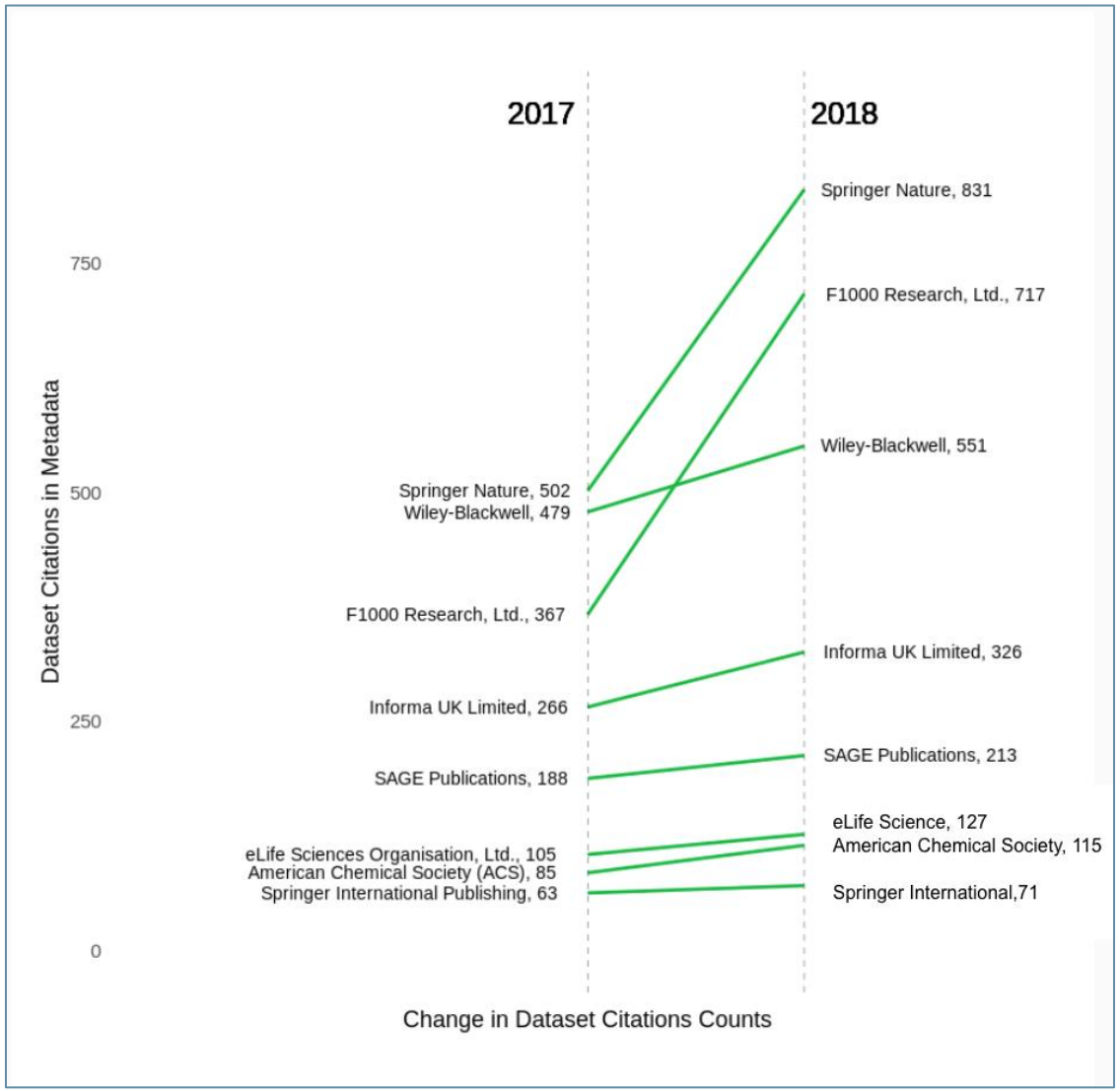

Natasha Simons (ANDS), Simon Goudie (Wiley), Azhar Hussain (Jisc), Iain Hrynaszkiewicz (Springer Nature)

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



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

But are journal and publisher policies increasing data citation?

