

Data Standards

(through ecological goggles
- i.e. the practitioners view)

Anders Gravbrøt Finstad

Professor

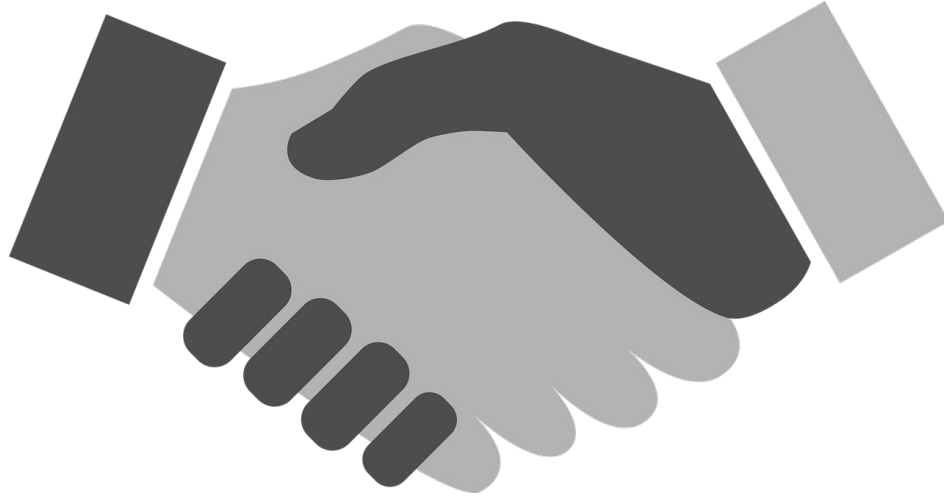
Department of Natural History and Centre for Biodiversity

Dynamics, NTNU

<https://orcid.org/0000-0003-4529-6266>

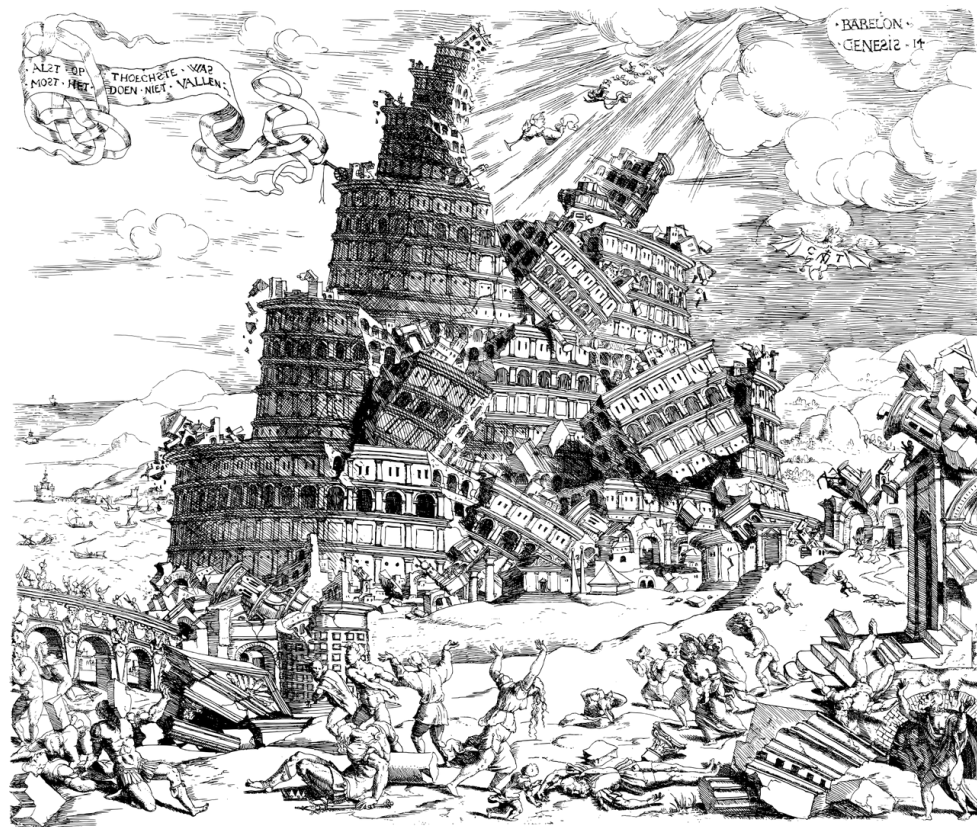
<https://www.ntnu.edu/employees/anders.finstad>

What are data standards?



Data standards are documented agreements on representation, format, definition, structuring, tagging, transmission, manipulation, use, and management of data

And why do we want them?



Communication

The FAIR principles

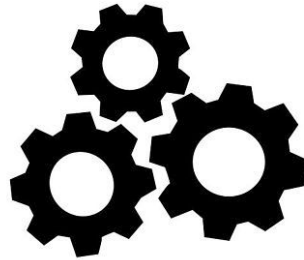
F_{indable}



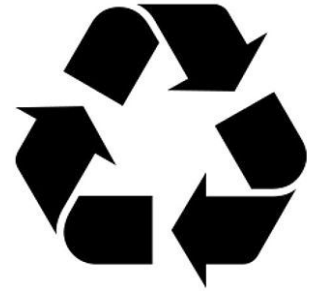
A_{ccessible}



I_{nteroperable}



R_{eusable}





Findable
Accessible
Interoperable
Reusable

Data standards in a narrow sense is about making data interoperable

- Interoperable with your own existing data
- Interoperable with your future data
- Interoperable with your research groups data
- Interoperable with the institutional, national, global etc. data pool

PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS *THE* CORRECT WAY TO WRITE NUMERIC DATES:

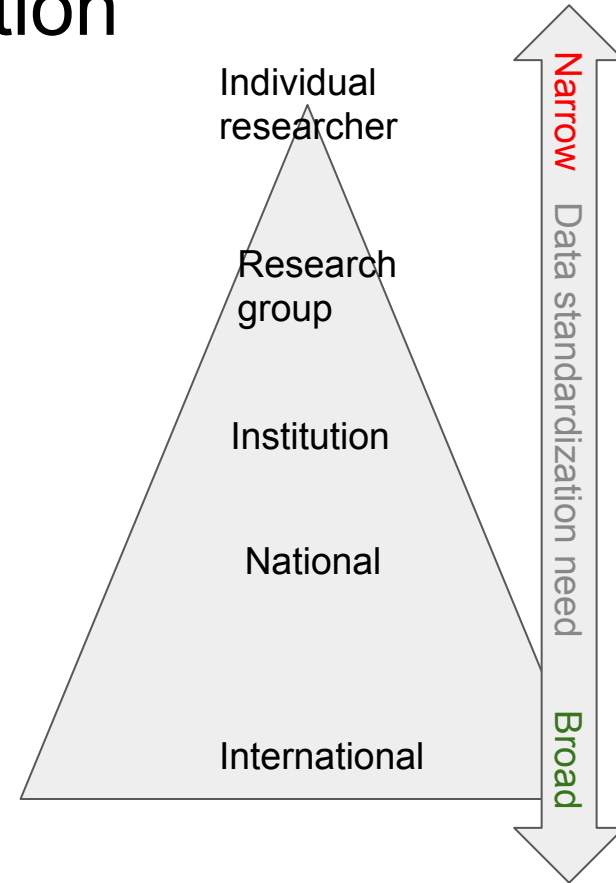
2013-02-27

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013.II.27. $27\frac{1}{2}$ -13 2013.158904109
MMXIII-II-XXVII MMXIII ^{LVII}/_{CCCLXV} 1330300800
 $((3+3) \times (111+1) - 1) \times 3/3 - 1/3^3$ 2013
10/11011/1101 02/27/20/13 0²1³2¹3⁴7⁵8⁶ 

Standardization recognition

- **Data standardization**: common data definitions, formats, representation and structures - broad or narrow recognition
- We all use data standards as part of making data and results useful for our self
- More advantages
 - **Reproducibility** = others (and yourself) may actually understand what your done
 - **Quality** = data integrity





Heterogeneity of ecological data perceived as barrier to publishing and re-use (FAIRification)



