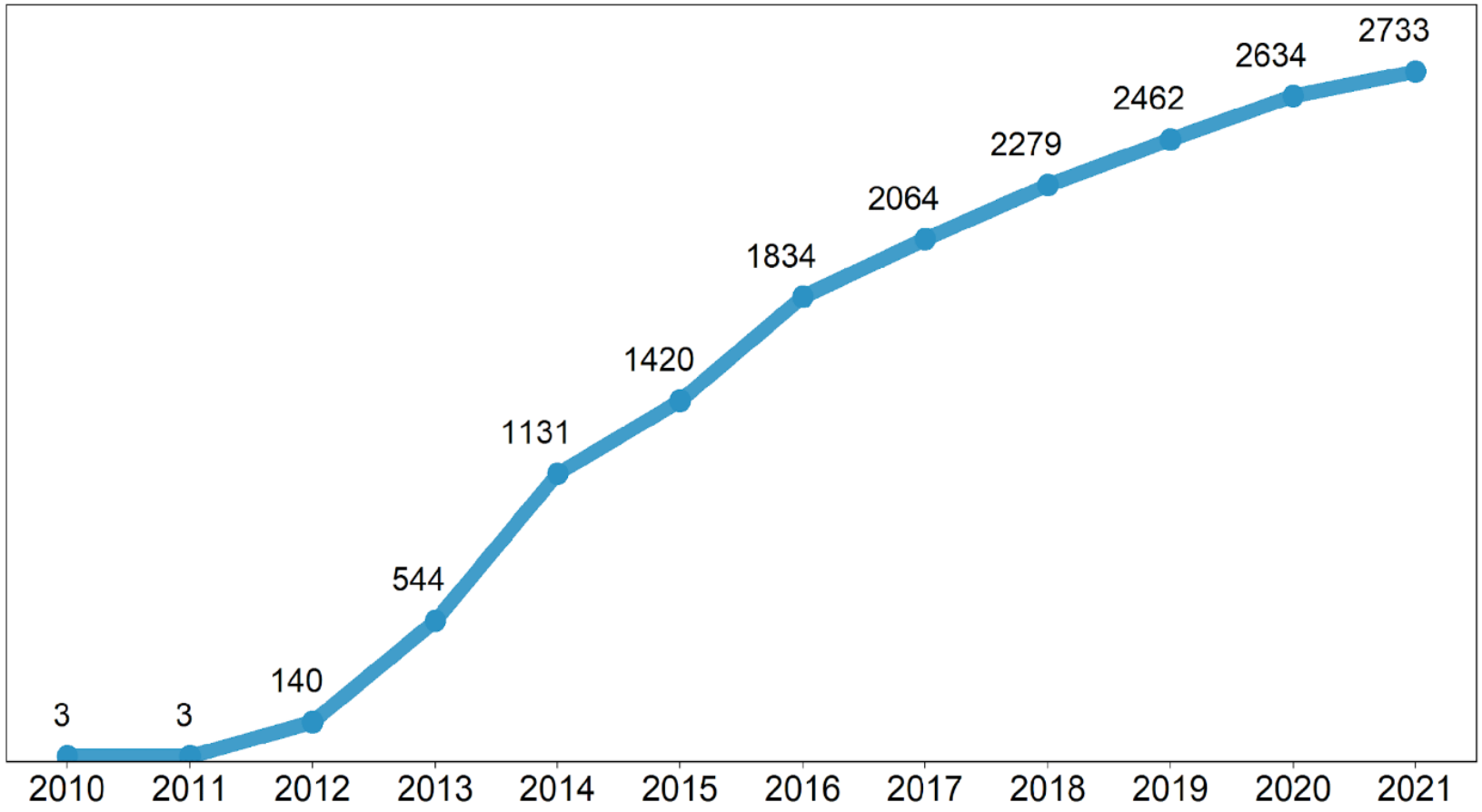
Gateways 2021
Working Session
October 21, 2021
11:45 am - 12:30 pm EST

Introducing re3data – the Registry of Research Data Repositories

Rouven Schabinger (Karlsruher Institut für Technologie (KIT))
Dorothea Strecker (Humboldt-Universität zu Berlin)
Yi Wang (Karlsruher Institut für Technologie (KIT))
Nina Weisweiler (Helmholtz-Gemeinschaft, Helmholtz Open Science Office)

Number of repositories indexed in re3data per year:



October 2021:
More than **2700** research data repositories

DataCite

KIT Karlsruher Institut für Technologie

PURDUE UNIVERSITY

Global Registry of Research Data Repositories

re3data.org

REGISTRY OF RESEARCH DATA REPOSITORIES

Search... Search

across **all research 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 Origins

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

- **2012-2015:** DFG-funded project (Humboldt University, Karlsruhe Institute of Technology, Helmholtz Open Science Office)
- **2013:** Merge with DataBib (new partner: Purdue University, new International Editorial Board)
- **2015:** Official partner service of DataCite
- **2020-2022:** Current DFG-funded project → re3data coref

- Technical maintenance and development of the service is financed and managed by Karlsruhe Institute of Technology (KIT) and DataCite.

DataCite HELMHOLTZ Open Science HUMBOLDT-UNIVERSITÄT ZU BERLIN KIT Karlsruher Institut für Technologie PURDUE UNIVERSITY

What is a “repository” in re3data?

- We generally refer to all re3data entries as “repositories”.
- Repositories in re3data are defined as “[...] a subtype of a sustainable information infrastructure providing long-term storage of and access to research data.”(Metadata Schema for the Description of Research Data Repositories, Version 3.1, p. 5)
- The re3data registration policy is intentionally inclusive to cover a broad range of infrastructures for research data.

Registration Policy

Search Browse Suggest Resources Contact

re3data.org Registration Policy

To be registered in re3data.org a research data repository must

- be run by a legal entity, such as a sustainable institution (e.g. library, university)
- clarify access conditions to the data and repository as well as the terms of use
- **have focus on research data**

A research data repository is a subtype of a sustainable information infrastructure which provides long-term storage and access to research data that is the basis for a scholarly publication. Research data means information objects generated by scholarly projects for example through experiments, measurements, surveys or interviews.

Provider types in re3data

An infrastructure listed in re3data is:

- a **dataProvider** primarily offering research data and their metadata (ideally exposing metadata via application programming interfaces) **and with data upload functionality**,

and/or

- a **serviceProvider** primarily harvesting metadata of research data from data providers, e.g. a **data discovery** portal.

re3data coref

COMMUNITY DRIVEN OPEN REFERENCE FOR RESEARCH DATA REPOSITORIES

Main project goal:
Positioning re3data as a central reference for research data repositories.