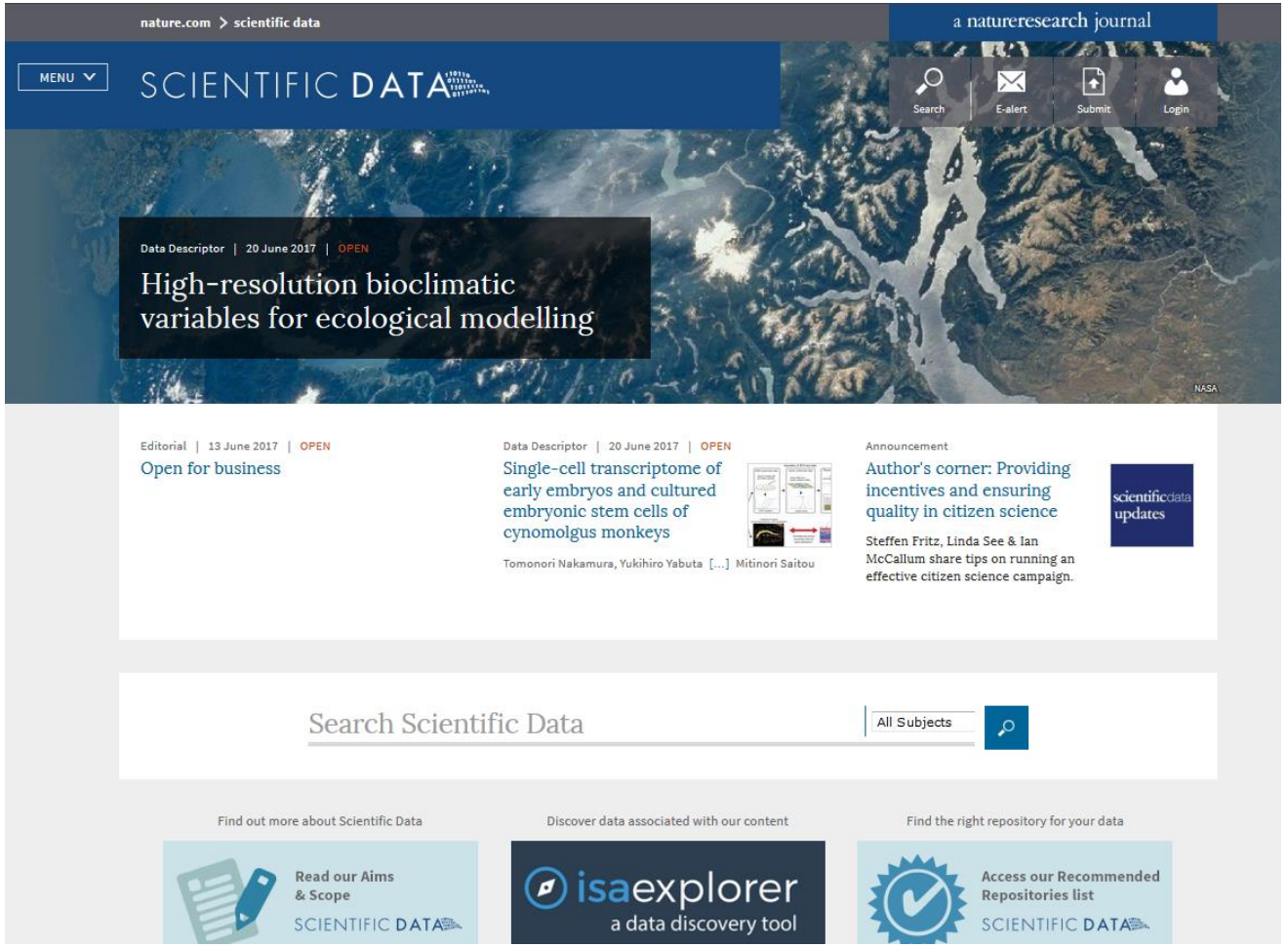


Yes, but progress has been slow

SPRINGER NATURE

Image credit: Kristian Garza and Martin Fenner https://doi.org/10.5438/h16y-3d72 available under CC BY

Scientific Data – at the cutting edge of data citation?