Forum

Ecological Data Should Not Be So Hard to Find and Reuse

Timothée Poisot ¹ , Anne Bruneau ^{1, 2}, Andrew Gonzalez ³

Show more



Read & annotate PDF



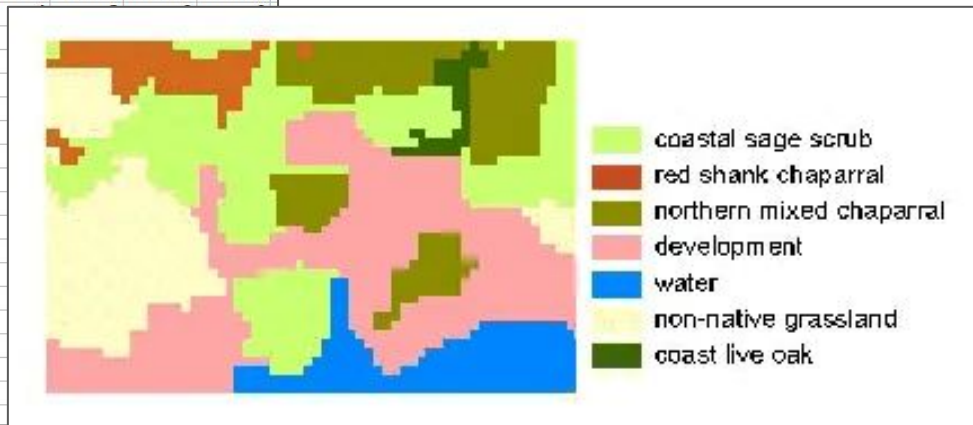
Add to wizdom.ai

<https://doi.org/10.1016/j.tree.2019.04.005>

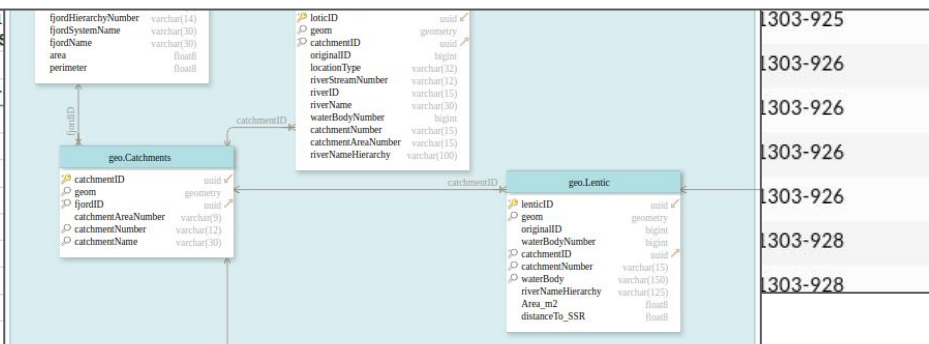
exemplified by Darwin Core or other specialized ontologies), but rather domain-specific data representations informed by real research use-cases, which can be made interoperable. The challenges of standardized data formats lie in creating a way of representing the many possible ways in which ecological data can be integrated, and finding which way is amenable to research work while ensuring the integrity of the information it represents. This requires a discussion at the interface between technical constraints and the needs of researchers, who act as both producers and consumers of data. Although adopting a domain-specific perspective

Heterogeneity of ecological data perceived as barrier to publishing and re-use (FAIRification)

D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
løpnr	dato	lenke	garn	maskevidde	length	kandverg	vekt g	kjønn	stadiu	mage	cyster	alder	redskå	
1	###	NA	NA	NA	382	1	607	2	5					
2	###	2	2	10	117	NA	13,65	2	1					
3	###	1	2	10	115	NA	10,8	1	1					
4	###	1	2	10	84	NA	5,6	2	1					
5	###	2	2	8	105	NA	9,8	2	1					
6	###	2	2	8	75	NA	3,87	2	1					
7	###	2	3	15,5	167	NA	32,88	1	1					
8	###	4	2	29	410	1	700	2	5					
9	###	4	2	29	335	1	379	2	3					
10	###	2	3	10	107	NA	10,47	2	1					
11	###	5	2	10	320	1	302	2	3					
12	###	3	3	8	82	NA	4,17	1	1					
13	###	4	3	8	109	NA	10,18	2	1					
14	###	4	3	8	124	NA	15,25	2	1					
15	###	3	2	10	113	NA	12,12	2	1					
16	###	4	2	15,5	169	NA	3808	2	1					
17	###	2	3	8	108	NA	9,6	1	1					
18	###	2	3	8	112	NA	11,7	2	1					
19	###	5	2	24	443	1	801	1	3					

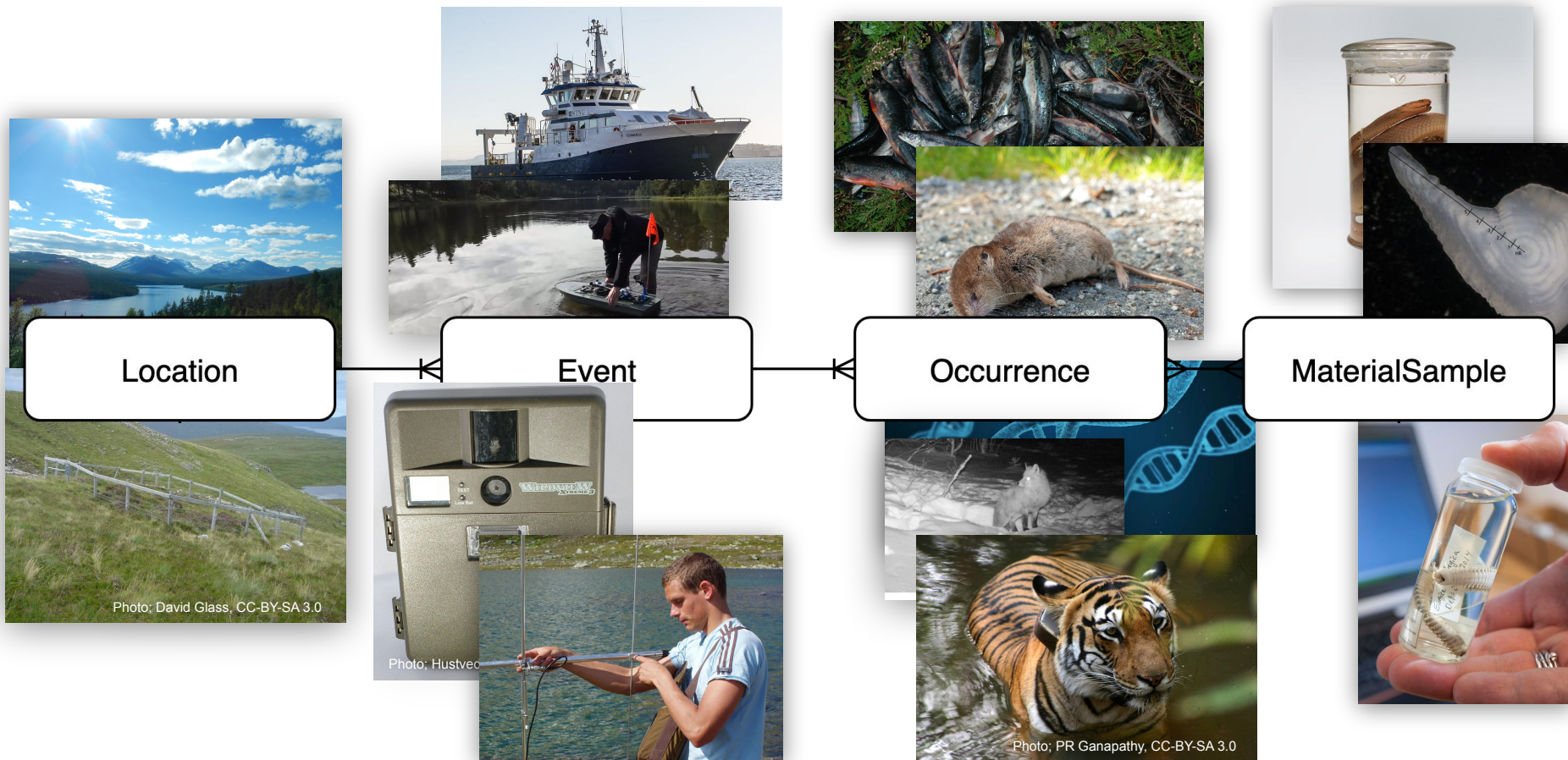


Stasjon	Gml-stasj	Areal	Omganger	L-0+	L-E	A-0+	A-E	L-0+
E-04								
E-05								
E-06	Boj-4	100	3	63	13	66	4	
E-07								
E-08								
E-09	Boj-5	75	1	19	4	13	1	
E-10								
E-11								
E-12	Boj-6	75	1	11	0	8	1	
E-13	Boj-8							
E-14								
E-15								
E-16	Boj-9							
E-17	Side-1							
E-18	Side-2							
E-19	H-4							
E-20	H-2							
E-21	H-8	85	1	6	38	0	0	



However, ecological data are not disparate, just different traditions of expressing data

A conceptual schema for (field) ecological data



The Darwin core “standard” and extensions extensively used for sharing ecological data



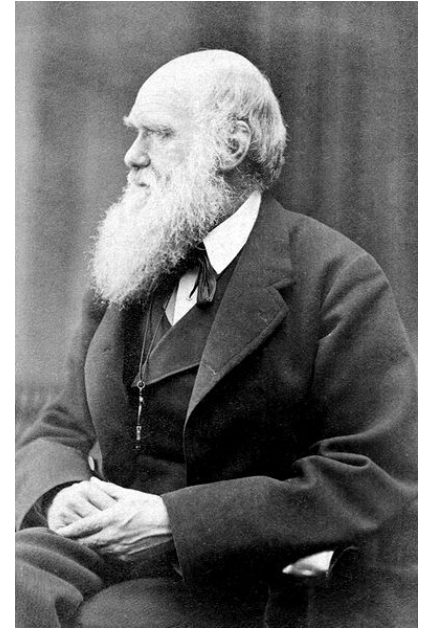
What is Darwin Core?

