re3data coref

Research data quality assurance as a research topic in re3data COREF

Maxi Kindling[1] (maxi.kindling@hu-berlin.de)
0000-0002-0167-0466
Dorothea Strecker[1]
0000-0002-9754-3807

[1]Berlin School of Library and Information Science, Humboldt-Universität zu Berlin

Berliner Bibliothekswissenschaftliches Kolloquium, 30 June 2020

DataCite FIND, ACCESS, AND REUSE DATA
HELMHOLTZ Open Science
KIT Karlsruher Institut für Technologie
DFG

Except for logos, this presentation is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License.

Outline

re3data coref

1. re3data.org
2. re3data COREF
3. Concepts of Research Data Quality and Research Data Quality Assurance
4. Research in re3data COREF
5. Outlook

Mission

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

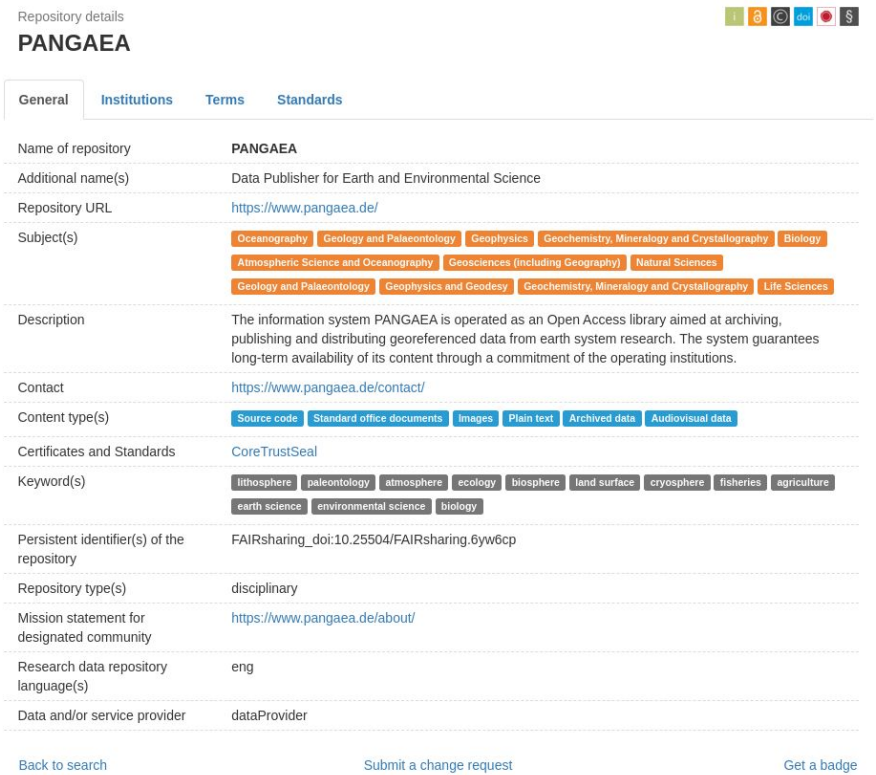
- Global registry of research data repositories (2,500+ repositories registered to date)
- Covers all academic disciplines
- Presents repositories and portals for the permanent storage and access of research data sets to researchers, funding bodies, publishers and scholarly institutions.
- Promotes a culture of sharing, increased access and better visibility of research data

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Metadata

41 properties on

- General information
- Responsibilities
- Policies
- Legal aspects
- Technical standards
- Quality standards
- Icon system