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

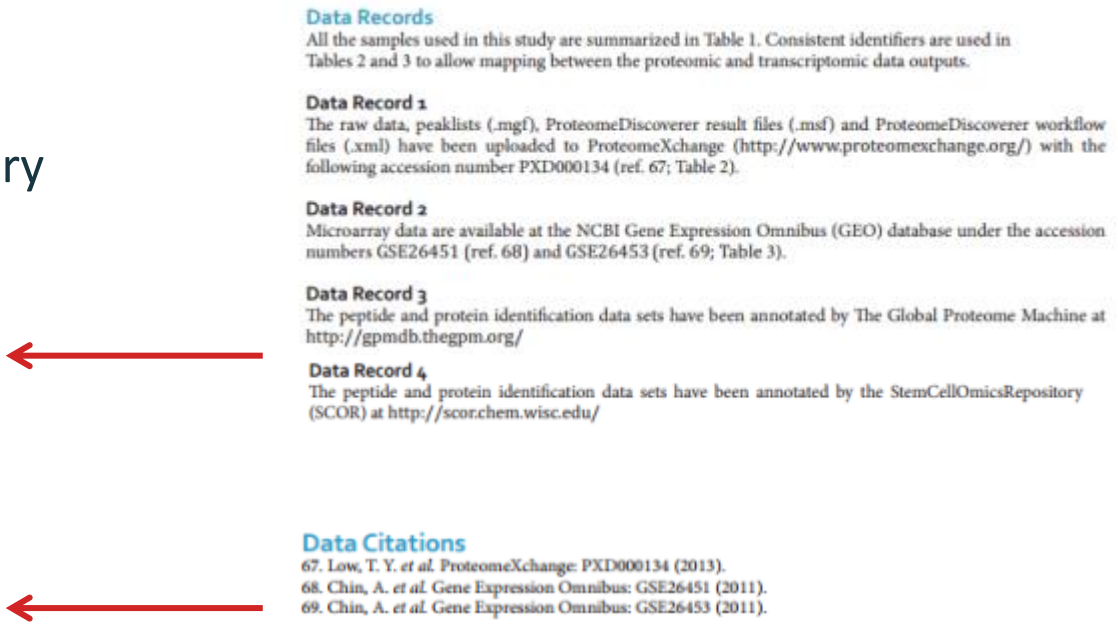
What can we learn from its first 4 years of publishing?

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).

