How to ensure "good" data? Quality assurance by research data repositories

Open Repositories 2021, 10 June 2021

Authors:

Maxi Kindling[1] (maxi.kindling@hu-berlin.de) iD 0000-0002-0167-0466, Dorothea Strecker[1] iD 0000-0002-9754-3807, Vivien Petras[1] iD 0000-0002-8113-1509, Yi Wang[1] iD 0000-0003-1354-3461

re3data COREF team:

Gabriele Kloska[2], Heinz Pampel[4], Kirsten Elger[3], Lea Maria Ferguson[4], Michael Witt[5], Margarita Trofimenko[2], Nina Weisweiler[4], Paul Vierkant[6], Robert Ulrich[2], Roland Bertelmann[4], Rouven Schabinger[2], Thanh Binh Nguyen[2]

[1]Berlin School of Library and Information Science, Humboldt-Universität zu Berlin
[2]Karlsruher Institut für Technologie / Karlsruhe Institute of Technology (KIT)
[3]GFZ German Research Centre for Geosciences / Helmholtz-Zentrum Potsdam – Deutsches GeoForschungsZentrum (GFZ)
[4]Helmholtz-Gemeinschaft / Helmholtz Association, Helmholtz Open Science Office
[5]Purdue University, United States
[6]DataCite - International Data Citation Initiative e.V.

Outline

1. re3data and re3data COREF

2. Research data quality as a research topic in re3data COREF

3. Survey: Research data quality management at research data repositories (preliminary)

4. Outlook and conclusion

re3data mission

- Global registry of research data repositories (2,600+ repositories registered to date)

- Covers all academic disciplines

- Presents repositories and portals for the permanent storage of and access to research data sets to researchers, funding bodies, publishers and scholarly institutions.

- Promotes a culture of sharing, increased access and better visibility of research data

- Service launched in 2012 (funded by German Research Foundation)

re3data
coref

number of repositories indexed in re3data by year
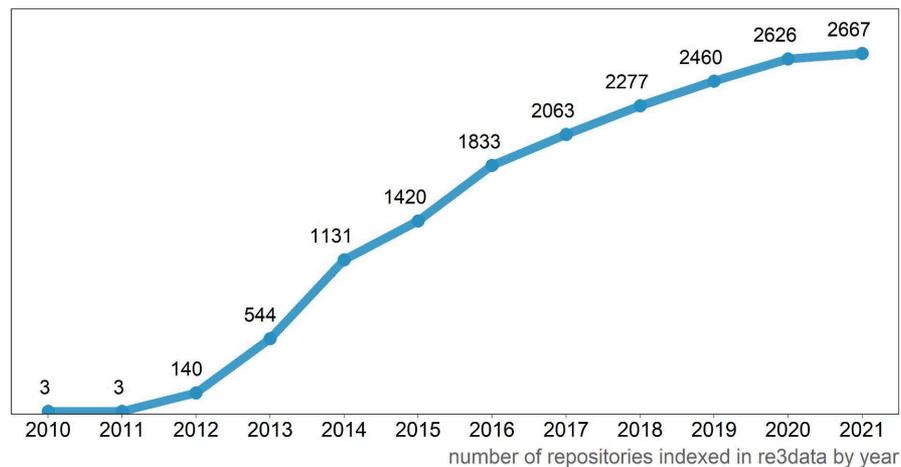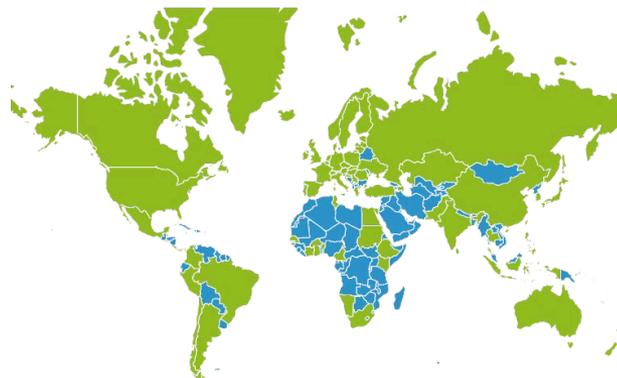
re3data metadata

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 COREF - objective

Increasing re3data's value for all stakeholders, further establishing the service as a tool for finding and referencing research data repositories.