Repository details

PANGAEA

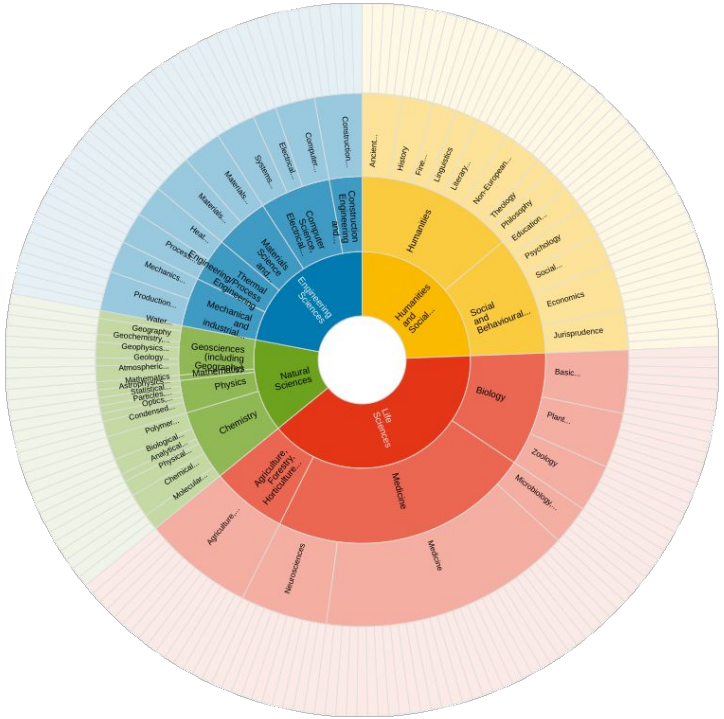
General | Institutions | Terms | Standards

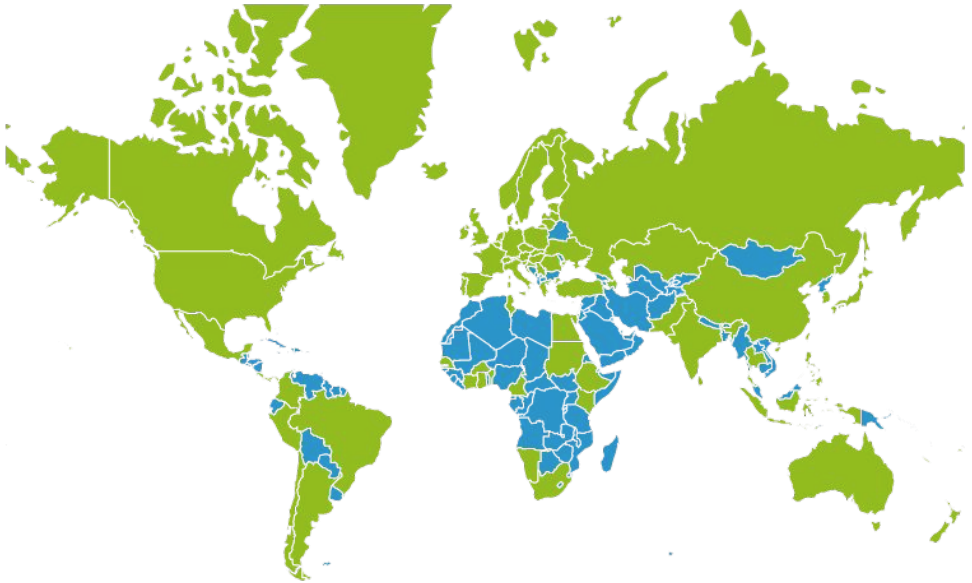
Name of repository	PANGAEA
Additional name(s)	Data Publisher for Earth and Environmental Science
Repository URL	https://www.pangaea.de/
Subject(s)	Oceanography; Geology and Palaeontology; Geophysics; Geochemistry, Mineralogy and Crystallography; Biology; Atmospheric Science and Oceanography; Geosciences (including Geography); Natural Sciences; Geology and Palaeontology; Geophysics and Geodesy; Geochemistry, Mineralogy and Crystallography; Life Sciences
Description	The information system PANGAEA is operated as an Open Access library aimed at archiving, publishing and distributing georeferenced data from earth system research. The system guarantees long-term availability of its content through a commitment of the operating institutions.
Contact	https://www.pangaea.de/contact/
Content type(s)	Source code; Standard office documents; Images; Plain text; Archived data; Audiovisual data
Certificates and Standards	CoreTrustSeal
Keyword(s)	lithosphere; paleontology; atmosphere; ecology; biosphere; land surface; cryosphere; fisheries; agriculture; earth science; environmental science; biology
Persistent identifier(s) of the repository	FAIRsharing_doi:10.25504/FAIRsharing.6yw6cp
Repository type(s)	disciplinary
Mission statement for designated community	https://www.pangaea.de/about/
Research data repository language(s)	eng
Data and/or service provider	dataProvider