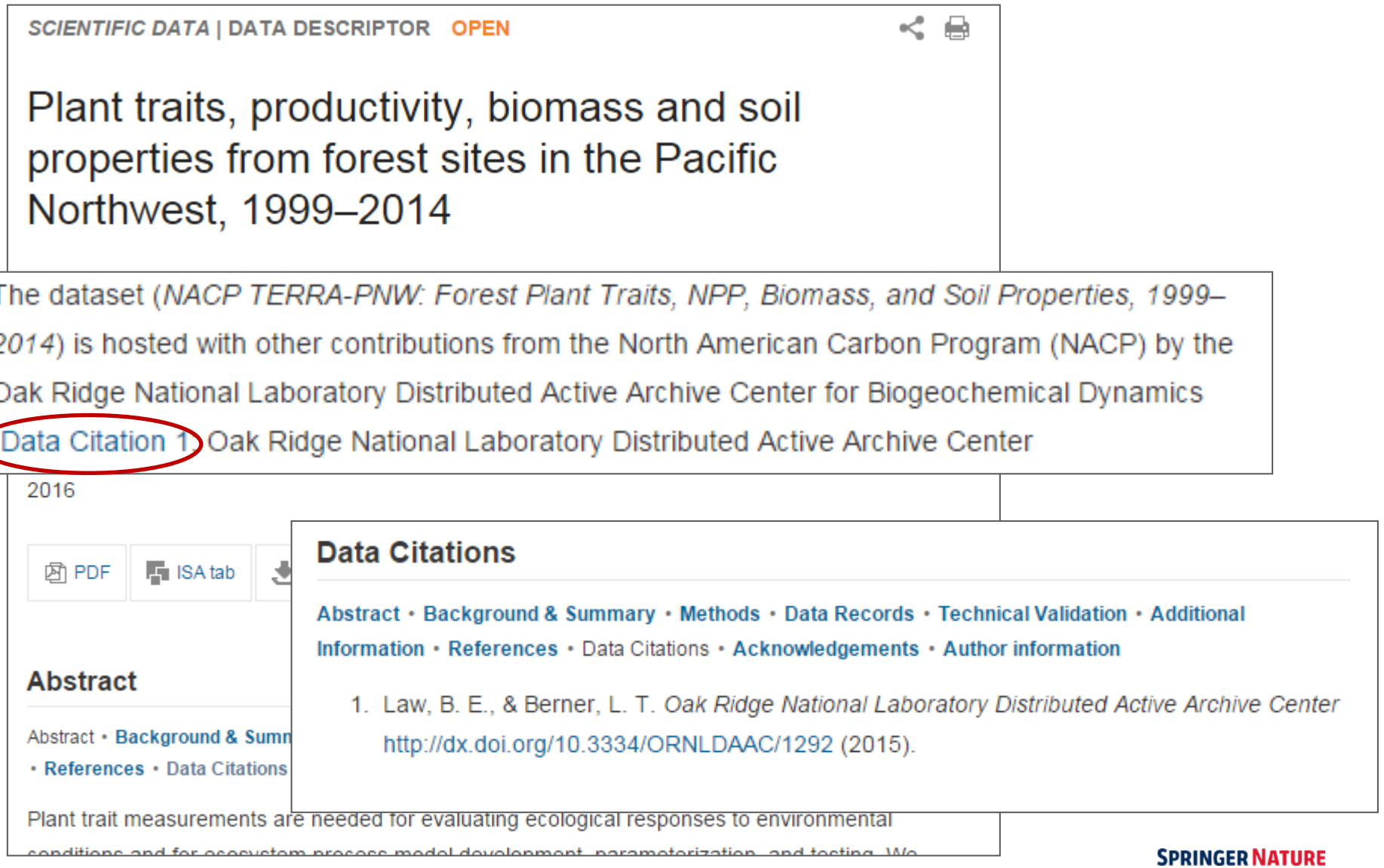
Data Record 3
The peptide and protein identification data sets have been annotated by The Global Proteome Machine at http://gpmdb.thegpm.org/

Data Record 4
The peptide and protein identification data sets have been annotated by the StemCellOmicsRepository (SCOR) at http://scor.chem.wisc.edu/

