How to choose a data repository

Dominique Roche

@dom_roche

MARIE CURIE **ACTIONS**

UNIVERSITÉ DE NEUCHÂTEL

DRYAD

Joint Data Archiving Policy (JDAP)

The Joint Data Archiving Policy (JDAP) describes a requirement that data supporting publications be publicly available. This policy was adopted in a joint and coordinated fashion by many leading journals in the field of evolution in 2011, and JDAP has since been adopted by additional journals across various disciplines.

Whitlock, M. C., M. A. McPeek, M. D. Rausher, L. Rieseberg, and A. J. Moore. 2010. Data Archiving. American Naturalist. 175(2):145-146, https://doi.org/10.1086/650340

Rieseberg, L., T. Vines, and N. Kane. 2010. Editorial and retrospective 2010. Molecular Ecology. 19(1):1-22, https://doi.org/10.1111/j.1365-294X.2009.04450.x

Moore, A. J., M. A. McPeek, M. D. Rausher, L. Rieseberg, and M. C. Whitlock. 2010. The need for archiving data in evolutionary biology. Journal of Evolutionary Biology. https://doi.org/10.1111/j.1420-9101.2010.01937.x

http://datadryad.org/pages/jdap

Journals that require open data

Examples:

- The American Naturalist
- Biological Journal of the Linnean Society
- Biology Letters
- BMC Ecology
- BMC Evolutionary Biology
- BMJ
- BMJ Open
- Ecological Applications
- Ecological Monographs
- Ecology
- Ecosphere
- Evolution
- Evolutionary Applications
- Frontiers in Ecology and the Environment
- Functional Ecology
- Genetics
- Heredity

…

AAAS
ADVANCING SCIENCE, SERVING SOCIETY

nature

JOURNAL OF Evolutionary Biology

THE ROYAL SOCIETY

Journal of Ecology
Journal of Applied Ecology
Methods in Ecology and Evolution
Functional Ecology
Journal of Animal Ecology

PLoS
PUBLIC LIBRARY of SCIENCE

EVOLUTION
INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

Living database of journal data policies in E&E

Contributors: Ilias Berberi, **Dominique Roche**

Date created: 2021-11-15 12:15 PM | Last Updated: 2021-11-16 03:37 PM

Identifier: DOI 10.17605/OSF.IO/D6SP3

Category: Data

40 of 199 journals in E&E (20%) require open data

33 (17%) require a data availability statement
74 (37%) recommend open data
55 (28%) have no data policy