- Provide **customizable and extendable core repository descriptions** that are **persistently identifiable** and can be appropriately **referenced and cited**.
- Update and expand the **metadata schema**.
- Develop further options for **automated data exchange** (M2M) between re3data and other services.
- Improve the **(re-)use** of re3data, its API and metadata. Build **new widgets** that take recent requirements of stakeholders into account.
- Provide more functions for **monitoring and recommendation**, for example of repositories that enable FAIR data practices (cooperation with EU project FAIRsFAIR).
- All the above activities are carried out according to a **new conceptual model** for re3data.

DataCite

HELMHOLTZ Open Science

Humboldt-Universität zu Berlin

KIT Karlsruher Institut für Technologie

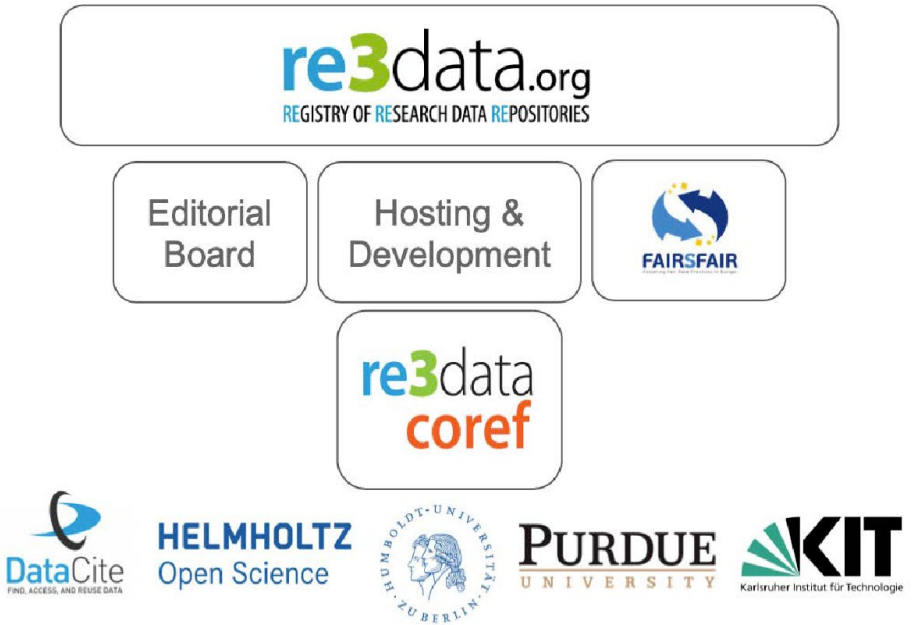
GFZ Helmholtz Centre Potsdam

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation

re3data COREF Project

Funded by: DFG Deutsche Forschungsgemeinschaft

- 36 Months, started in January 2020
- Berlin School of Library and Information Science at the Humboldt-Universität zu Berlin
- DataCite - International Data Citation Initiative e. V.
- Helmholtz Association's Open Science Office (located at the GFZ German Research Centre for Geosciences
- KIT Library at the Karlsruhe Institute of Technology
- The project partners actively participate in national and international initiatives on research data management (DINI, RDA).



Visit the project blog:
coref.project.re3data.org

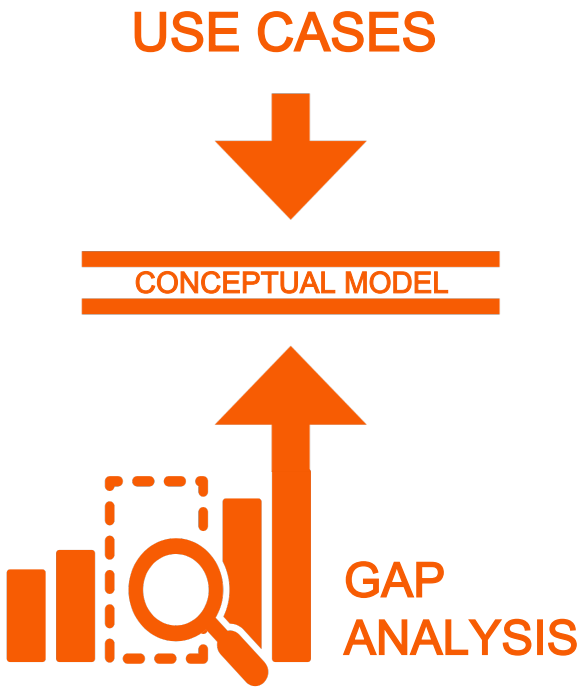
re3data coref

re3data Conceptual Model for User Stories: Most common use cases

1. Search and discovery of research data repositories
2. Reuse re3data metadata
3. Administrate re3data metadata
4. Reference re3data metadata

Feedback was collected via a **survey** and **workshops** with 21 stakeholders to identify relevant **use cases** and **requirements**.

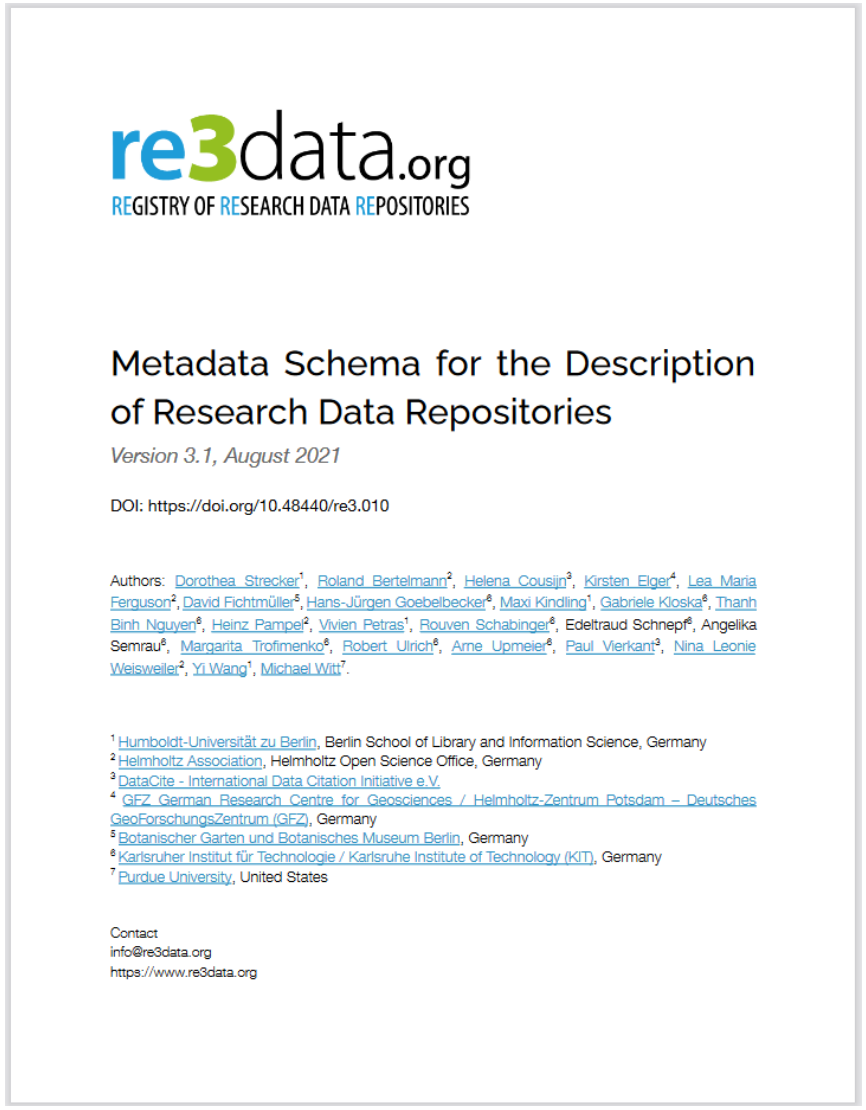
Read the report:
https://doi.org/10.48440/re3.013