Key Facts

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

Revising the re3data Metadata Schema

Version 3.1

ca. 150 comments in the RFC - Thank you!

Two major changes:

- revision of r3d:certificate
- new property r3d:profile

Next steps:

- Version 4.0
- 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 [c], 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/

Updating the re3data service model

Identify **relevant use cases and requirements** (survey and workshops with 21 stakeholders)

Use cases include:

- search and discovery of research data repositories
- reusing re3data metadata
- administrating re3data metadata
- referencing research data repositories

USE CASES

CONCEPTUAL SERVICE MODEL

GAP ANALYSIS

Making research data quality assurance visible

re3data
coref

making **efforts of ensuring data quality at research data repositories** visible in re3data

- multi-method empirical research into data quality and data quality assurance practices
- revising metadata element r3d:qualityManagement based on the findings

Research design

Data journal guidelines

**Quality definitions
Repository recommendations**

⬇

Metadata collection on data journal characteristics of 142 data journals

Qualitative content analysis of journal guidelines (author guidelines, peer review guidelines, etc)

Core Trust Seal documents

**Quality definitions
Quality management activities at repositories**

⬇

Qualitative content analysis of 99 self-description documents from Core Trust Seal certification 2017—2019

RDR indexed in re3data

**Quality management in practice
Applied quality criteria**

⬇

Online survey conducted from 18 May until 14 June 2021

Survey

Questionnaire
- repository characteristics, formal assessment of data quality, data curation, data review, data evaluation by repository users, and indicators of data quality (Overview)
- 24 questions

Sampling
- Mailing list of 1897 re3data repository contacts

Field phase
- Individualized online survey with Lime Survey
- Started 18 May, deadline extended until 14 June

Survey

Response rate (preliminary)
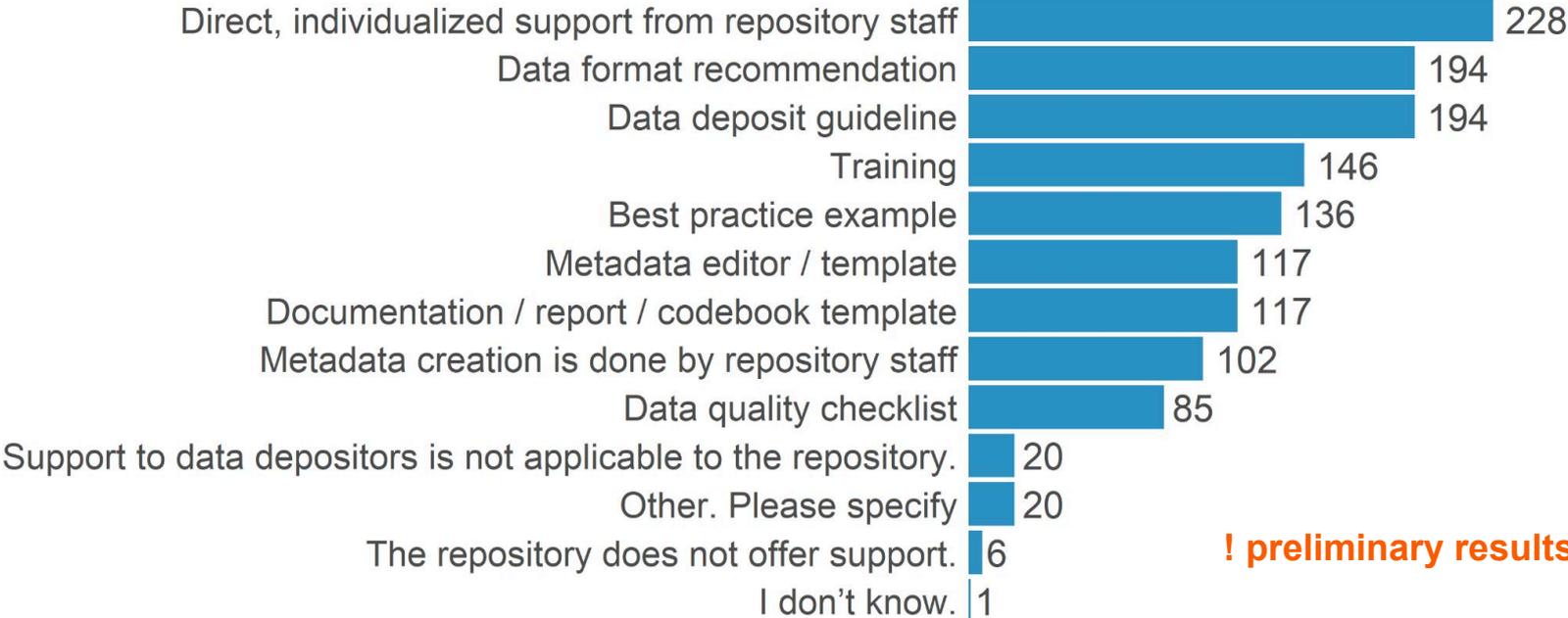
- More than 300 repositories participated so far