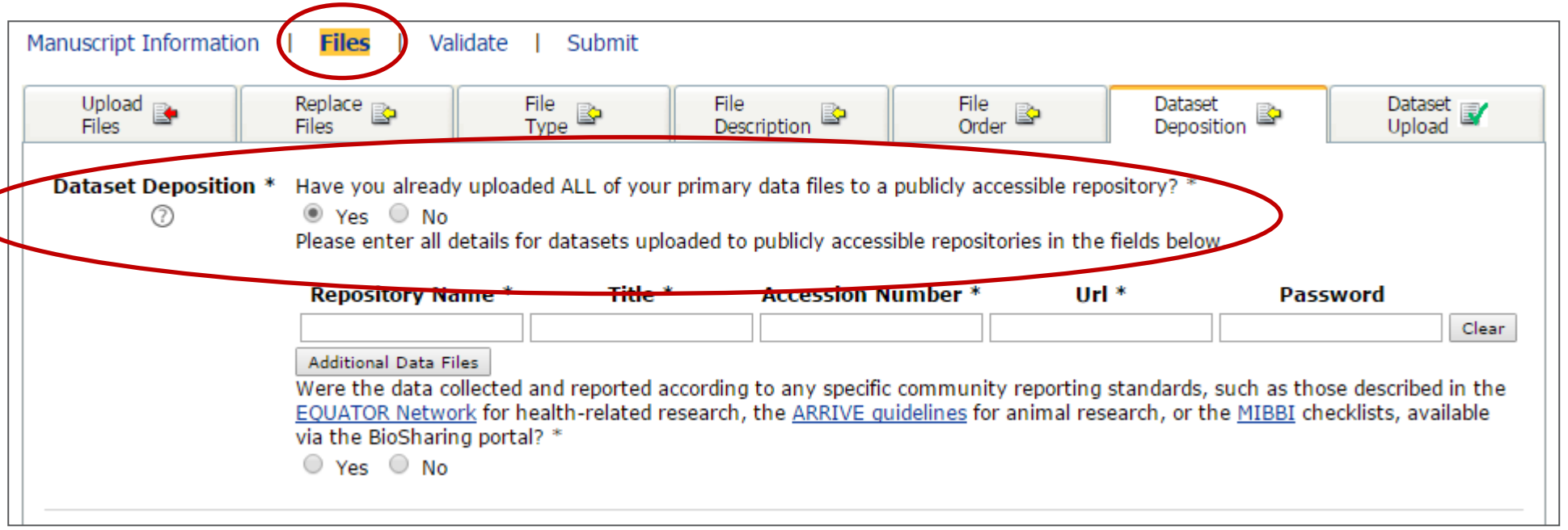
Data Citations

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



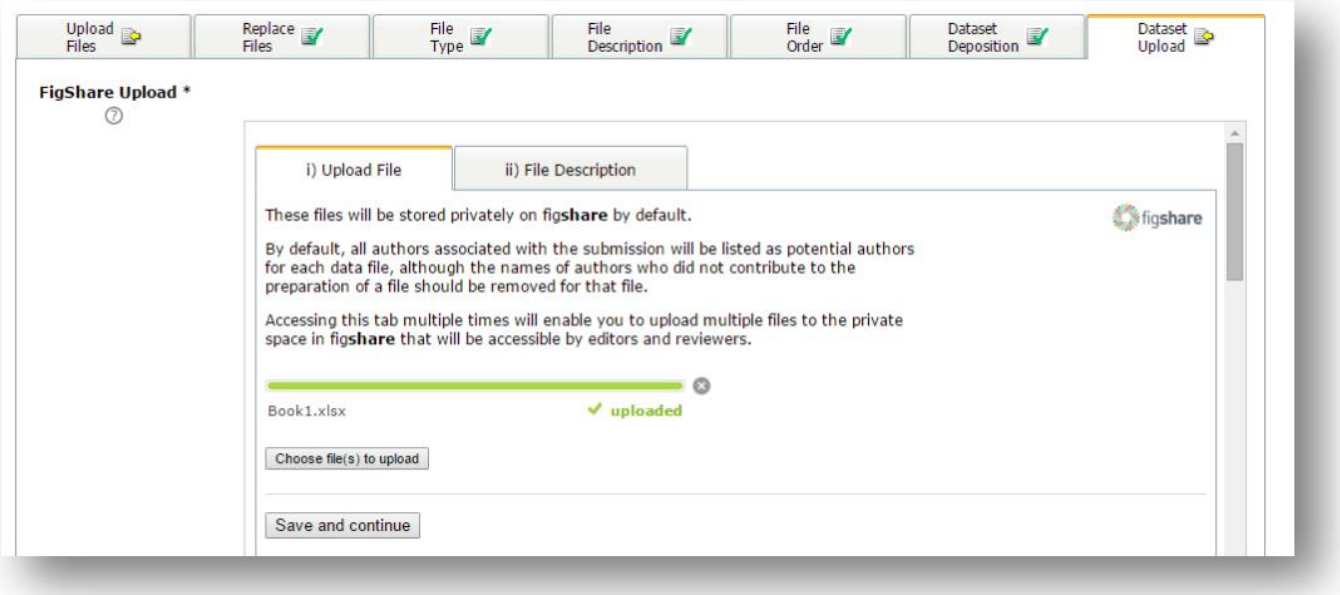
How does *Scientific Data* do it?