Download the model:
https://doi.org/10.48440/re3.012

re3data coref

Metadata Schema Version 3.1



re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Metadata Schema for the Description of Research Data Repositories

Version 3.1, August 2021

DOI: https://doi.org/10.48440/re3.010

Authors: Dorothea Strecker[1], Roland Bertelmann[2], Helena Cousijn[3], Kirsten Elger[4], Lea Maria Ferguson[2], David Fichtmüller[5], Hans-Jürgen Goebelbecker[6], Maxi Kindling[1], Gabriele Kloska[6], Thanh Binh Nguyen[6], Heinz Pampel[2], Vivien Petras[1], Rouven Schabinger[6], Edeltraud Schnepf[6], Angelika Semrau[6], Margarita Trofimenko[6], Robert Ulrich[6], Arne Upmeier[6], Paul Vierkant[3], Nina Leonie Weisweiler[2], Yi Wang[1], Michael Witt[7].

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

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

Strecker S., et al (2021): Metadata Schema for the Description of Research Data Repositories: version 3.1, https://doi.org/10.48440/re3.010

42 Properties on…

- General information
- Responsibilities
- Policies
- Legal aspects
- Technical standards
- Quality standards
- DOI for repository descriptions
- Since 2020: ROR IDs for institutions

Version 3.1 of the Metadata Schema:
https://www.re3data.org/schema

Revision of the metadata schema: New Version 3.1

- **Schema Revision Step 1 –** Version 3.1 introduces two major changes:

"certificate": Version 3.1 will expand descriptions of repository certification. References to a document outlining details of the certificate as well as certification start and end date will describe the certification status in greater detail.

"profile": The new property *profile* provides the opportunity to identify a subset of repositories indexed in re3data pre-defined by a third party based on community-developed criteria (e.g. facilitation of the FAIR principles). Each profile has to provide a URL outlining the selection criteria.

UPCOMING:

- **Schema Revision Step 2** – Version 4.0 (towards the end of the project)
- Mapping re3data metadata to **schema.org**
- **RDF** implementation

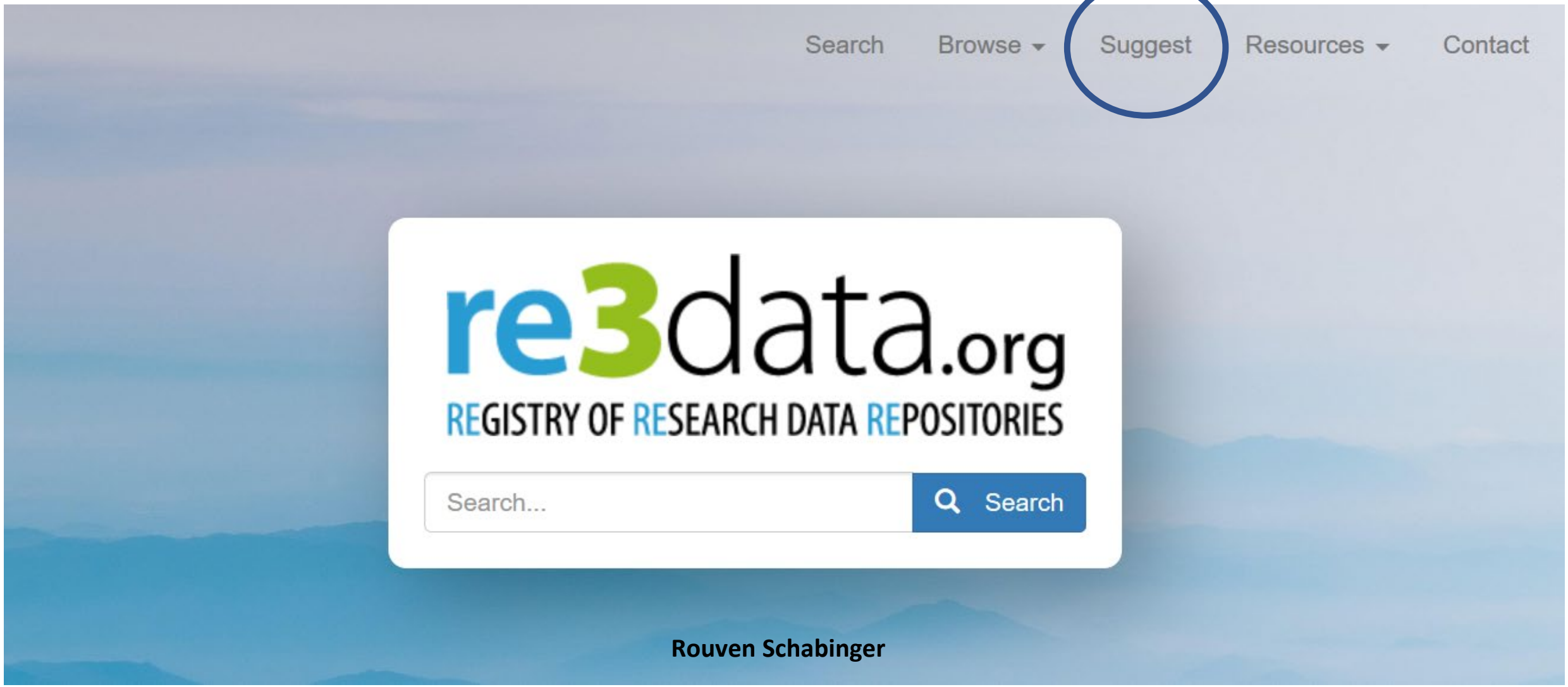
re3data coref

Quality & Transparency:
Making research data quality assurance visible

- Analysis of **data journal** guidelines (sample of 142 data journals and 173 guideline documents)
- **Survey on data quality management at research data repositories** (330 complete responses)
  - Results will be considered in the **update of the metadata schema** to provide information on quality assurance standards and measures for research data repositories
- Review of the re3data **editorial process**
- Building a **Model of Trust for metadata editing**
- Implementing the usage of **authority files** (ORCID)