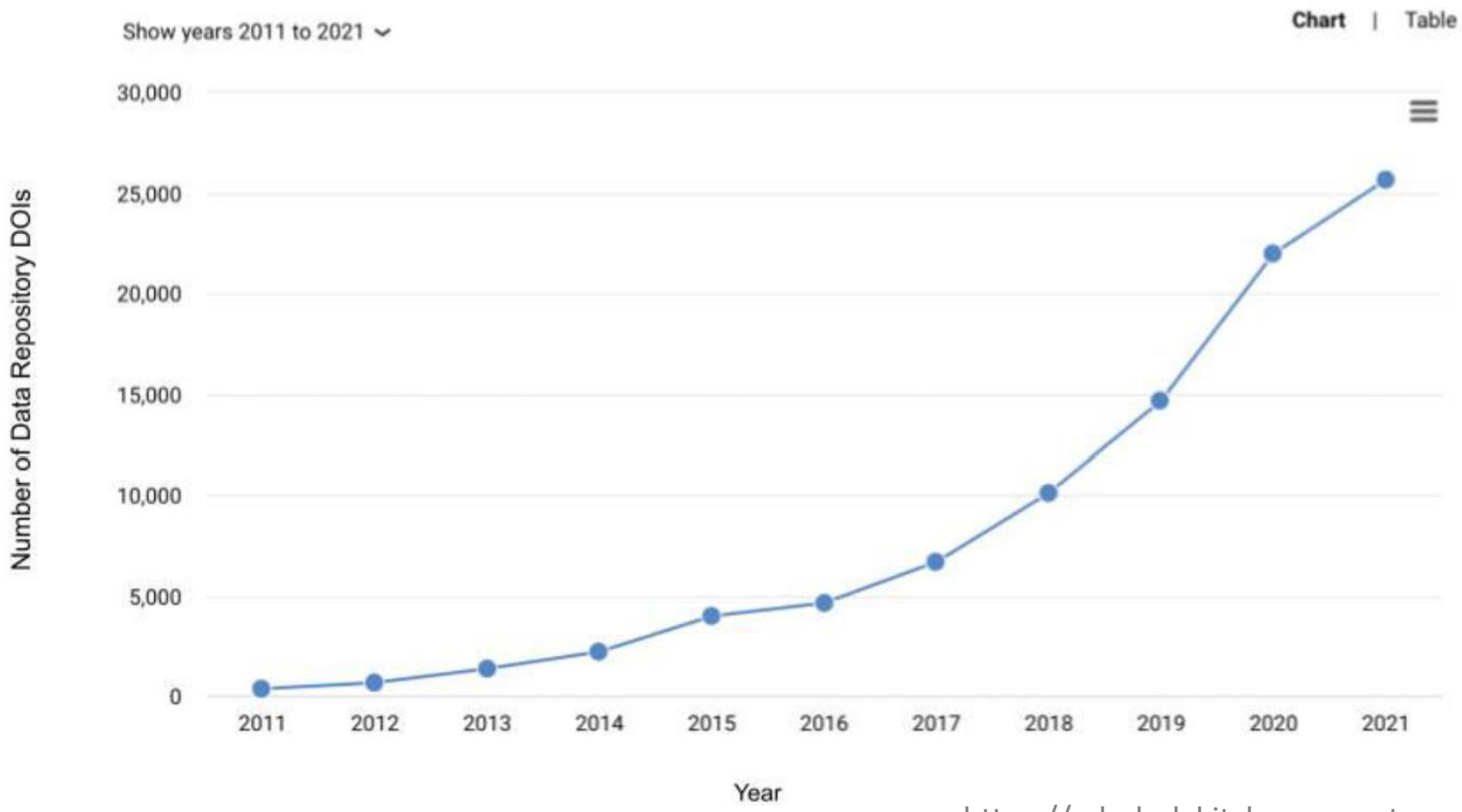
THE SCHOLARLY kitchen

What's Hot and Cooking In Scholarly Publishing

Guest Post: A Decade of Open Data in Research – Real Change or Slow Moving Compliance?

Growth in Figshare, Dryad, Dataverse, Mendeley Data, Open Science Framework and Vivli DOIs in the published literature (2011-2021)

By MARK HAHNEL | MAR 30, 2022



https://scholarlykitchen.sspnet.org

Why share data in a repository?

- avoids data being lost to science (and you!)
- websites are ephemeral (journals too)
- supplementary material = paywalled & not findable
- boost collaboration, get a DOI, gain credit

Open data = more citations

2007

Sharing Detailed Research Data Is Associated with Increased Citation Rate

Heather A. Piwowar*, Roger S. Day, Douglas B. Fridsma

PLOS one

2013

Data reuse and the open data citation advantage

Heather A. Piwowar[1,2] and Todd J. Vision[1,2,3]