Broadly definition: A body of standard practices to facilitate biodiversity data sharing

Narrow definition: A library and definitions of common biodiversity data terms (ca. 200 of them)

Maintained by

Biodiversity
Information
Standards
T D W G



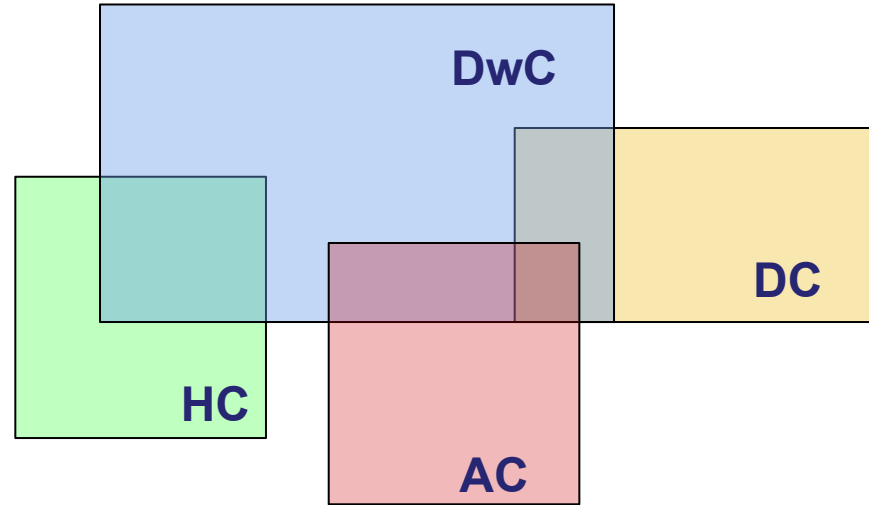
Some online resources: [The Darwin Core Standard](#), [History of DwC and biodiversity informatics in general](#), [Quick reference guide \(the DwC “human-friendly” access point\)](#), [Darwin Core Questions & Answers site](#), [Darwin Core Hour Webinars](#), [academic reference](#)

DwC builds upon and share terms with other standards (examples)

Dublin Core (DC): Describing resources in general (e.g. type, modified)

Audubon Core (AC): Terms describing biodiversity media resources

Humbolt Core: Terms for biodiversity inventory



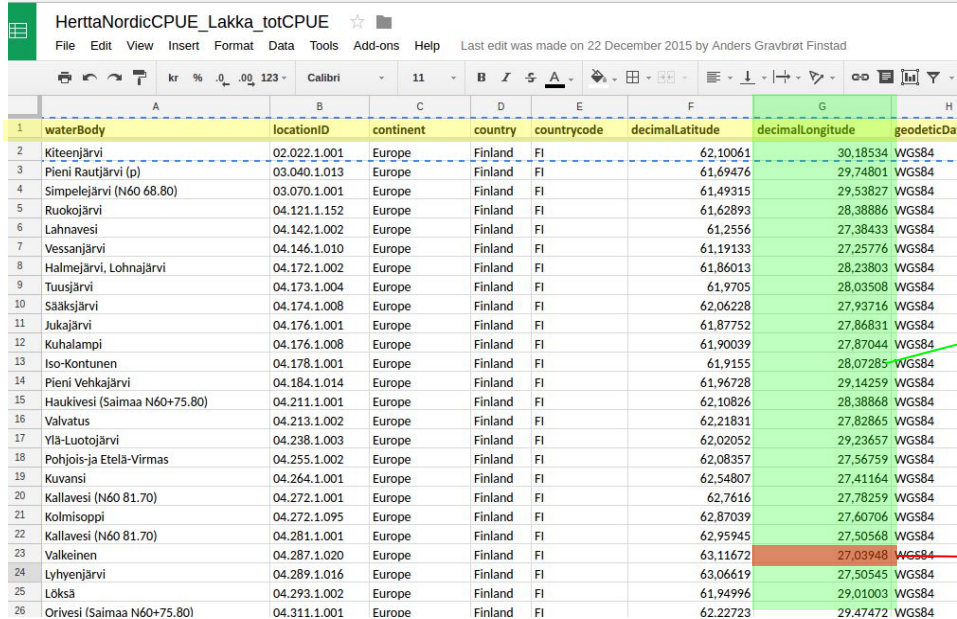
Example: describing an occurrence

occurrenceID: "urn:uuid:e122e87f-3207-458c-9f21-c5231960c674"
recordNumber: "12244"
organismID: "orcid.org/0000-0003-4529-6266"
organismName: "Anders Gravbrøt Finstad"
scientificName: "Homo sapiens Linnaeus, 1758"
occurrenceStatus: "present"

organismQuantity: 101
organismQuantityType: "biomassKg"
establishmentMeans: "invasive"

recordedBy: "Zaphod Beeblebrox"
identifiedBy "Ford Prefect"

DwC and the data?



1	waterBody	locationID	continent	country	countrycode	decimalLatitude	decimalLongitude	geodeticDa
2	Kiteenjärvi	02.022.1.001	Europe	Finland	FI	62.10061	30.18534	WGS84
3	Pieni Rautjärvi (p)	03.040.1.013	Europe	Finland	FI	61.69476	29.74801	WGS84
4	Simpelejärvi (N60 68.80)	03.070.1.001	Europe	Finland	FI	61.49315	29.53827	WGS84
5	Ruokojärvi	04.121.1.152	Europe	Finland	FI	61.62893	28.38886	WGS84
6	Lahnavesi	04.142.1.002	Europe	Finland	FI	61.2556	27.38433	WGS84
7	Vessanjärvi	04.146.1.010	Europe	Finland	FI	61.19133	27.25776	WGS84
8	Halmejärvi, Lohnejärvi	04.172.1.002	Europe	Finland	FI	61.86013	28.23803	WGS84
9	Tuusjärvi	04.173.1.004	Europe	Finland	FI	61.9705	28.03508	WGS84
10	Sääksjärvi	04.174.1.008	Europe	Finland	FI	62.06228	27.93716	WGS84
11	Jukajärvi	04.176.1.001	Europe	Finland	FI	61.87752	27.86831	WGS84
12	Kuhalampi	04.176.1.008	Europe	Finland	FI	61.90039	27.87044	WGS84
13	Iso-Kontunen	04.178.1.001	Europe	Finland	FI	61.9155	28.07285	WGS84
14	Pieni Vehkajärvi	04.184.1.014	Europe	Finland	FI	61.96728	29.14259	WGS84
15	Haukivesi (Saimaa N60+75.80)	04.211.1.001	Europe	Finland	FI	62.10826	28.38868	WGS84
16	Valvatus	04.213.1.002	Europe	Finland	FI	62.21831	27.82865	WGS84
17	Ylä-Luotojärvi	04.238.1.003	Europe	Finland	FI	62.02052	29.23657	WGS84
18	Pohjois-ja Etelä-Virmas	04.255.1.002	Europe	Finland	FI	62.08357	27.56759	WGS84
19	Kuvansi	04.264.1.001	Europe	Finland	FI	62.54807	27.41164	WGS84
20	Kallavesi (N60 81.70)	04.272.1.001	Europe	Finland	FI	62.7616	27.78259	WGS84
21	Kolmisoppi	04.272.1.095	Europe	Finland	FI	62.87039	27.60706	WGS84
22	Kallavesi (N60 81.70)	04.281.1.001	Europe	Finland	FI	62.95945	27.50568	WGS84
23	Valkeinen	04.287.1.020	Europe	Finland	FI	63.11672	27.03948	WGS84
24	Lyhyenjärvi	04.289.1.016	Europe	Finland	FI	63.06619	27.50545	WGS84
25	Löksä	04.293.1.002	Europe	Finland	FI	61.94996	29.01003	WGS84
26	Orivesi (Saimaa N60+75.80)	04.311.1.001	Europe	Finland	FI	62.22723	29.47472	WGS84