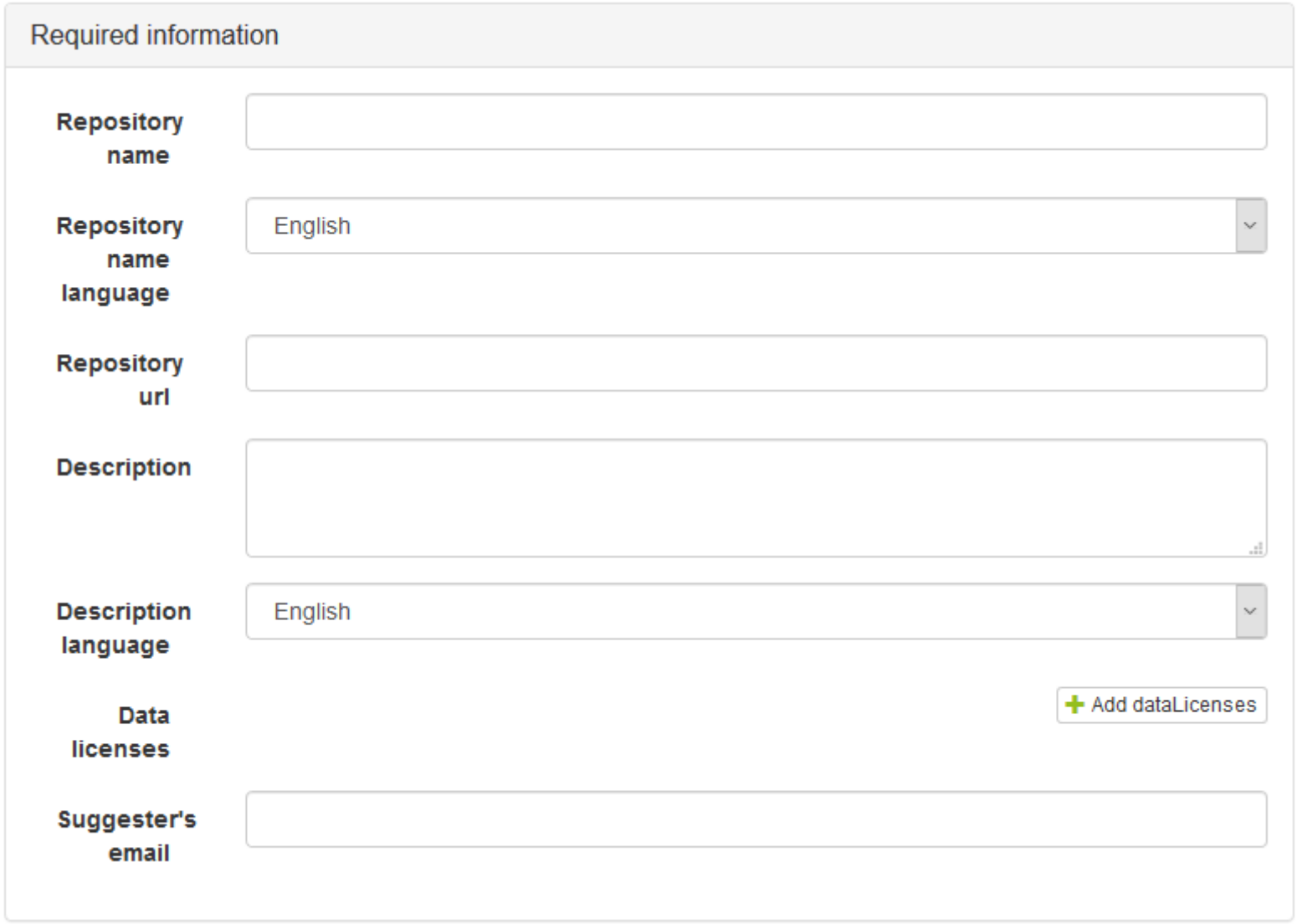
re3data coref

What makes a good registry record?