Data analysis

- Data will be matched with re3data metadata information
- Results will be shared with the community, without references to single repositories

Survey: Data quality management at research data repositories (preliminary)

re3data coref

What types of support are offered to data depositors?
Choose all that apply.

Type of support	Count
Direct, individualized support from repository staff	228
Data format recommendation	194
Data deposit guideline	194
Training	146
Best practice example	136
Metadata editor / template	117
Documentation / report / codebook template	117
Metadata creation is done by repository staff	102
Data quality checklist	85
Support to data depositors is not applicable to the repository.	20
Other. Please specify	20
The repository does not offer support.	6
I don't know.	1

! preliminary results !

1366 responses from 306 repositories

Survey: Data quality management at research data repositories (preliminary)

re3data
coref

Are formal criteria applied to data before publication?

Yes, formal criteria are applied. — 198

No, formal criteria are not applied. — 57

Formal criteria are not applicable to the repository. — 30

I don't know. — 21

! preliminary results !

306 responses from 306 repositories

Survey: Data quality management at research data repositories (preliminary)

re3data
coref
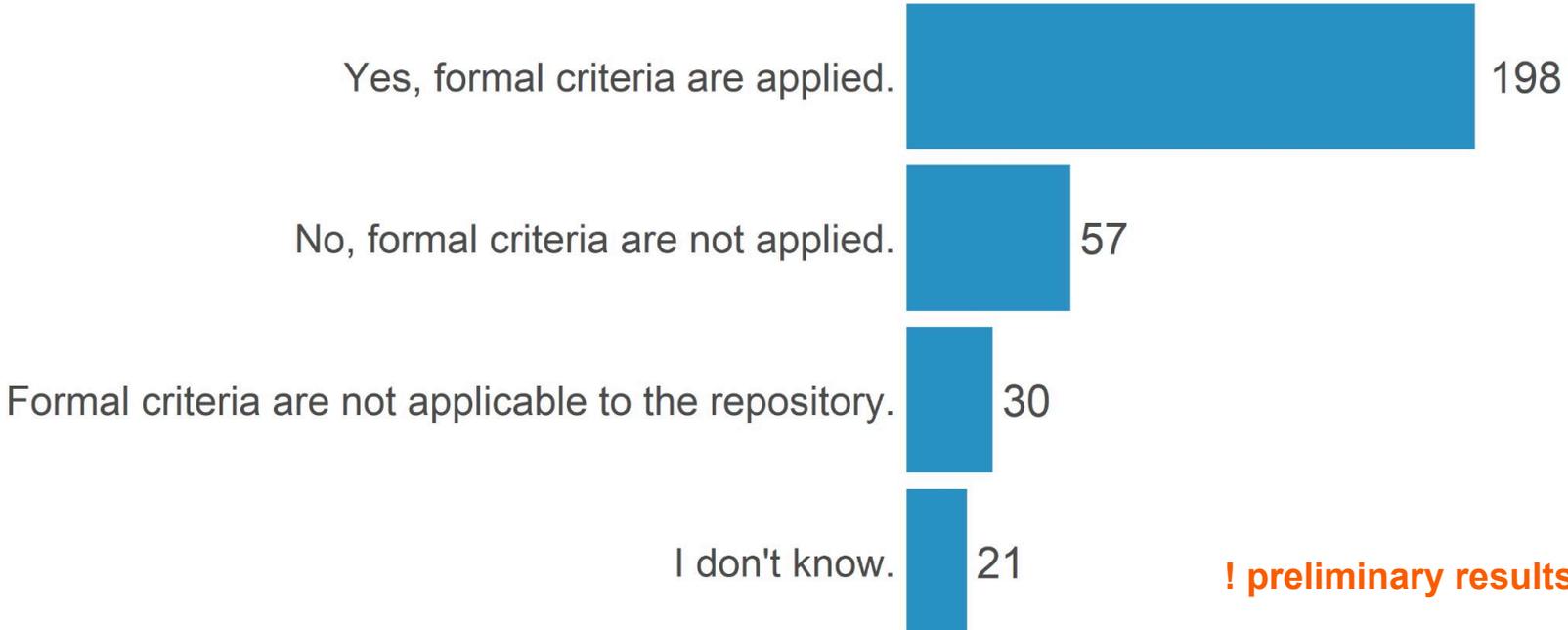
Are data reviewed beyond the application of formal criteria?
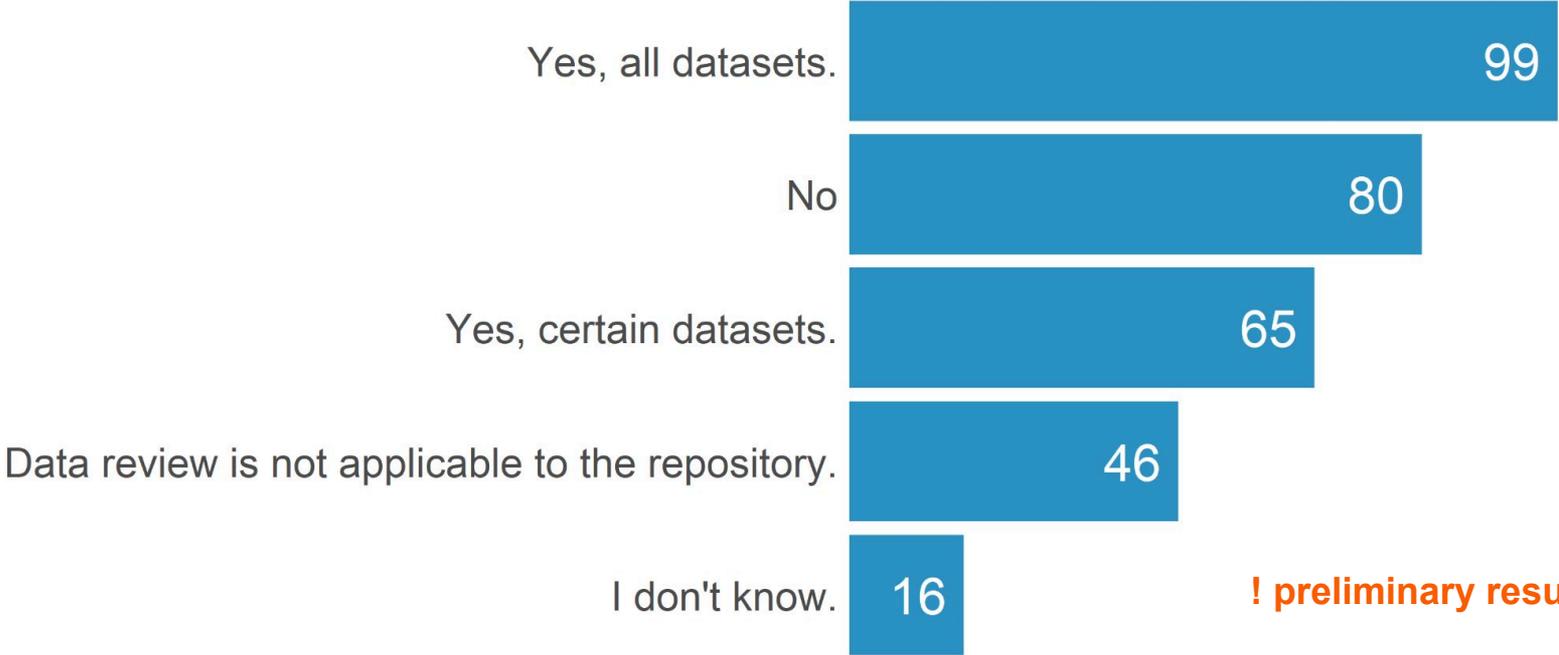
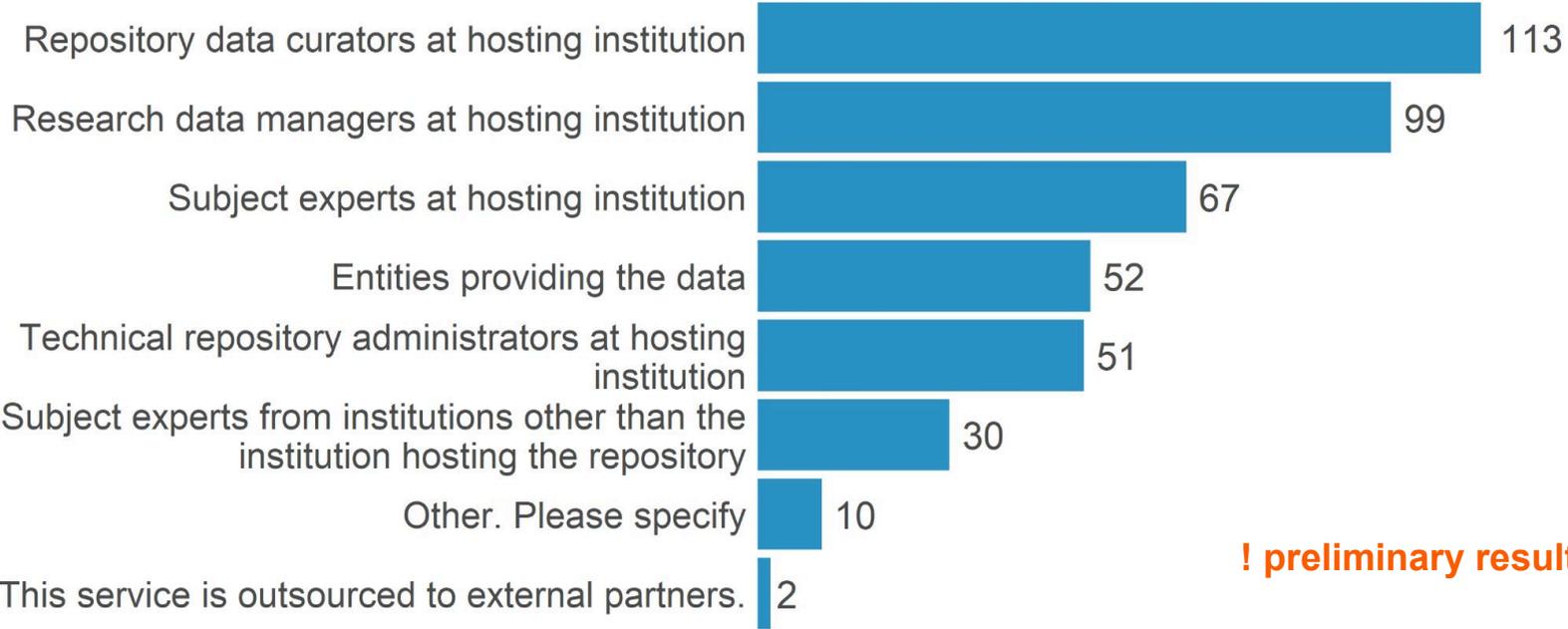| Response	Count
Yes, all datasets.	99
No	80
Yes, certain datasets.	65
Data review is not applicable to the repository.	46
I don't know.	16