A dataset consists of a set of **records** (a.k.a rows for tabular representation)

Data are organized in **fields** (a.k.a columns for tabular representation). Content described by standardized terms

Fields are populated by **values** that may (or may not) be restricted by controlled vocabularies or datatypes

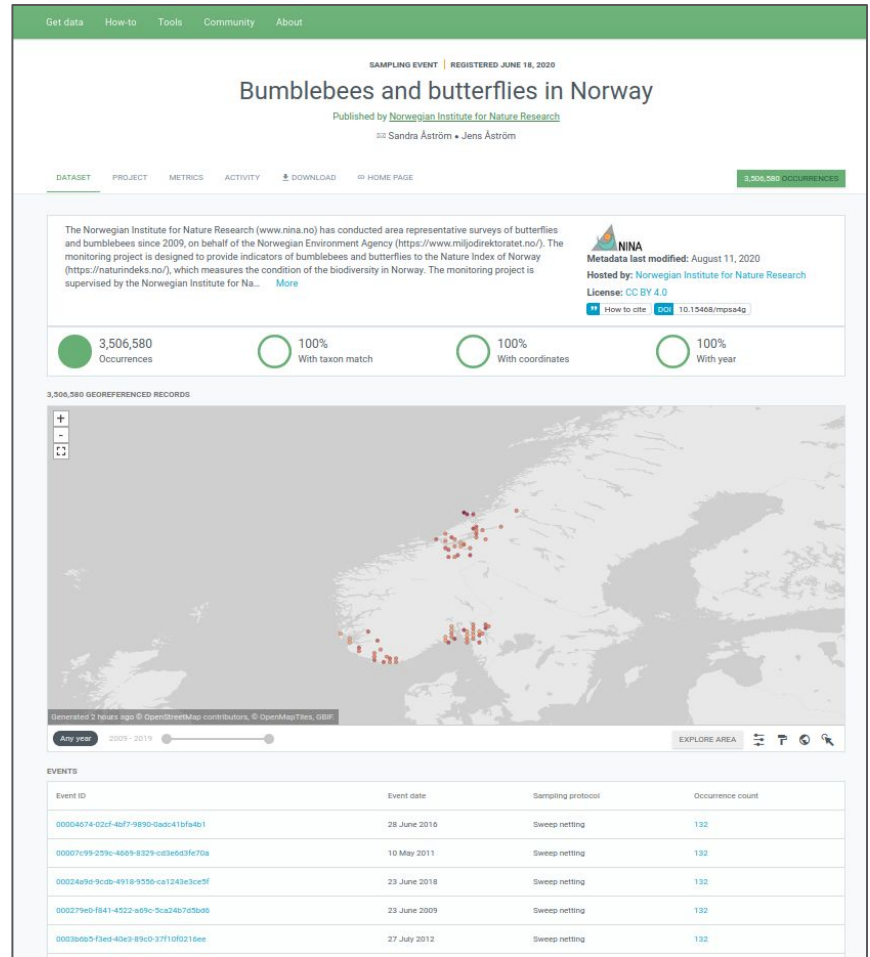
Examples of ecological data types documented
using Darwin Core and extensions