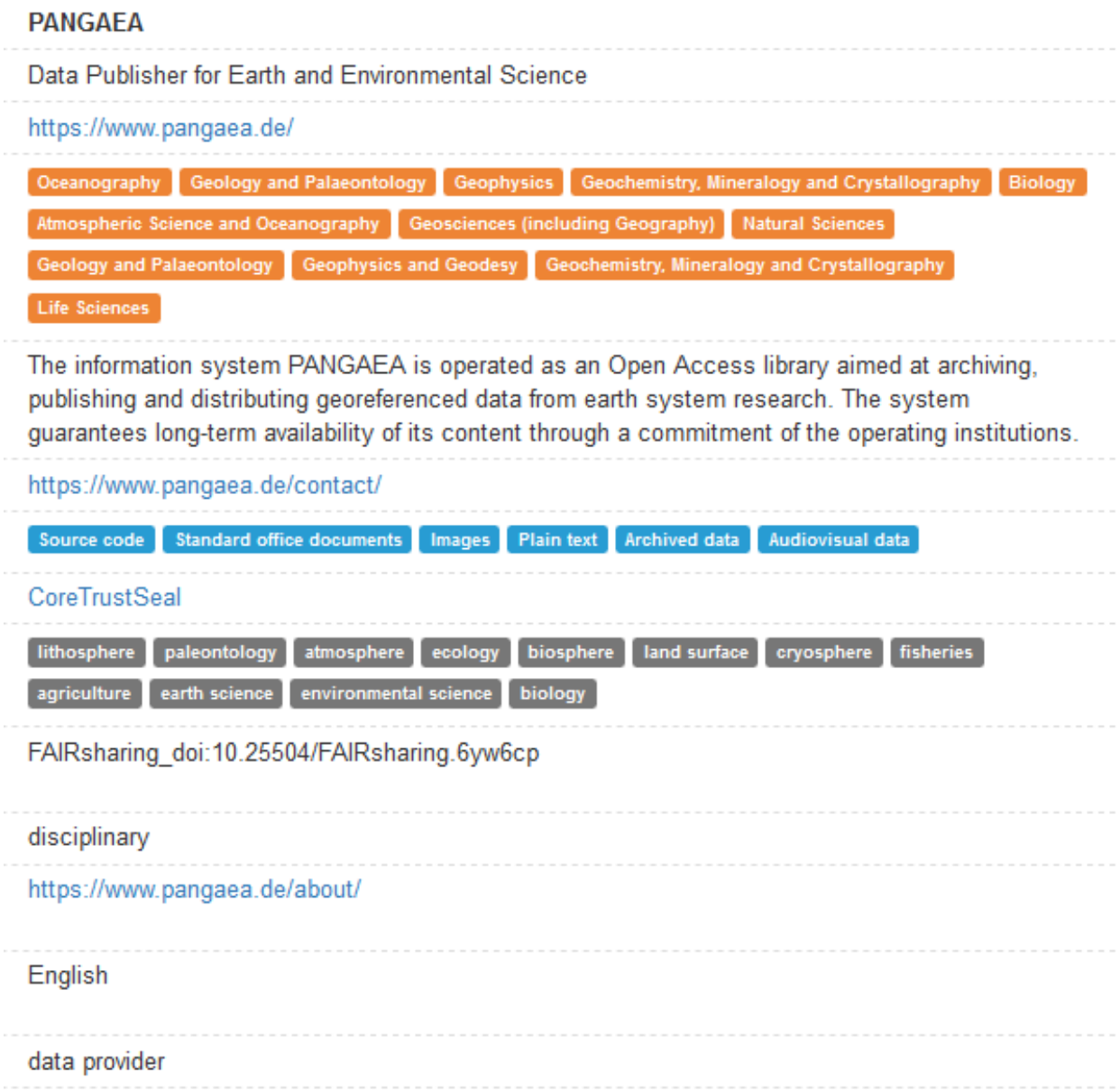
Rouven Schabinger

Required information



General

- Additional names
- **Subjects**
- **Repository contacts**
- Content types
- **Certificates**
- Keywords
- **Repository identifiers**
- Size
- **Types**
- Mission statement URL
- Start date
- Repository languages
- Provider types



Institutions

- Name
- Name language
- **Country**
- Type
- URL
- Responsibility start date
- Responsibility end date
- Additional names
- Contacts
- Identifiers
- **Responsibility types**

Terms

- Policies
- **Database access**
- **Data accesses**
- **Data uploads**
- Data upload licenses

Policies (2)

Policy Name	CoreTrustSeal assessment
Policy Name	Data policy of the information system PANGAEA

Database access

Type of access to research data repository	open

Database licenses (1)

Database License	CC0

Data access (1)

Type of access to data	open

Data licenses (4)

DataLicense	CC
DataLicense	other

Access (property)	*Open Access*		*Restricted Access*		*Closed Access*
Access to Repository (20.1 databaseAccessType)	open		open or restricted		closed
Access to Data (22.1 dataAccessType)	open (embargoed, restricted, closed)		restricted (embargoed, closed)		closed
Data Upload (24.1 dataUploadType)	open or restricted	closed	open or restricted	closed	-

Standards

- Software
- Versioning
- **PID systems**
- Citation guideline URL
- AID systems
- Enhanced publication
- Quality management
- **APIs**
- **Metadata standards**
- Syndications
- Remarks

DOI

other

yes

https://wiki.pangaea.de/wiki/Citation

ORCID

unknown

yes

interfaces (1)