Back to search | Submit a change request | Get a badge

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Browsing





re3data COREF - Objectives

- Improve the (re-)use of re3data, its API and metadata.
- Extend re3data to be a reference for research data repositories by setting up a re3data PID.
- Extend the scope of re3data by addressing research funding bodies and publishers.
- Update the re3data service model and metadata schema to current needs.

Key Facts

- Starting date - 01/01/2020
- Duration - 36 months
- Funded by the German Research Foundation

WP Metadata Schema

- Revision of the Metadata Schema and Update
  - Revision of repositoryType
  - Revision of qualityManagement
  - FAIR alignment
  - [...]
- Open and transparent request for comments process
- Mapping to schema.org
- RDF implementation

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Metadata Schema for the Description of Research Data Repositories

Version 3.0
December 2015
doi: http://doi.org/10.2312/re3.008

Authors: Jessika Rücknagel [b], Paul Vierkant [a], Robert Ulrich [c], Gabriele Kloska [c], Edeltraud Schnepf [c], David Fichtmüller [d], Evelyn Reuter [c], Angelika Semrau [c], Maxi Kindling [b], Heinz Pampel [a], Michael Witt [e], Florian Fritze [b], Stephanie van de Sandt [b], Jens Klump [f], Hans-Jürgen Goebelbecker [c], Michael Skarupianski [c], Roland Bertelmann [a], Peter Schirmbacher [b], Frank Scholze [c], Claudia Kramer [c], Claudio Fuchs [a], Shaked Spier [b], Agnes Kirchhoff [d]

[a] GFZ German Research Centre for Geosciences, Library and Information Services (LIS), Germany
[b] Humboldt-Universität zu Berlin, Berlin School of Library and Information Science (BLIS), Germany
[c] Karlsruhe Institute of Technology (KIT), KIT Library, Germany
[d] Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin, Germany
[e] Purdue University Libraries, West Lafayette, USA
[f] Commonwealth Scientific and Industrial Research Organisation (CSIRO), Mineral Resources, Kensington WA, Australia

Contact
info@re3data.org
http://www.re3data.org

Except where otherwise noted, this work is licensed under http://creativecommons.org/publicdomain/zero/1.0/

WP Quality & Transparency

- Study on status quo of research data quality assurance measures and standards
- Update of metadata schema to provide information on research data repository quality assurance standards and measures
- Implement the use of authority files (ORCID)
- Review of the re3data editorial process
- Build model of trust for metadata editing

Research Data Quality (RDQ)

Working definition

- No formal, generalizable definition, but many theoretical and practical approaches (very often referring to (information) management in business)
- Pragmatic approach:
  - Set of properties of (digital) research data needed for scientific work (data generation, processing, provision and (re-)use etc.)
  - Properties to be determined and operationalised in a transparent and consistent manner
  - Agreement on explicit criteria within the scientific community, a research project etc. that
  - To assess/control/evaluate criteria often data documentation is the medium of choice

Cf. Kindling, 2013, p. S. 140ff.; RfII, 2017, S. A-8f.