PeerJ

2020

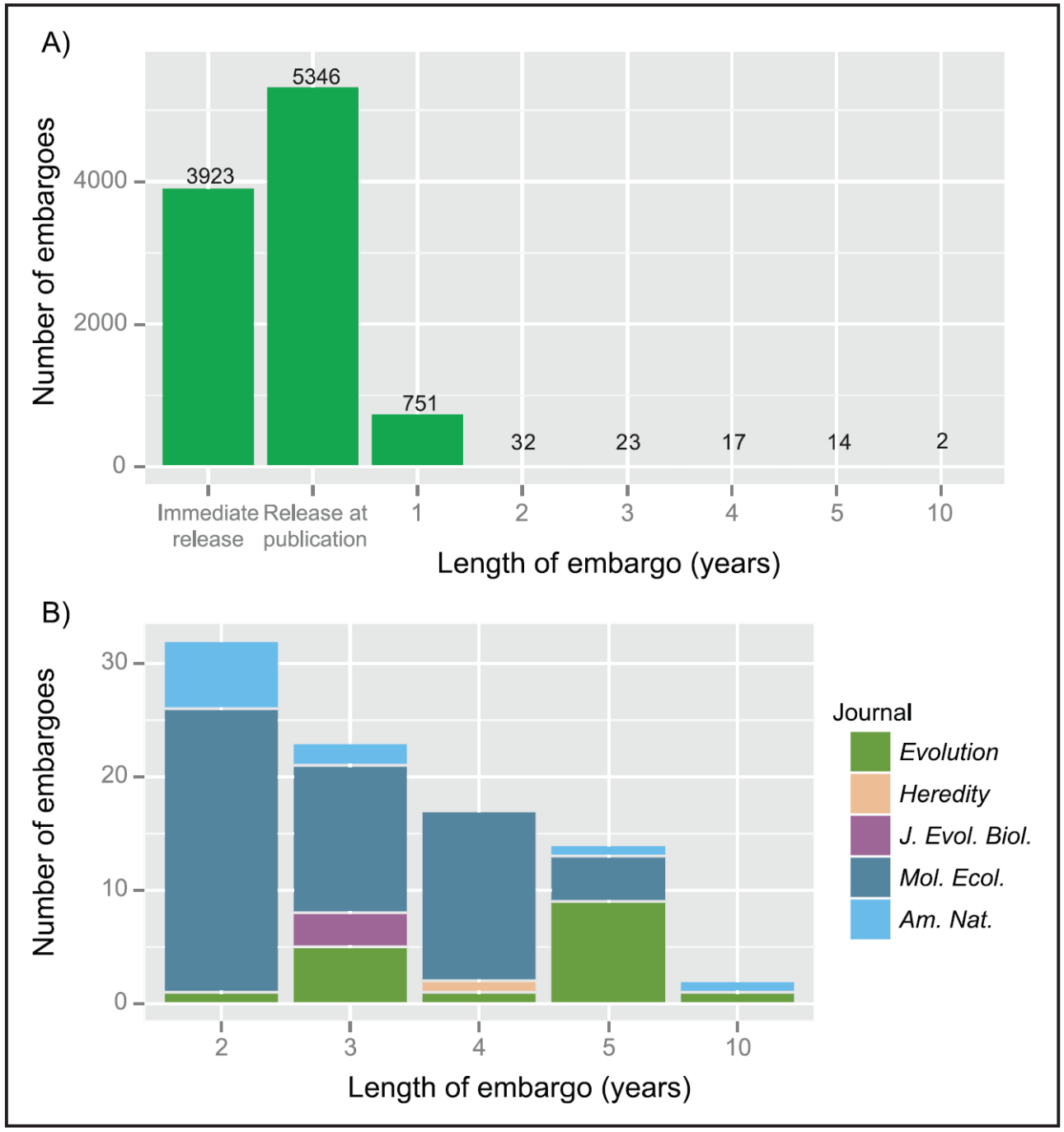
The citation advantage of linking publications to research data

Giovanni Colavizza[1,2], Iain Hrynaszkiewicz[3,4], Isla Staden[1,5], Kirstie Whitaker[1,6], Barbara McGillivray[1,6]*

PLOS ONE

Why share data in a repository?

- avoids data being lost to science (and you!)
- websites are ephemeral (journals too)
- supplementary material = paywalled & not findable
- get a DOI, boost collaboration, gain credit
- comply with policy + safeguard data you plan to reuse



Embargoes

10,108 archived datasets (2011-2013)

(Roche et al 2014 *PLOS Biol*)

Data Repositories: Institutional vs. ‘International’

Institutional

- For researchers at institution

Research Data JCU

SWISS base

Boston University Libraries
BU Research Data

‘International’

- For all researchers

GBIF

GenBank NCBI

zenodo

OSF

Hybrid

- Dataverse & Figshare for institutions

HARVARD Dataverse

figshare for institutions

Data Repositories: Disciplinary vs. Generalist

View data repositories

- **Biological sciences:** Nucleic acid sequence; Protein sequence; Molecular & supramolecular structure; Neuroscience; Omics; Taxonomy & species diversity; Mathematical & modelling resources; Cytometry and Immunology; Imaging; Organism-focused resources