OAI-PMH

Darwin Core

DCC

ISO 19115

DCC

Dublin Core

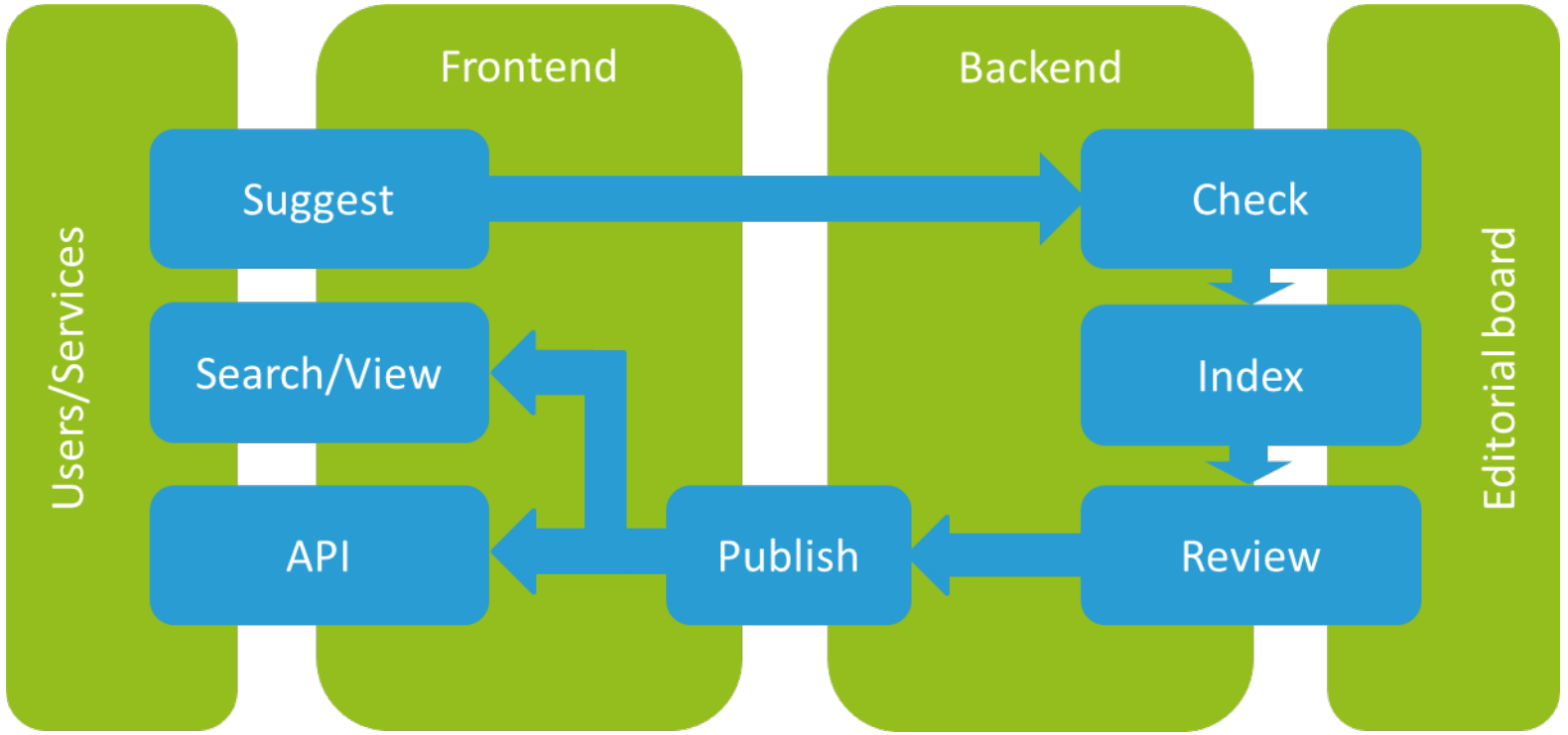
DCC

DIF - Directory Interchange Format

DCC

Icons – facilitating the selection process of appropriate research data repositories

Workflow



re3data Research Data Repository Registration

Editorial Board

Keep your record up-to-date

- automated (e.g. CTS)
- Check ups by Editorial Team
- your change request?

Contact	https://nanohub.org/about/contact
Content type(s)	Databases · Audiovisual data · Software applications
Keyword(s)	Nanomaterial Registry · nanoBIO · nanoscience · nan
Persistent identifier(s) of the repository	RRID:SCR_013963 OMICS_27120
Repository size	6.139 resources
Repository type(s)	disciplinary
Mission statement for designated community	https://nanohub.org/about
Research data repository language(s)	English
Data and/or service provider	data provider service provider

Back to search

Submit a change request

Submit a change request

Make changes to the properties that need an update. The editorial board will review the submitted record and put it online

General

pre-filled metadata fields

Repository name nanoHUB

Repository name language English

Additional names — **Text** nanoHUB.org

Language English

Repository url https://nanohub.org/

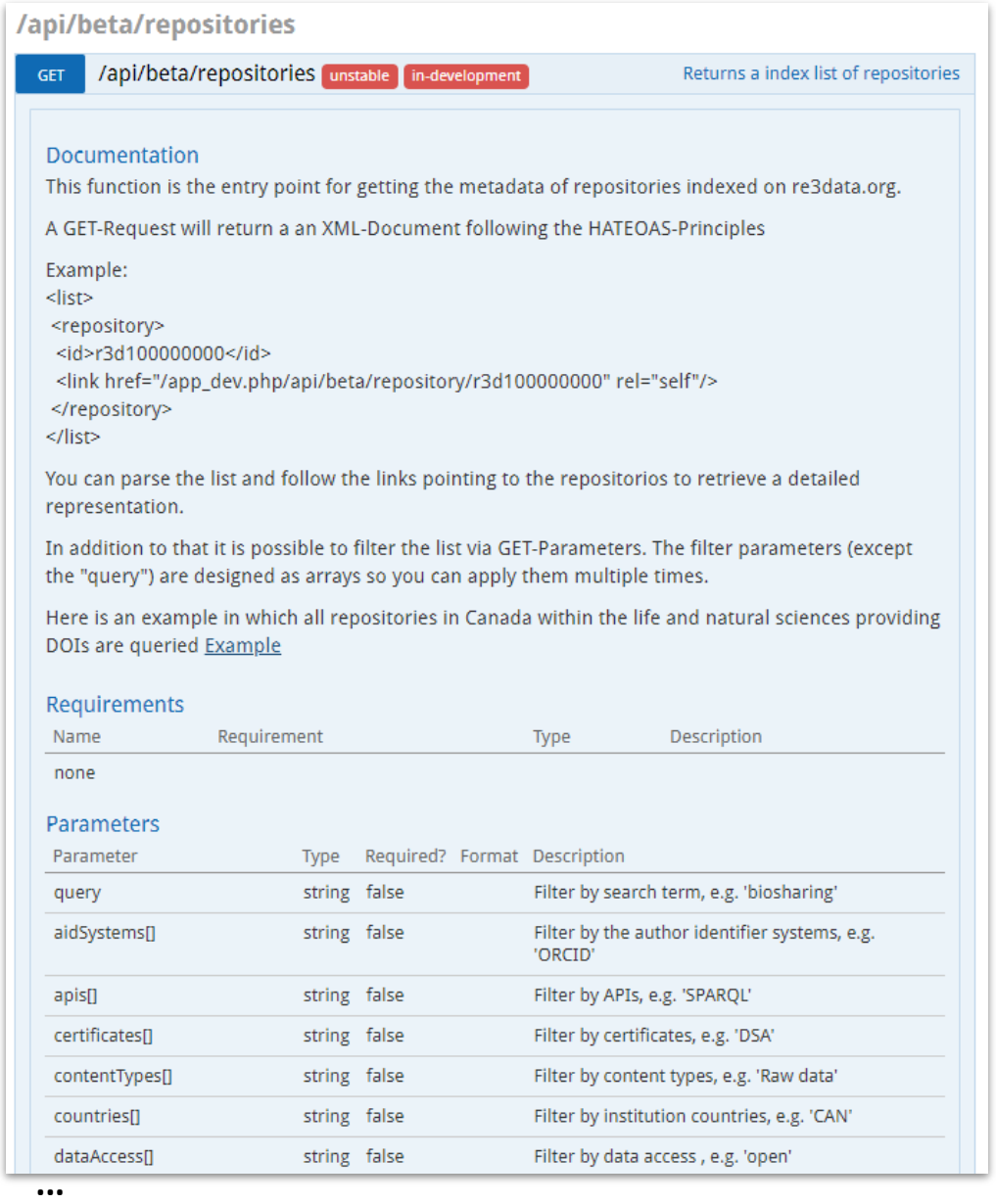
Subjects — **Name** 31 Chemistry

Scheme DFG

Get a badge

The re3data API

- REST API
- results are returned in XML
- several interfaces for getting information on all (**api/v1/repositories**) or specific repositories (**api/v1/repository**)
- query parameters



/api/beta/repositories

GET /api/beta/repositories unstable in-development — Returns a index list of repositories

Documentation

This function is the entry point for getting the metadata of repositories indexed on re3data.org.