Species inventories



Amateur Entomologists' Society (<http://www.amentsoc.org>), CC-BY 2.0

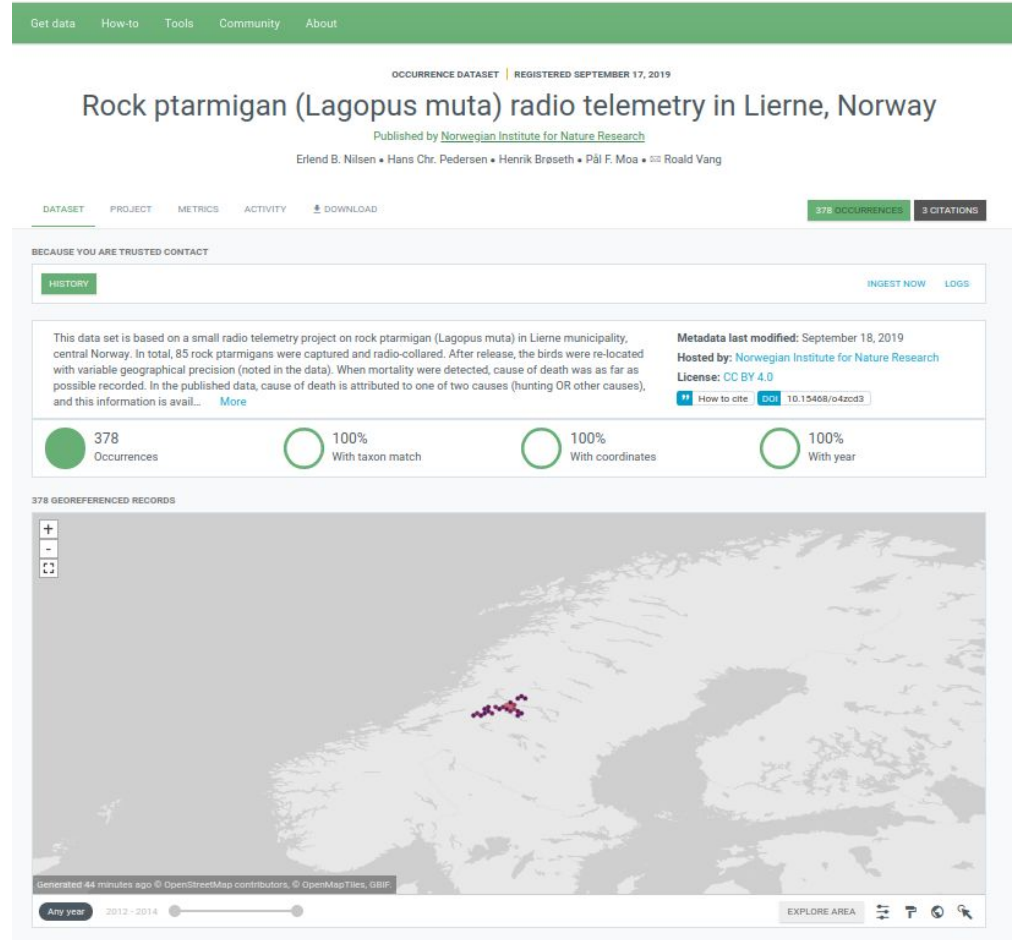


<https://www.gbif.org/dataset/aea17af8-5578-4b04-b5d3-7adf0c5a1e60>

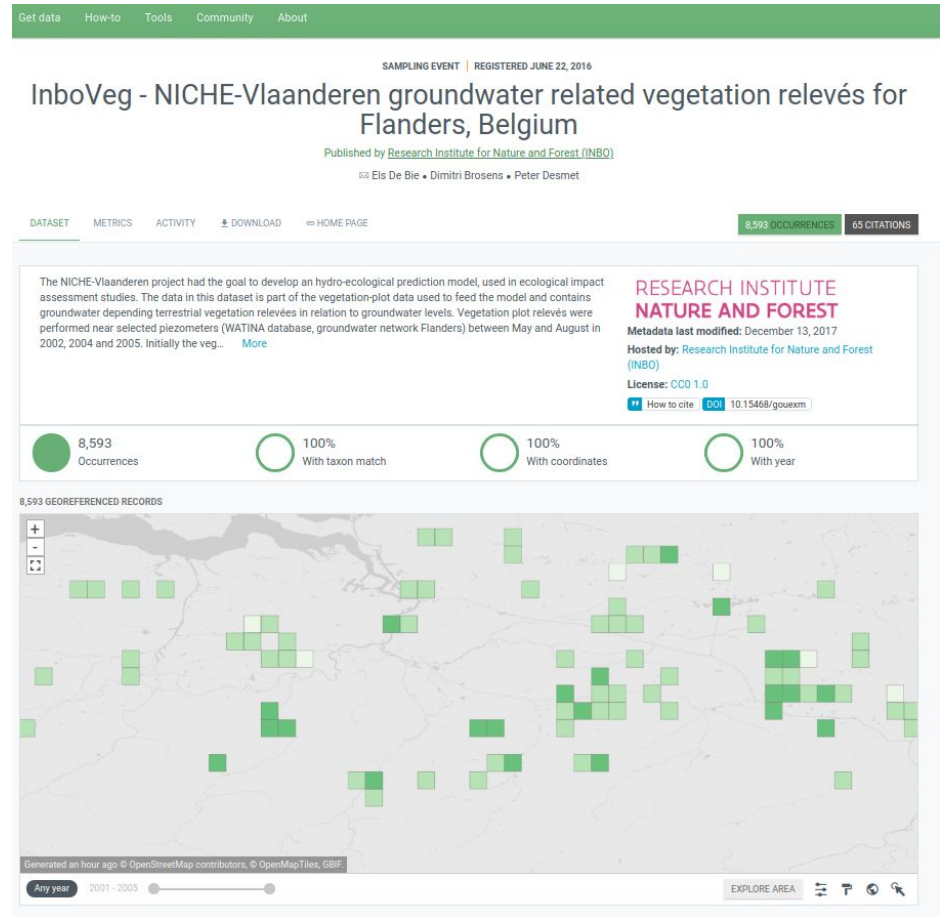
Panel data



Foto: Erlend B. Nilsen



Vegetation plots



Line transects

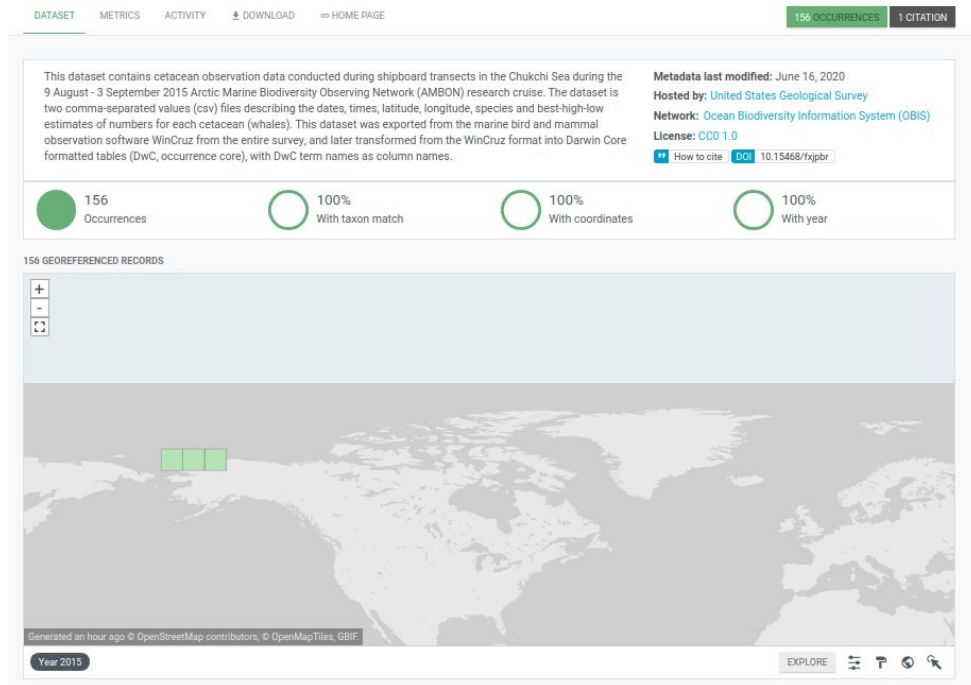


Foto: John Seys CC-BY-NC-ND

Vessel line-transect surveys of Arctic cetaceans in the Chuckchi Sea, Arctic Marine Biodiversity Observing Network (AMBON) research cruise on the vessel Norseman II, 2015-08-09 - 2015-09-03

Published by [United States Geological Survey](#)

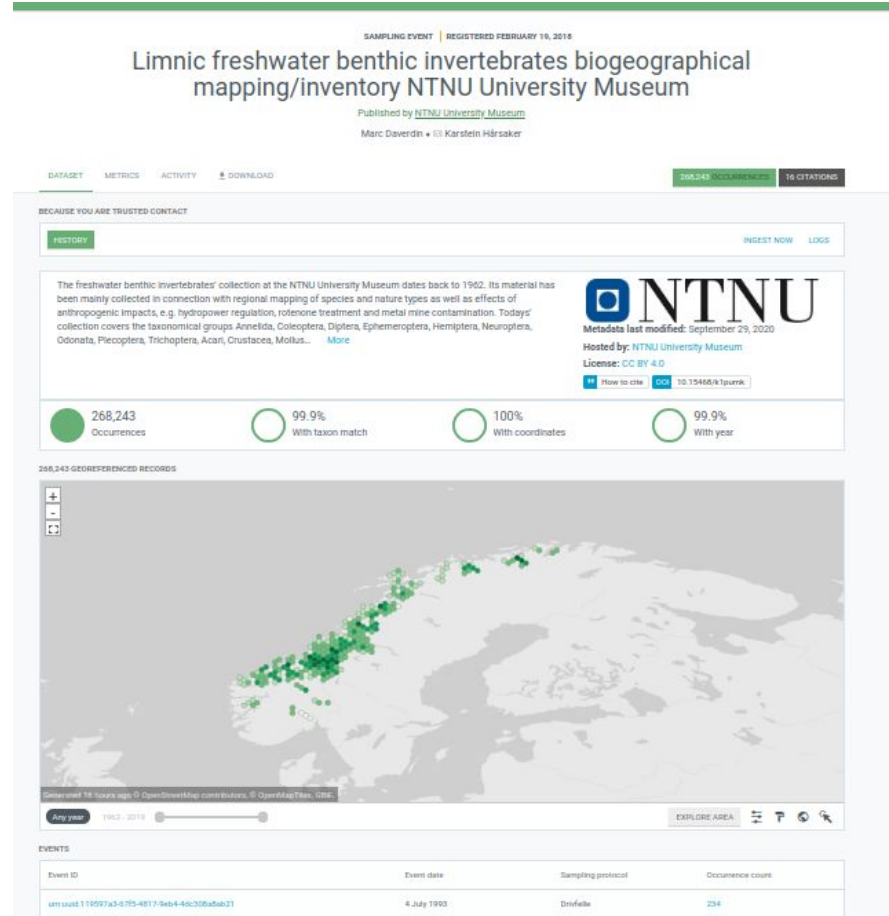
Kate Stafford • Sue Moore • Katrin Iken • Chris Turner