- **Generalist repositories**

SCIENTIFIC DATA

https://www.nature.com/sdata/policies/repositories

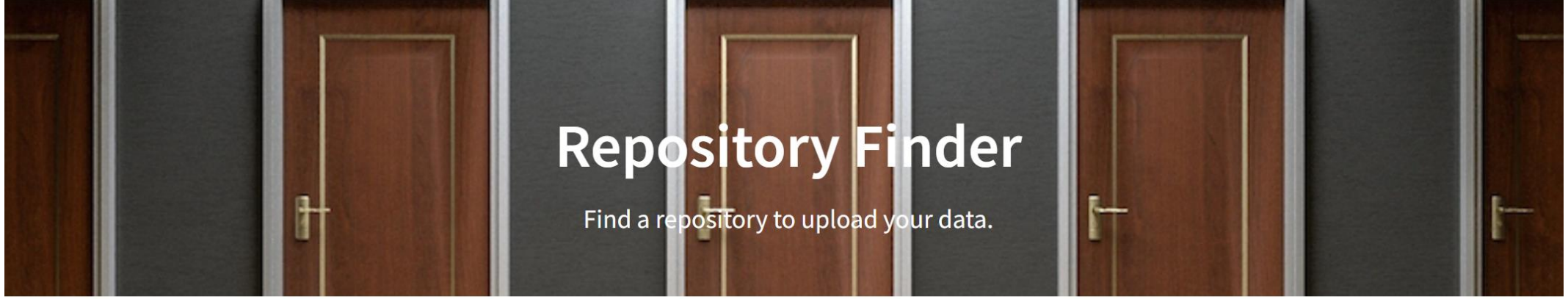
Data Repositories: Disciplinary vs. Generalist

Disciplinary

- GBIF
- Polar Data Catalogue
- GenBank
- DataStream
- KNB
- Pangaea
- Sequence Read Archive

Generalist

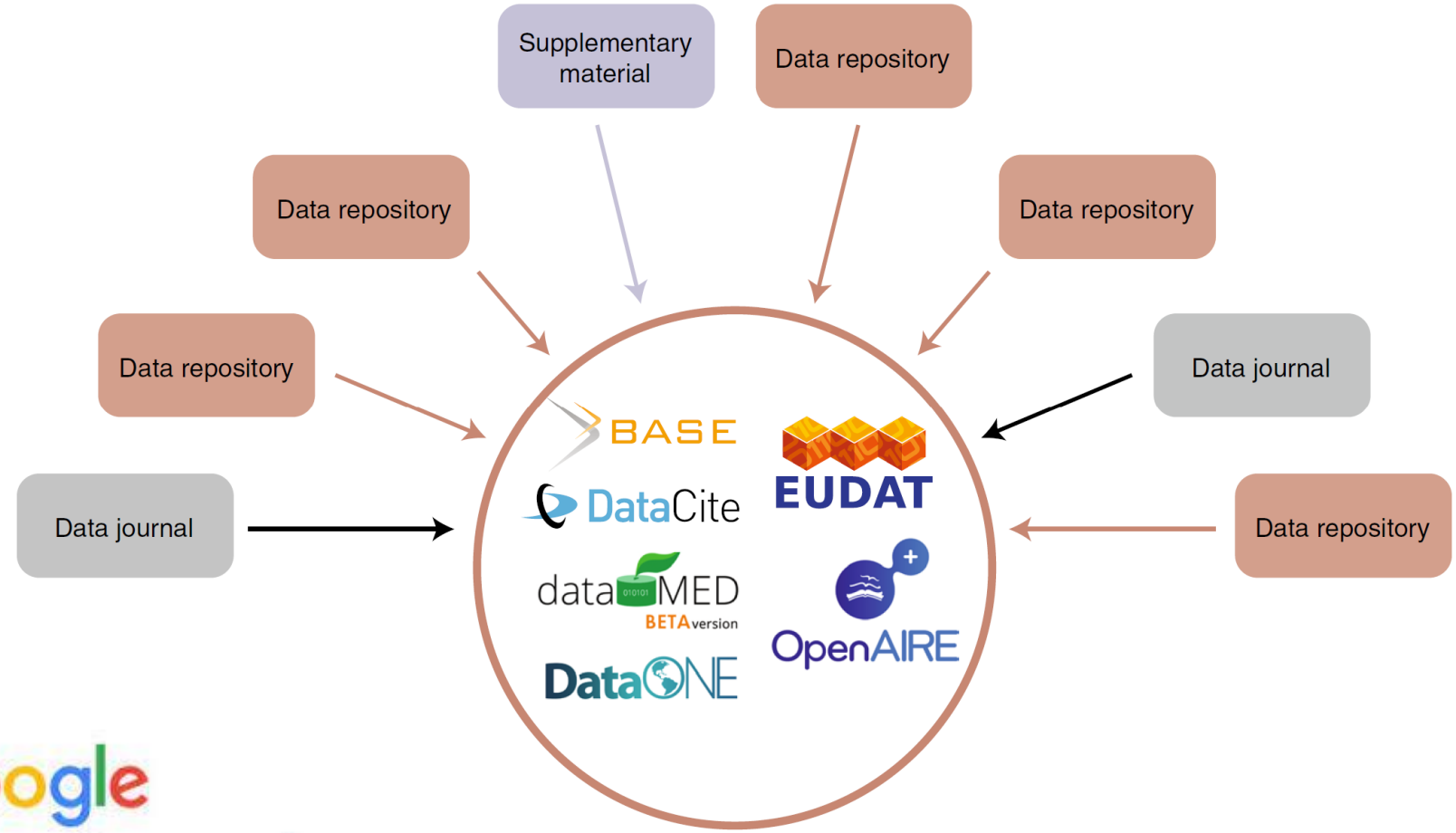
- Dryad
- Figshare
- Zenodo
- Open Science Framework (OSF)
- Dataverse
- Mendeley Data