! preliminary results !

306 responses from 306 repositories

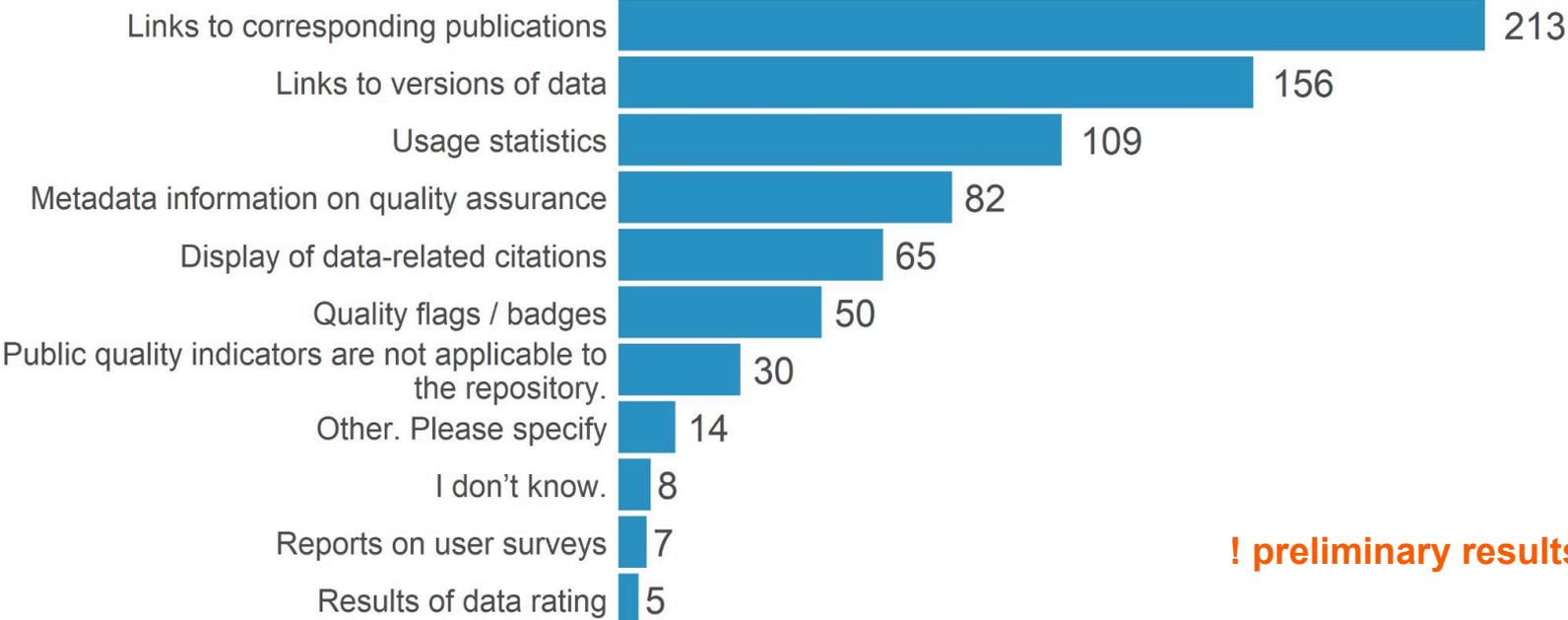# Survey: Data quality management at research data repositories (preliminary)

re3data
coref

Who is responsible for data review? Choose all that apply.
(Please note that the roles listed can overlap.)

Role	Count
Repository data curators at hosting institution	113
Research data managers at hosting institution	99
Subject experts at hosting institution	67
Entities providing the data	52
Technical repository administrators at hosting institution	51
Subject experts from institutions other than the institution hosting the repository	30
Other. Please specify	10
This service is outsourced to external partners.	2

! preliminary results !

424 responses from 164 repositories

Survey: Data quality management at research data repositories (preliminary)

re3data
coref

What types of quality indicators are made public to data users? Choose all that apply.

Quality Indicator	Count
Links to corresponding publications	213
Links to versions of data	156
Usage statistics	109
Metadata information on quality assurance	82
Display of data-related citations	65
Quality flags / badges	50
Public quality indicators are not applicable to the repository.	30
Other. Please specify	14
I don't know.	8
Reports on user surveys	7
Results of data rating	5

! preliminary results !

739 responses from 306 repositories

Outlook and conclusion

- RDR can play a significant role in research data quality assurance
- Visibility of quality management needs to be increased
- As complex as the concept of quality is, so are the applied processes of quality management at repositories
- re3data COREF plans to propose "soft" description/vocabulary for a more sophisticated quality management property in the next version of the metadata schema (4.0)
- re3data COREF plans follow-up activities

Thank You!

re3data
coref

www.re3data.org

@re3data

doi.org/10.1045/march2017-kindling