- 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

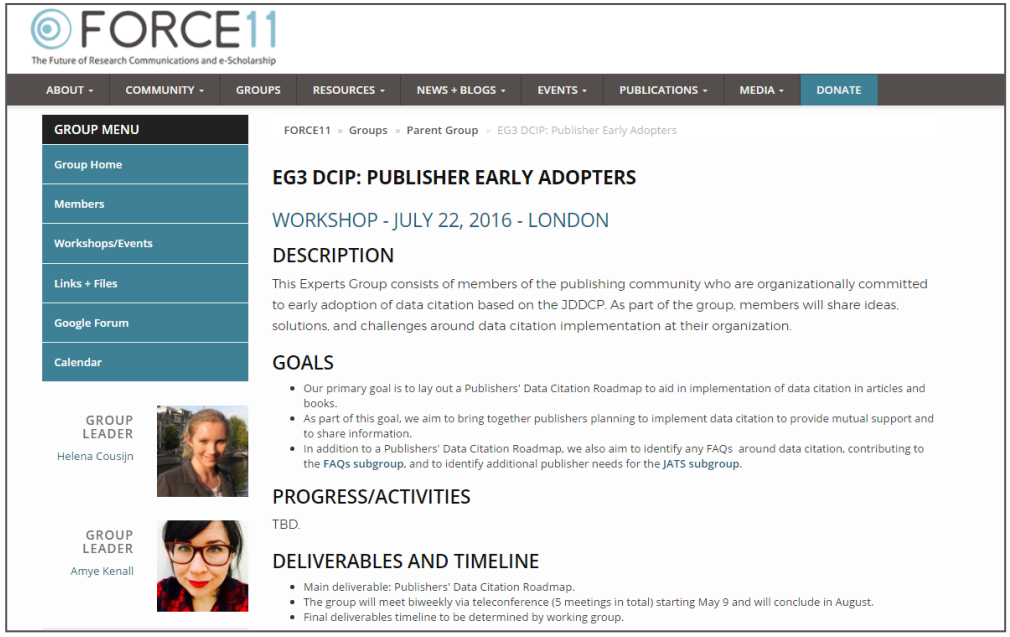
Scientific Data maintains (and shares) a list of recommended data repositories and requires deposition into a community (discipline specific) repository, where these repositories are available for a researcher's dataset