Research Data Quality (RDQ)

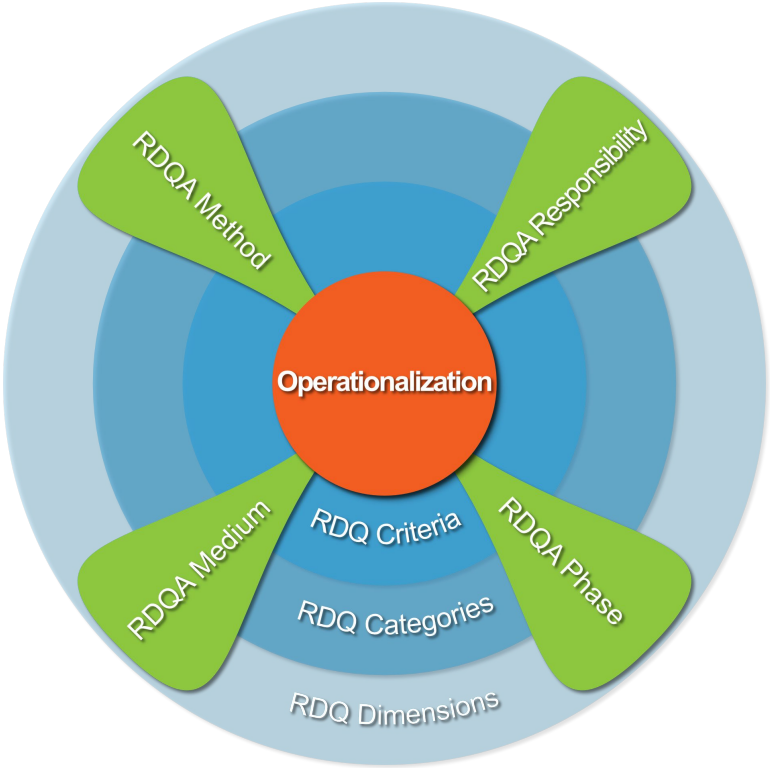
Quality Dimension	Category	Example Criteria
Research specific	Context	Innovativeness
Technical	Representation	Open data format
Legal	Access	Availability of data license
Ethical	Intrinsic	Anonymization
Methodological	Context	Suitability of methods
Organizational	Context	Provenance metadata
Communicational	Preservation	RDR deposit

Research Data Quality Assurance (RDQA)

Source: Kindling, Maxi, 2020, Preliminary extract; PhD thesis (work in progress) at BSLIS, supervised by Prof. Dr. Peter Schirmbacher (i.R.))

Research Data Quality Framework (RDQF)

re3data coref



Source: Kindling, Maxi, 2020, Preliminary extract; PhD thesis (work in progress) at BSLIS, supervised by Prof. Dr. Peter Schirmbacher (i.R.))

RDQ(A) in re3data COREF

re3data coref

Research interests

Beyond Open and FAIR:

- How can we contribute to the visibility of RDQA?
- What can we do to strengthen RDQA?
- What are best practices of RDQA?
- What can we learn from a multidisciplinary perspective?

RDQ(A) in re3data COREF

Repositories

RDQ criteria in practice
RDQA of repositories

"Formal (standards) and informal community requirements and their implementation on the level of repositories; e.g. CoreTrustSeal, FAIR Data Principles [...]"

- RDQA and certification processes of RDR
- Certificates, Standards, (Community) Requirements
- Data journal repository recommendations
- CTS self assessment documents

Research data

RDQ criteria in practice
RDQA of data sets, data collections

"Technical and scientific as well as interdisciplinary requirements and their implementation on the level of data files and data collections; e.g. peer review [...]"

- Data journal author and review guidelines
- Data journal review methods
- RDQ criteria of repositories
- RDQA methods of repositories
- CTS self assessment documents

Metadata

RDQ criteria in practice
RDR metadata in re3data

"[...] revision of metadata element "Quality Management"
→ "Improve the visibility of RDAQ"

- Metadata standards in use
- Metadata requirements of repositories
- **RDQ(A) metadata in re3data**

Analysis of Data Journal Guidelines

Why Data Journals?
They publish Data Papers:

- Citable, peer-reviewed publications describing reusable data sets and / or data sets that are of significance for the field.