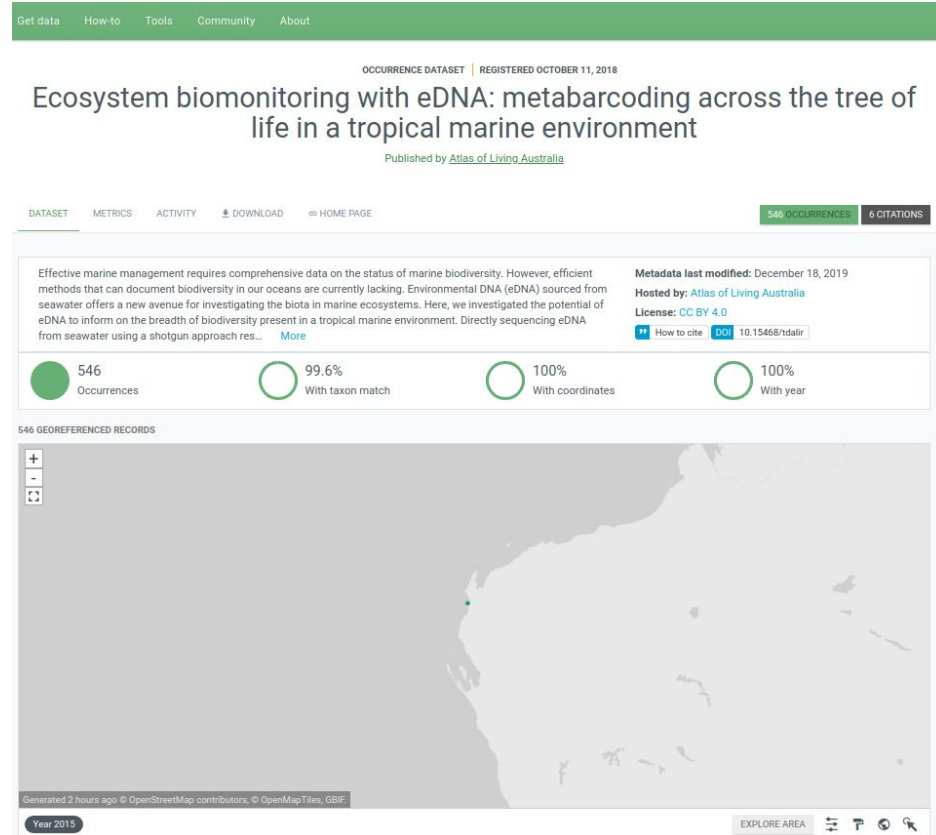
Quantitative monitoring



Foto: Kjell A. Aune



eDNA based inventories

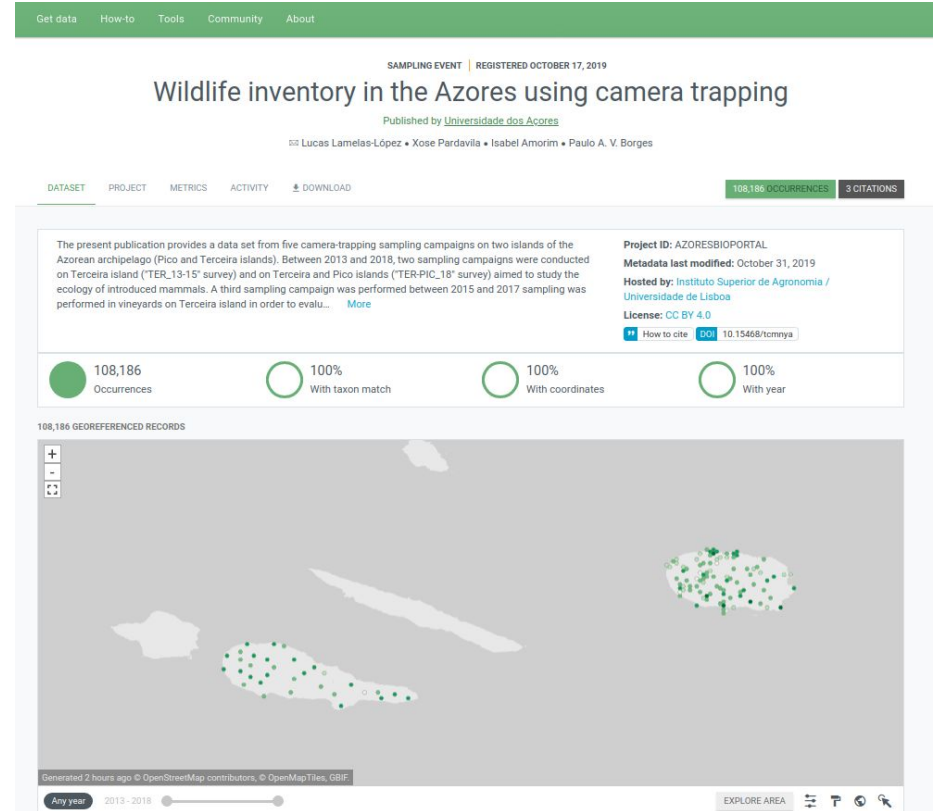


<https://www.gbif.org/dataset/f8ceb4e6-82ff-4325-afc8-eb5e64b5f842>

Camera traps



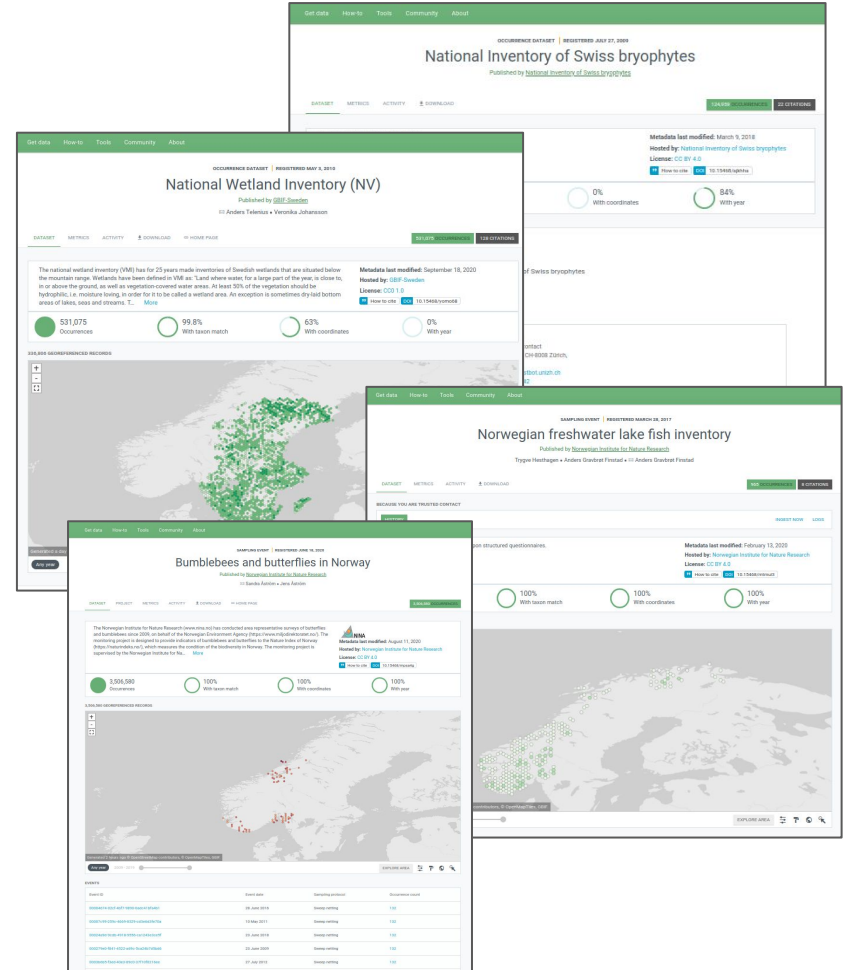
Illustration from Lamelas-López et al. (2020) <https://doi.org/10.3897/BDJ.8.e4>



<https://www.gbif.org/dataset/7d6b90d2-14c0-4ba6-9e45-449b56bab878>

Challenges

- Heterogeneity in application how DwC are used to code ecological data



Challenges

- Heterogeneity in application how DwC are used to code ecological data
- More coherent data model (coming ...?)
- Need of (more) community derived recommendations and guides to publishing and re-use of ecological data

 **GBIF** Global Biodiversity Information Facility

Other Formats

[PDF file](#)

Contribute

[Improve this document](#)

[Create an issue](#)

[Edit on GitHub](#)

Table of Contents

Colophon

[Suggested citation](#)

[Authors](#)

[Contributors](#)

[License](#)

[Persistent URI](#)

[Document control](#)

[Abstract](#)

[Cover image](#)

Preface

1. Introduction

1.1. Rationale

1.2. Target audiences

1.3. Introduction to sequence-derived occurrence data

1.4. Introduction to biodiversity publishing

1.5. Processing workflows: from sample to ingestible data

1.6. Taxonomy of sequences

1.7. Outputs

2. Data packaging and mapping

2.1. Categorization of your data

2.1.1. Category I: Sequence-

Publishing sequence-derived data through biodiversity data platforms [Community review draft]

Anders F. Andersson · Andrew Bissett · Anders G. Finstad · Frode Fossey · Marie Grosjean · Michael Hope · Urmaz Köljalg · Daniel Lundin · R. Henrik Nilsson · Maria Prager · Thomas Stjernegaard · Cecilie Svenningsen · Dmitry Schigel – Version 2a2a602, 2020-10-06 08:28:24 UTC

This document is also available in [PDF format](#).