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



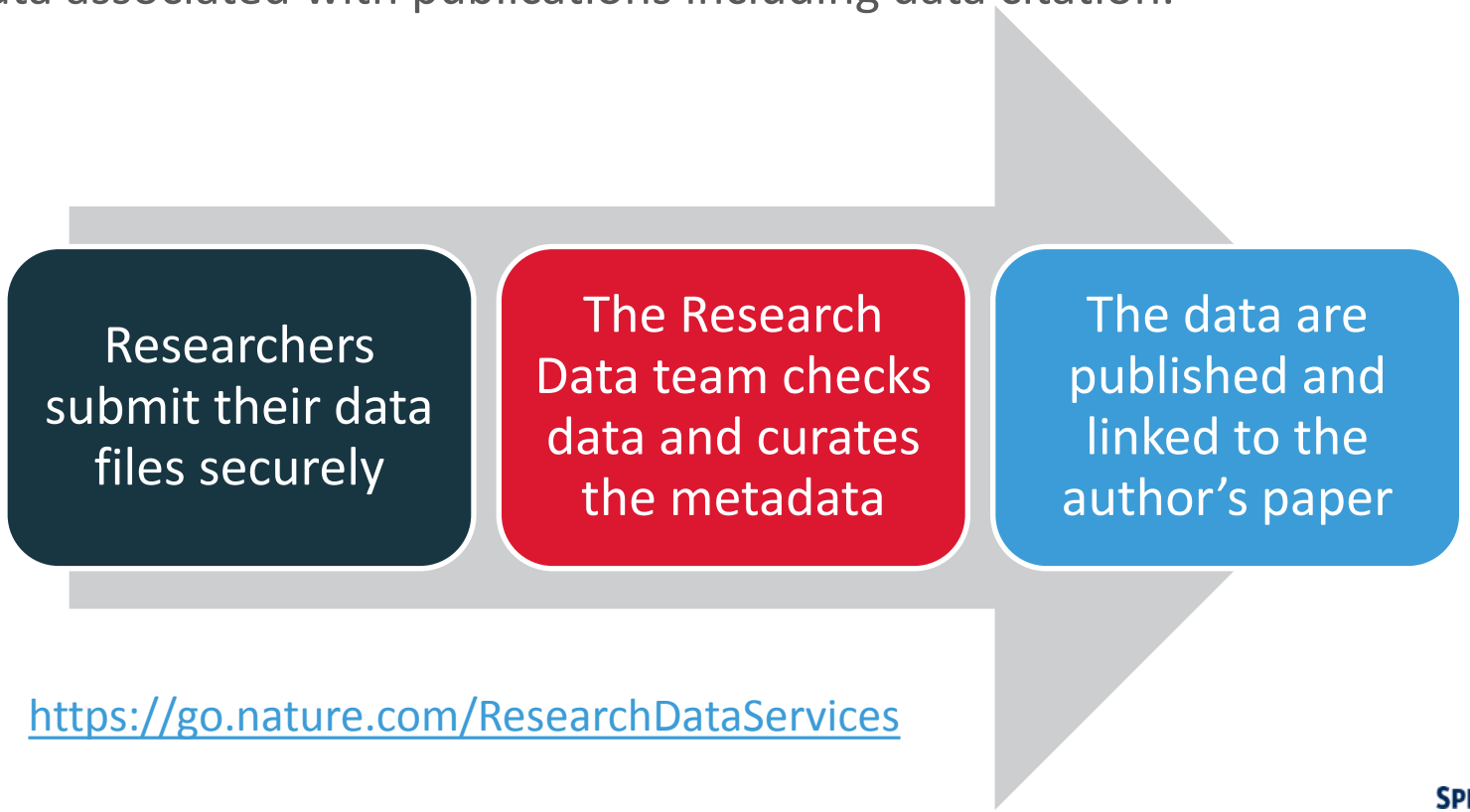
https://www.force11.org/group/dcip/eg3publisherearlyadopters

- 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.