Analysis of Data Journal Guidelines

Sampling and data collection

- Data Journals mentioned in various sources (see Zotero Library)
- Exclusion of discontinued journals
- Exclusion of journals that do not mention 'Data Paper' or similar article types in their scope or submission guidelines

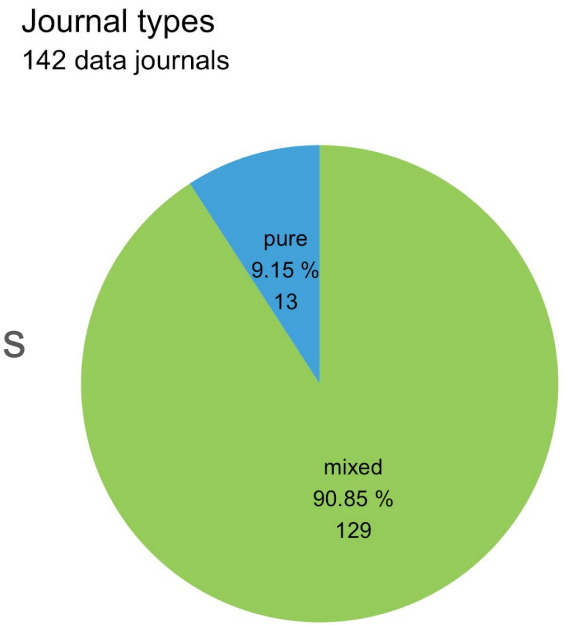
Result: sample of 142 Data Journals.

- Adding metadata via crossref API
- Manually adding information from journal website

Analysis of Data Journal Guidelines

Preliminary results: Data Journal type

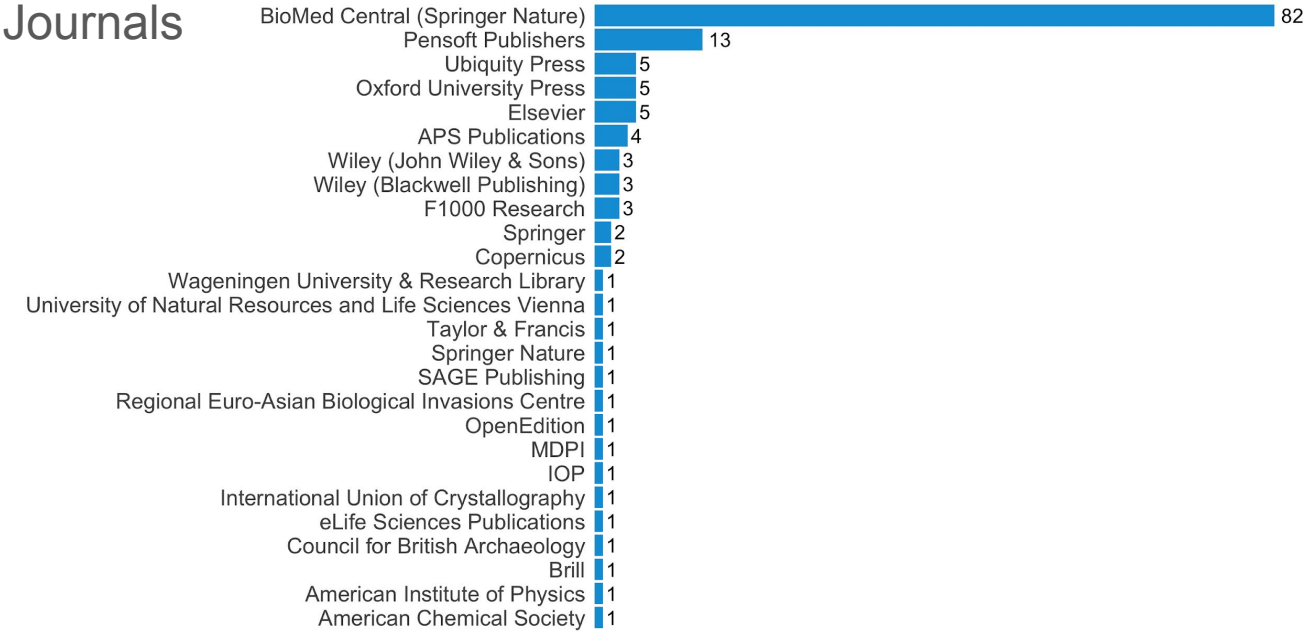
- 13 (9.15 %) journals are 'pure' Data Journals (exclusively publishing Data Papers or similar publication types)
- Comparison:
  - Candela et al. (2015) found 7 'pure' Data Journals (6 % of 116 Journals)
  - Schöpfel (2019) found 28 'pure' Data Journals (34 % of 82 Journals)