See the repositories in re3data that meet the criteria of the Enabling FAIR Data Project.

See the repositories in re3data that meet the criteria of the FAIRsFAIR Project.

Aggregators of data repositories



Google Dataset Search Beta

(Culina *et al* 2018 *Nature Ecol Evol*)

Data repositories should meet all of the following requirements:

- Ensure long-term persistence and preservation of datasets
- Be supported by a research community or research institution
- Provide deposited datasets with stable and persistent identifiers
- Allow access to data without unnecessary restrictions.
- Provide clear terms of data use and data access (or licence)
- Facilitate anonymous reviewer access for embargoed data
- Are there costs or restrictions for archiving?
- Are there limits on file size?
- Is version control supported?
- Is there an embargo option?

SCIENTIFIC DATA

www.nature.com/sdata/policies/repositories

SCIENTIFIC DATA

OPEN

COMMENT

The TRUST Principles for digital repositories

Dawei Lin[1]✉, Jonathan Crabtree[2], Ingrid Dillo[3], Robert R. Downs[4], Rorie Edmunds[5], David Giaretta[6], Marisa De Giusti[7], Hervé L'Hours[8], Wim Hugo[9], Reyna Jenkyns[10], Varsha Khodiyar[11], Maryann E. Martone[12], Mustapha Mokrane[3], Vivek Navale[13], Jonathan Petters[14], Barbara Sierman[15], Dina V. Sokolova[16], Martina Stockhause[17] & John Westbrook[18]

Box 1 The TRUST Principles

Principle	Guidance for repositories
Transparency	To be transparent about specific repository services and data holdings that are verifiable by publicly accessible evidence.
Responsibility	To be responsible for ensuring the authenticity and integrity of data holdings and for the reliability and persistence of its service.
User Focus	To ensure that the data management norms and expectations of target user communities are met.
Sustainability	To sustain services and preserve data holdings for the long-term.
Technology	To provide infrastructure and capabilities to support secure, persistent, and reliable services.

CORE TRUST SEAL

www.coretrustseal.org

(Lin *et al* 2020 *Sci Data*)

Additional resources

PERSPECTIVE
https://doi.org/10.1038/s41559-017-0458-2

nature ecology & evolution

Navigating the unfolding open data landscape in ecology and evolution

Antica Culina[1*], Miriam Baglioni[2], Tom W. Crowther[1,3], Marcel E. Visser[1], Saskia Woutersen-Windhouwer[1] and Paolo Manghi[2]

- nature.com/sdata/policies/repositories
- datacite.org
- re3data.org
- openaire.eu/how-to-make-your-data-fair
- fairsharing.org

SORTEE
Society for Open, Reliable, and Transparent Ecology and Evolutionary Biology

www.sortee.org