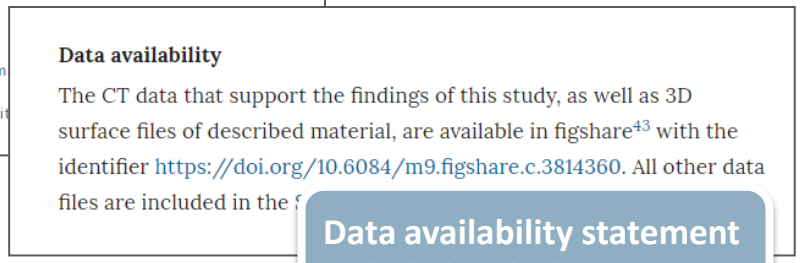


https://go.nature.com/ResearchDataServices

Example output of Research Data Support

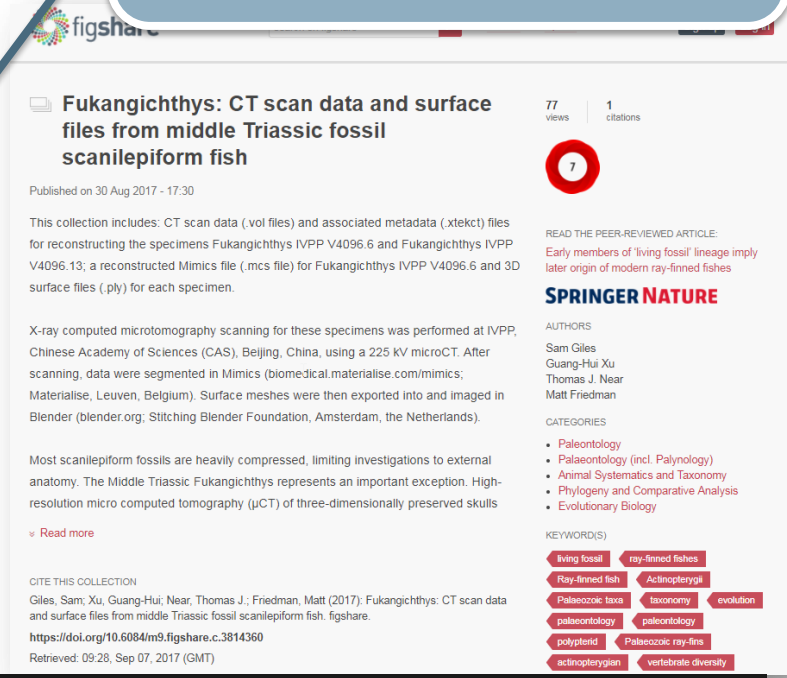
Paper published in **Nature** (https://doi.org/10.1038/nature23654)

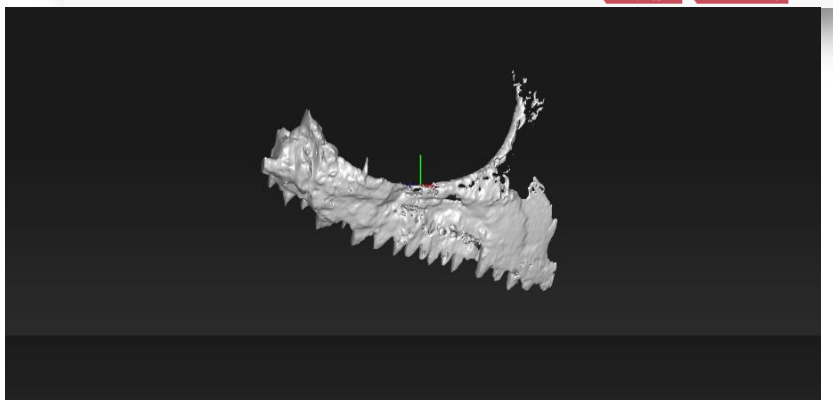
Dataset published in the **Springer Nature figshare repository** (https://doi.org/10.6084/m9.figshare.c.3814360)



Data availability statement included with the paper

Data citation included as Reference 43





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

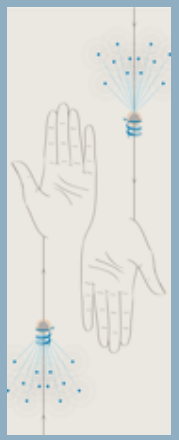
For more information on Research Data Support and other data-related activities at Springer Nature:

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.

SPRINGER NATURE