Analysis of Data Journal Guidelines

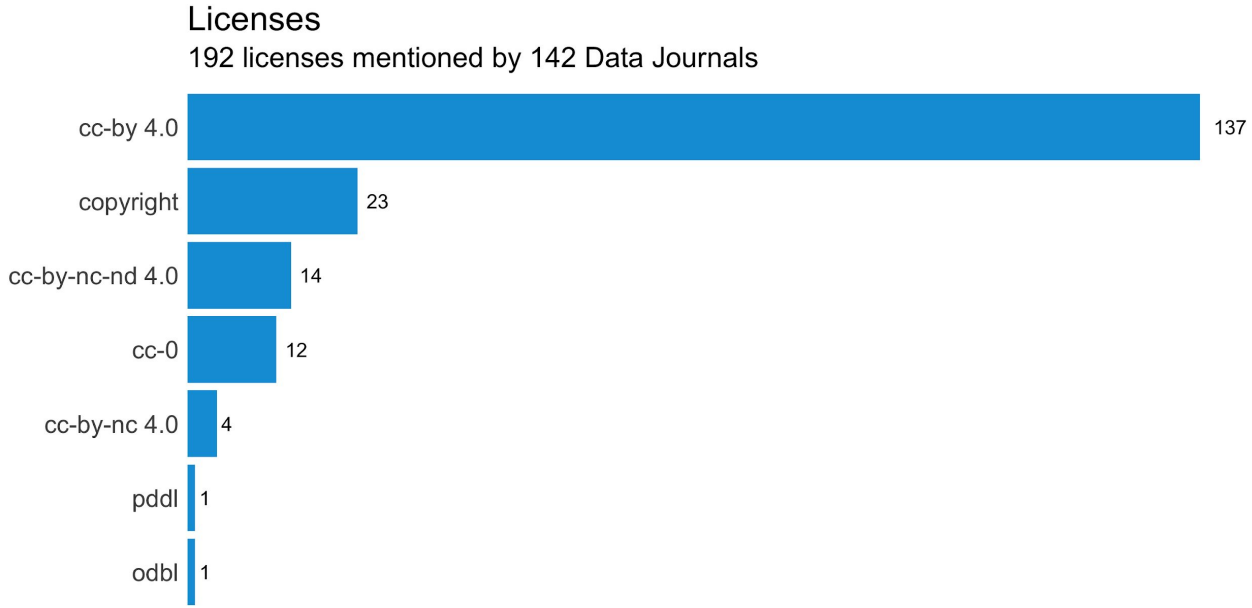
Preliminary results: Publishers

142 Data Journals



Analysis of Data Journal Guidelines

Preliminary results: Licenses



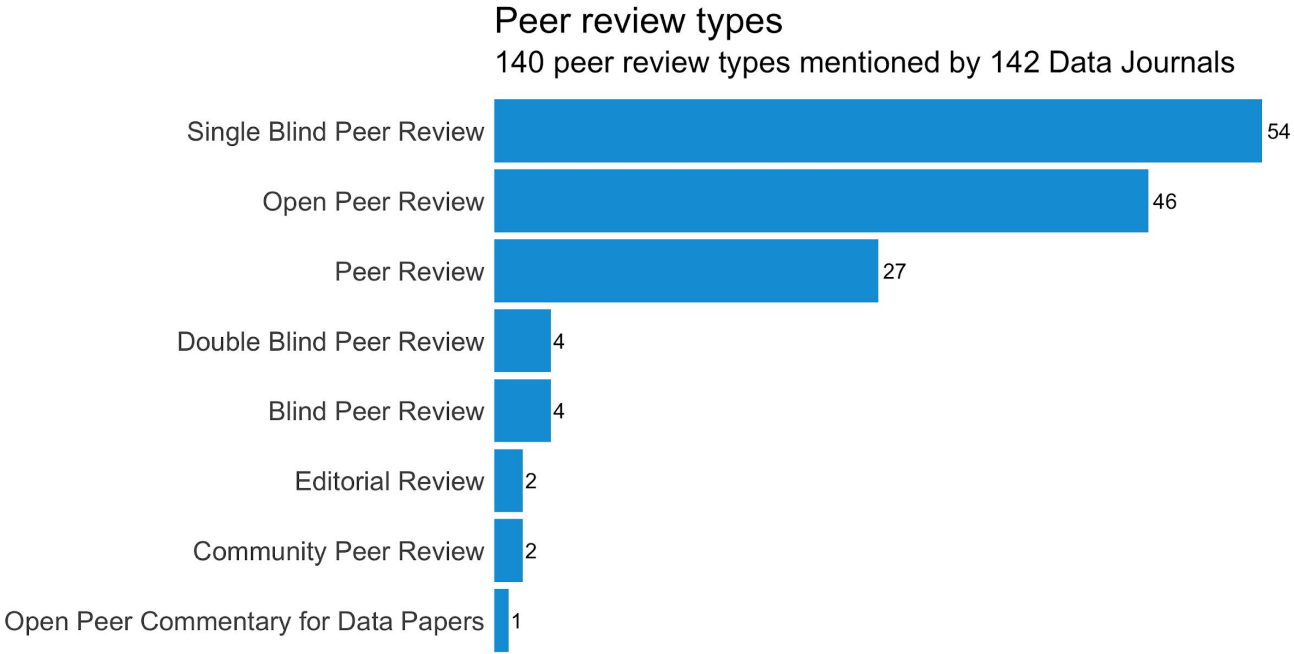
Analysis of Data Journal Guidelines

Preliminary results: Article types

909 article types mentioned by 138 journals

Analysis of Data Journal Guidelines

Preliminary results: Peer review types



Analysis of Data Journal Guidelines

Preliminary results:

- Diverse terms and slightly differing definitions of “data paper”
- Most data journals have a data sharing mandate (at the point of submission, acceptance or publication); some “only” encourage data sharing; all mandates offer opt-out
- Some recommend CC0 for data sets
- Most frequent requirements: long-term availability and PID assignment
- Recommended repositories (especially from biomedical journals) will be added to re3data

More research to come…

Analysis of Data Journal author guidelines

- Sample of 82 guidelines: partial overlap with repository recommendations

Analysis of repository certificates, standards, (community) requirements

- In close contact with stakeholders (CoreTrustSeal (CTS), FAIRsFAIR)

Analysis of CTS self assessment documents

(see right; wordcloud based on section R11 "Data Quality" of all 90 certified repositories)

Outlook

Survey among RDR operators on RDQA at repositories (autumn 2020)
based on results from previous research

Objectives

- Identification of best practices
- Revision of re3data metadata schema to increase visibility of RDQA methods and measures

Thank You!

www.re3data.org

@re3data

doi.org/10.1045/march2017-kindling

Except for logos, this presentation is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License.

References

Please find all references in our public Zotero library