Data sharing: the **when** choices

SHORT TERM

LONG TERM



**RESEARCH
PHASE**

- file formats
- ownership
- metadata
- storage
- backups

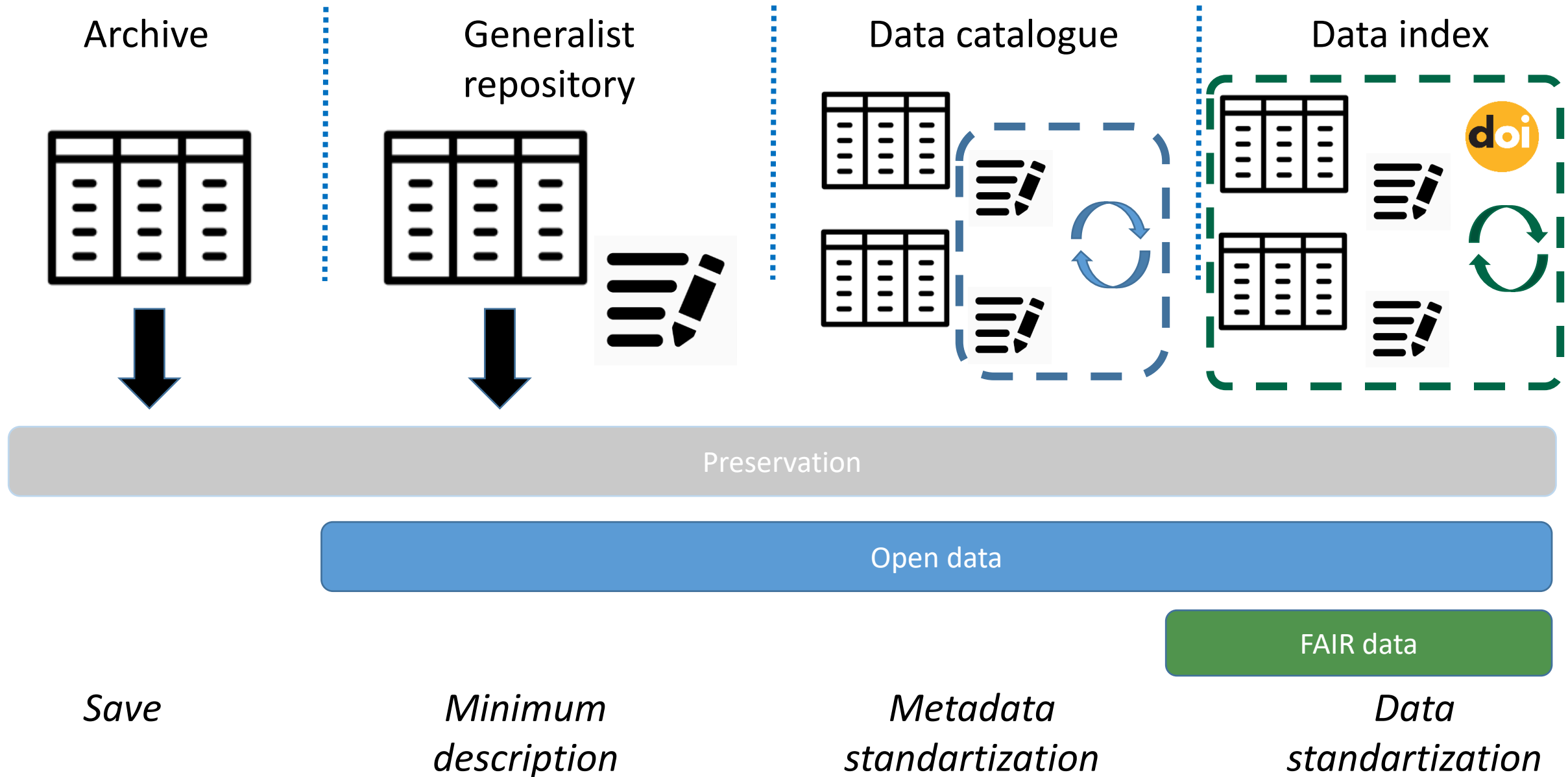
**DISSEMINATION
PHASE**

- share with whom?
- embargo?
- licensing
- metadata

**PRESERVATION
PHASE**

- repository?
- long-term manager?

Data sharing: the **where** choices



How to choose a data repository



(Roche et al 2014 PLOS Biol)



MARIE CURIE ACTIONS

Dominique Roche

@dom_roche

unine
UNIVERSITÉ DE
NEUCHÂTEL



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. McPeck, 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>

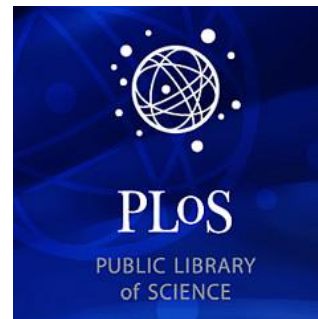
Moore, A. J., M. A. McPeck, M. D. Rausher, L. Rieseberg, and M. C. Whitlock. 2010. The need for archiving data in evolutionary biology. *Journal of Evolutionary Biology*. <https://doi.org/10.1111/j.1420-9101.2010.01937.x>

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

...