A GET-Request will return a an XML-Document following the HATEOAS-Principles

Example:

```
<list>
 <repository>
  <id>r3d100000000</id>
  <link href="/app_dev.php/api/beta/repository/r3d100000000" rel="self"/>
 </repository>
</list>
```

You can parse the list and follow the links pointing to the repositorios to retrieve a detailed representation.

In addition to that it is possible to filter the list via GET-Parameters. The filter parameters (except the "query") are designed as arrays so you can apply them multiple times.

Here is an example in which all repositories in Canada within the life and natural sciences providing DOIs are queried Example

Requirements

Name	Requirement	Type	Description
none			

Parameters

Parameter	Type	Required?	Format	Description
query	string	false		Filter by search term, e.g. 'biosharing'
aidSystems[]	string	false		Filter by the author identifier systems, e.g. 'ORCID'
apis[]	string	false		Filter by APIs, e.g. 'SPARQL'
certificates[]	string	false		Filter by certificates, e.g. 'DSA'
contentTypes[]	string	false		Filter by content types, e.g. 'Raw data'
countries[]	string	false		Filter by institution countries, e.g. 'CAN'
dataAccess[]	string	false		Filter by data access , e.g. 'open'

...

Using the API

- re3data metadata is free - database entries are licensed under CC0
- metadata is based on version 2.2 of the re3data Metadata Schema
- re3data metadata is currently used by several stakeholders for various use cases, including repository recommendation tools, tools for writing data management plans, and monitoring the repository landscape. (Weisweiler et al., 2021)

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Schema for the Description of Research Data Repositories

Version 2.2

December 2014

doi: http://doi.org/10.2312/re3.006

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

[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 Flagship, Kensington WA, Australia

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

query re3data API VS gather information of your friends' cats

set a goal

"As a research data portal, it is important for us to know which repositories offer an API.
We would like to aggregate API information, such as API endpoint, API type and general information about the repository."

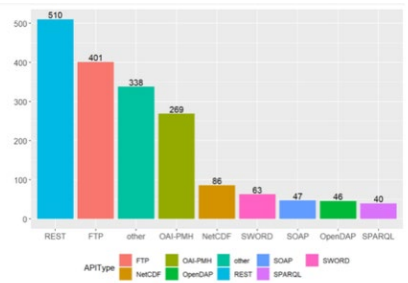
endpoint of url

https://www.re3data.org/api/v1/repository/<repository id>

define requirement

repositoryURL
re3data.orgIdentifier
repositoryName
api
apiType

analysis



set a goal

As a cat lover you want to find the detailed information of your friend's cat.

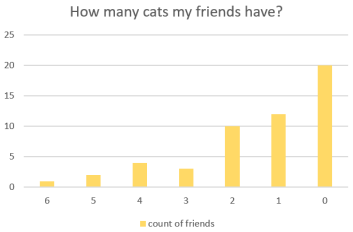
address of your friends

Get address of all your friends

define cat information

cats number
color of cats
cats breed

analysis



Step1: load packages

```
#install.packages("htttr")
#install.packages("xml2")
#install.packages("dplyr")
#install.packages("ggplot2")
library(httr)
library(xml2)
library(dplyr)
library(ggplot2)
```

The package **httr** includes the HTTP method GET, which will be used to request data from the re3data API.
xml2 includes functions for working with XML, for example parsing or extracting content of specific elements.
dplyr offer useful functions for data manipulation and reshaping.
ggplot2 is a package for data visualization.

Step2: obtain URLs for further API queries

The endpoint **/api/v1/repository** provides detailed information about individual repositories that can be accessed via re3data IDs.
Therefore, URLs for the next query are created by adding re3data IDs to the base URL.

```
re3data_request <- GET("http://re3data.org/api/v1/repositories")
re3data_IDs <- xml_text(xml_find_all(read_xml(re3data_request), xpath = "//id"))
URLs <- paste("https://www.re3data.org/api/v1/repository/", re3data_IDs, sep = "")
```

```
> read_xml(re3data_request)
{xml_document}
<list>
 [1] <repository>\n  <id>r3d100000001</id>\n  <name>Odum Institute Archive Dataverse</name>\n  <li ...
 [2] <repository>\n  <id>r3d100000002</id>\n  <name>Access to Archival Databases</name>\n  <link h ...
 [3] <repository>\n  <id>r3d100000004</id>\n  <name>Datenbank Gesprochenes Deutsch</name>\n  <link ...
```

Results

```
> head(re3data_IDs)
[1] "r3d100000001" "r3d100000002" "r3d100000004" "r3d100000005" "r3d100000006" "r3d100000007"
```

```
> head(URLs)
[1] "https://www.re3data.org/api/v1/repository/r3d100000001"
[2] "https://www.re3data.org/api/v1/repository/r3d100000002"
[3] "https://www.re3data.org/api/v1/repository/r3d100000004"
[4] "https://www.re3data.org/api/v1/repository/r3d100000005"
```

Step 3: define what information about the repositories should be requested

```
extract_repository_info <- function(xlm  API_index) {
  list(
   re3data_ID = xml_text(xml_find_all(repository_metadata_XML, "//r3d:re3data.orgIdentifier")),
    repositoryName = xml_text(xml_find_all(repository_metadata_XML, "//r3d:repositoryName")),
    repositoryUrl = xml_text(xml_find_all(repository_metadata_XML, "//r3d:repositoryURL")),
    api = paste(unique(xml_text(xml_find_all(repository_metadata_XML,paste0("//r3d:api", "[", as.character(API_index), "]"))))),
    apiType = paste(unique(xml_text(xml_find_all(repository_metadata_XML,paste0("//r3d:api", "[", as.character(API_index), "]", "/@apiTyp
e")))))
  )
}
```

Step 4: create a container for storing results

```
repository_info <- data.frame(matrix(ncol = 5, nrow = 0))

colnames(repository_info) <- c("re3data_ID", "repositoryName", "repositoryUrl", "api", "apiType")
```