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

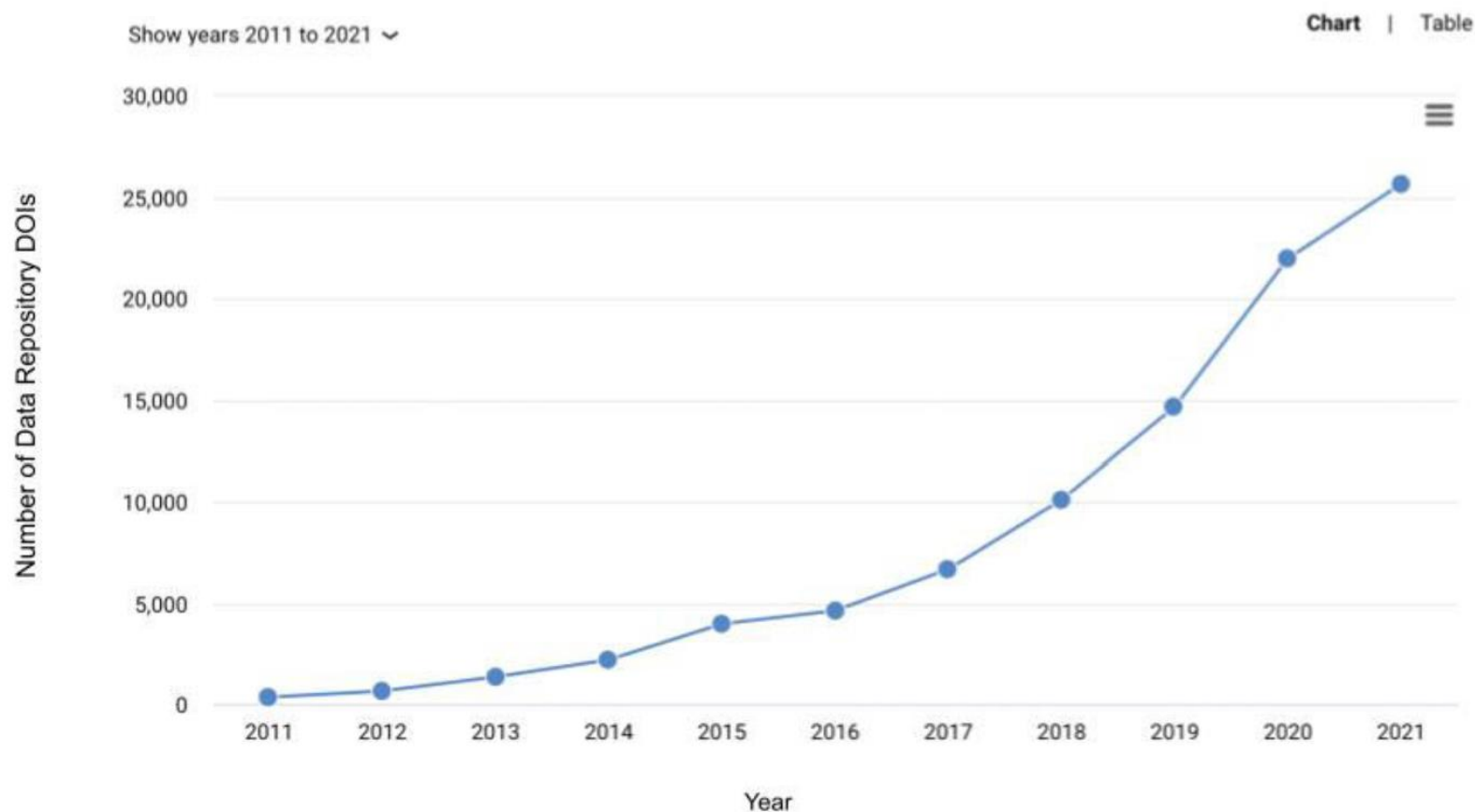
74 (37%) recommend open data

55 (28%) have no data policy

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



2013

Data reuse and the open data citation advantage

Heather A. Piwowar^{1,2} and Todd J. Vision^{1,2,3}



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*}



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

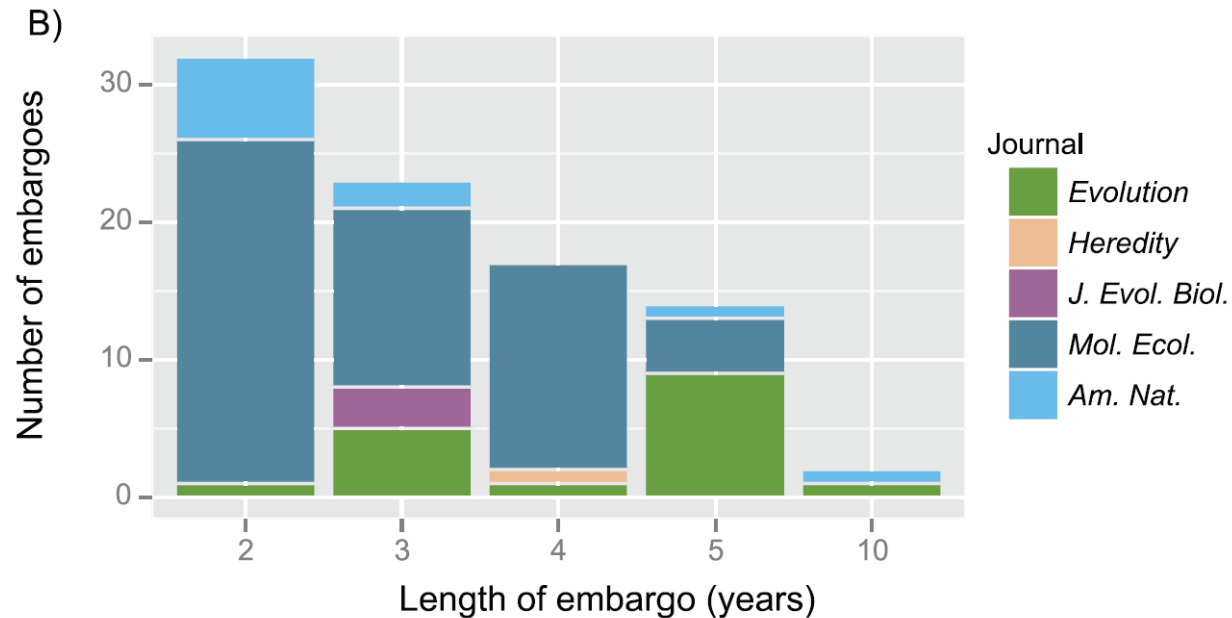
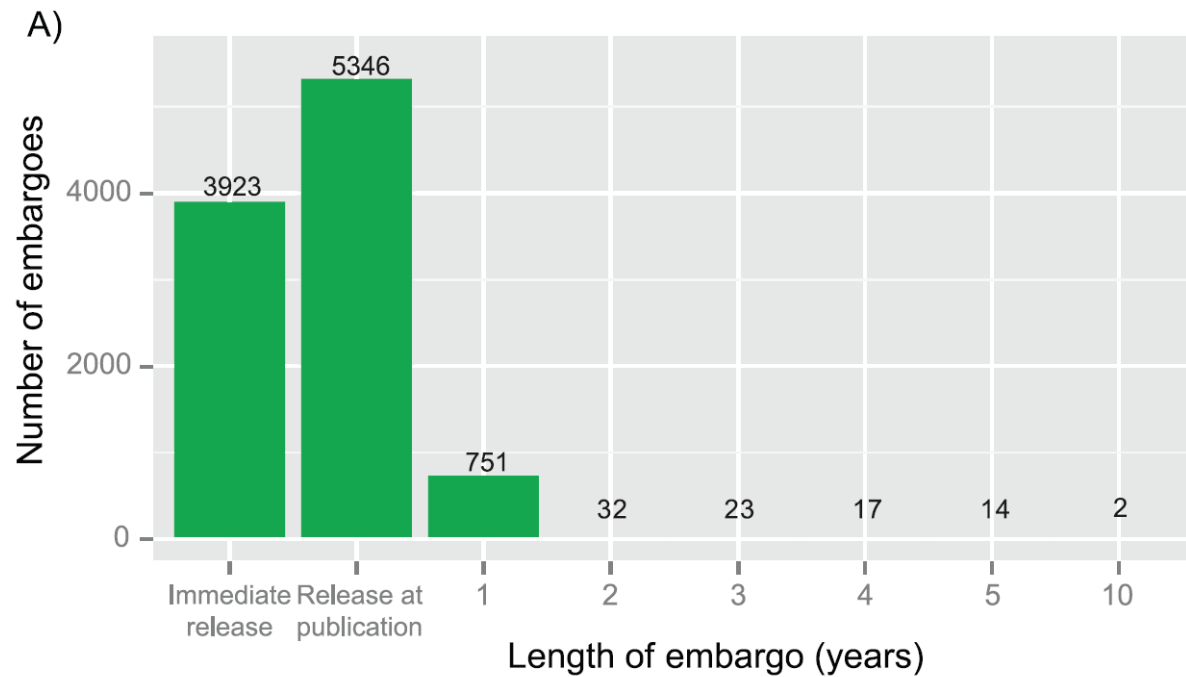


Embargoes



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



Hybrid

- Dataverse & Figshare for institutions



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



Repository Finder

Find a repository to upload your data.

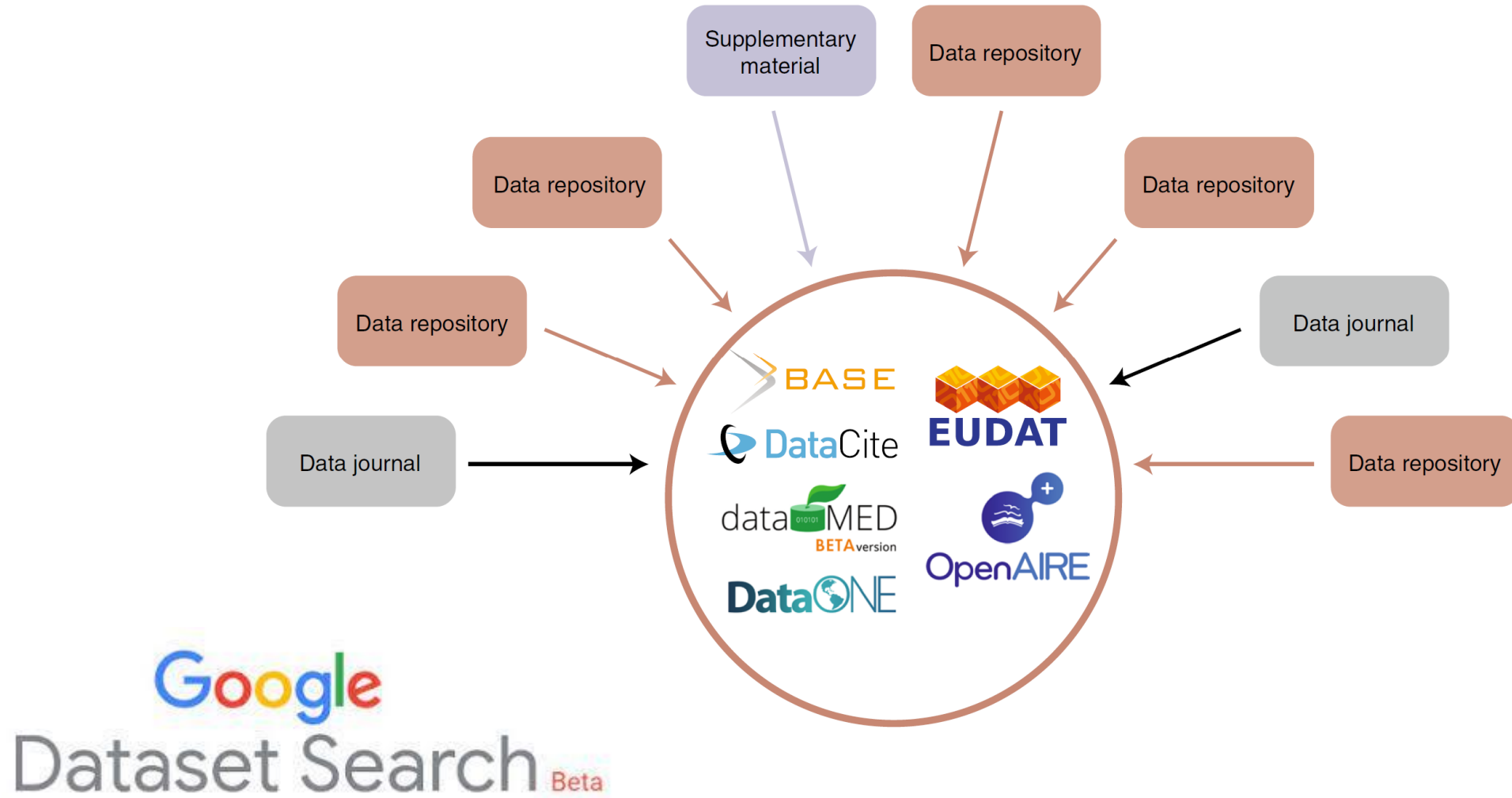
re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

 Search

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



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?

OPEN COMMENT The TRUST Principles for digital repositories

Dawei Lin¹✉, Jonathan Crabtree², Ingrid Dillo³, Robert R. Downs⁴, Rorie Edmunds⁵, David Giarretta⁶, Marisa De Giusti⁷, Hervé L'Hours⁸, Wim Hugo⁹, Reyna Jenkyns¹⁰, Varsha Khodiyar¹¹, Maryann E. Martone¹², Mustapha Mokrane³, Vivek Navale¹³, Jonathan Petters¹⁴, Barbara Sierman¹⁵, Dina V. Sokolova¹⁶, Martina Stockhause¹⁷ & John Westbrook¹⁸

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.



www.coretrustseal.org

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 Baglioni², Tom W. Crowther^{1,3}, Marcel E. Visser ¹,
Saskia Woutersen-Windhouwer¹ 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