Step 5: gather detailed information about repositories

```
for (url in URLs) {
  repository_metadata_request <- GET(url)
  repository_metadata_XML <-read_xml(repository_metadata_request)
  API_count <- length(paste(unique(xml_text(xml_find_all(repository_metadata_XML, "//r3d:api")))))
  if(API_count > 0){
    for(i in c(1:API_count)){
      results_list <- extract_repository_info(repository_metadata_XML, i)
      repository_info <- rbind(repository_info, results_list)
    }}else next

}
```

request data for specific url

XML responses is parsed with **read_xml**

extract all api elements

count the occurrence of the element api in this repository

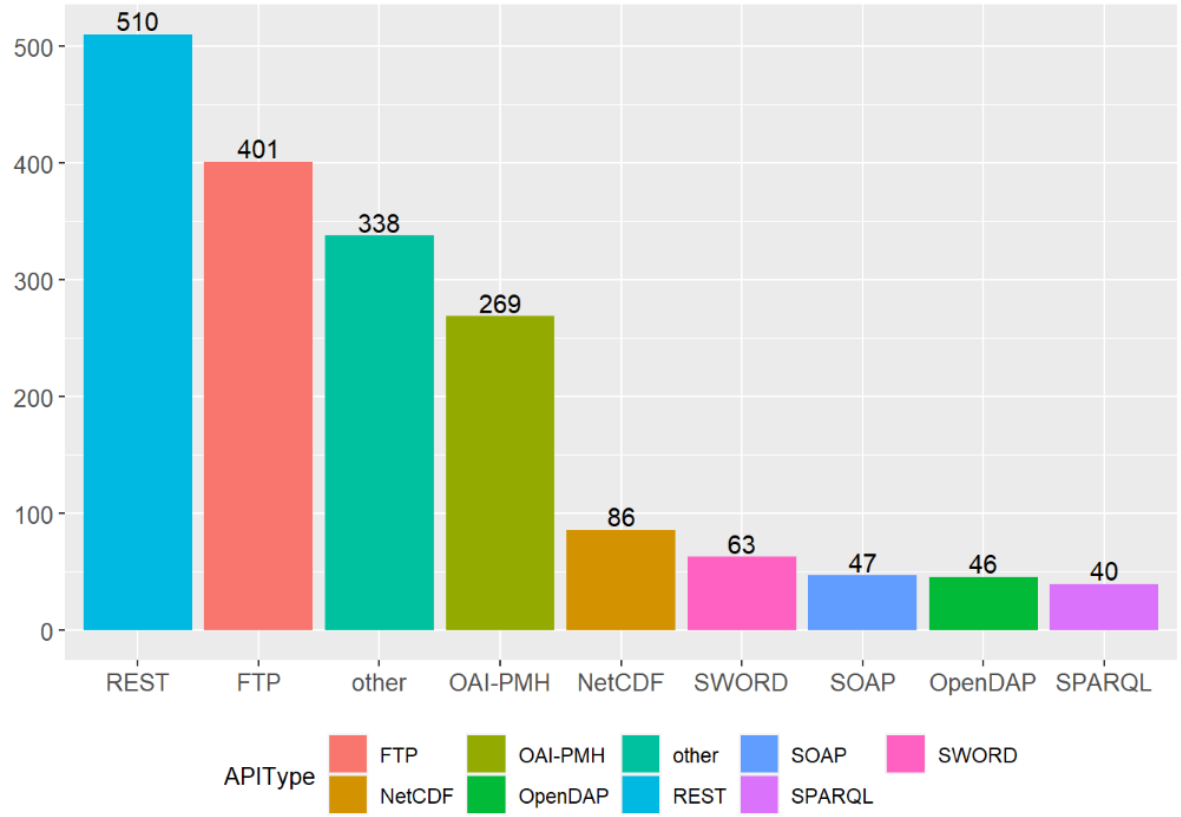
extract information for repositories that have at least 1 API

```
head(repository_info)
```

	re3data_ID	repositoryName	repositoryUrl	api	apiType
	<chr>	<chr>	<chr>	<chr>	<chr>
1	r3d100000002	Access to Archival Databases	https://aad.archives.gov/aad/	https://www.archives.gov/developer#toc-application-programming-interfaces-apis-	other
2	r3d100000005	UNC Dataverse	https://dataverse.unc.edu/	https://guides.dataverse.org/en/latest/api/native-api.html	REST
3	r3d100000005	UNC Dataverse	https://dataverse.unc.edu/	https://guides.dataverse.org/en/latest/api/sword.html	SWORD
4	r3d100000006	Archaeology Data Service	https://archaeologydataservice.ac.uk/	https://archaeologydataservice.ac.uk/about/endpoints.xhtml	OAI-PMH
5	r3d100000006	Archaeology Data Service	https://archaeologydataservice.ac.uk/	http://data.archaeologydataservice.ac.uk/query/	SPARQL
6	r3d100000011	The CEDA Archive	http://archive.ceda.ac.uk/	https://help.ceda.ac.uk/article/280-ftp	FTP

Step 6: visualize the results

```
repository_info %>%
  group_by(apiType) %>%
  summarise(type_count = n()) %>%
  ggplot(aes(x = reorder(apiType, -type_count), y = type_count, fill = apiType)) +
  geom_col() +
  geom_text(aes(label = type_count), position = position_dodge(width = 1),vjust = -0.2)+
  theme(axis.title = element_blank(),
        axis.text = element_text(size = 10),
        legend.position = "bottom",
        legend.title = element_text(size = 10))
```



Source code for this workshop and other examples

re3data / using_the_re3data_API Public

Gateways 2021
Working Session
October 21, 2021
11:45 am - 12:30 pm EST

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

THANK YOU!

Contact us!

Web: **www.re3data.org**
Mail: **info@re3data.org**
Twitter: **@re3data**

Visit the project blog:
coref.project.re3data.org

The Team:

Roland Bertelmann (Helmholtz-Gemeinschaft, Helmholtz Open Science Office)
Helena Cousijn (DataCite - International Data Citation Initiative e.V.)
Kirsten Elger (Deutsches GeoForschungsZentrum GFZ)
Lea Maria Ferguson (Helmholtz-Gemeinschaft, Helmholtz Open Science Office)
Maxi Kindling (Humboldt-Universität zu Berlin)
Thanh Binh Nguyen (Karlsruher Institut für Technologie (KIT))
Heinz Pampel (Helmholtz-Gemeinschaft, Helmholtz Open Science Office)
Vivien Petras (Humboldt-Universität zu Berlin)
Rouven Schabinger (Karlsruher Institut für Technologie (KIT))
Edeltraut Schnepf (Karlsruher Institut für Technologie (KIT))
Angelika Semrau (Karlsruher Institut für Technologie (KIT))
Dorothea Strecker (Humboldt-Universität zu Berlin)
Margarita Trofimenko (Karlsruher Institut für Technologie (KIT))
Robert Ulrich (Karlsruher Institut für Technologie (KIT))
Arne Upmeier (Karlsruher Institut für Technologie (KIT))
Paul Vierkant (DataCite - International Data Citation Initiative e.V.)
Yi Wang (Karlsruher Institut für Technologie (KIT))
Nina Weisweiler (Helmholtz-Gemeinschaft, Helmholtz Open Science Office)
Michael Witt (Purdue University, United States)