Lecture 21 – Repositories

Open Data Management & the Cloud

(Data Science & Scientific Computing / UniTS – DMG)

Real argument of the lesson

- Storage
- Repository
- Registry/catalogue
- Archive

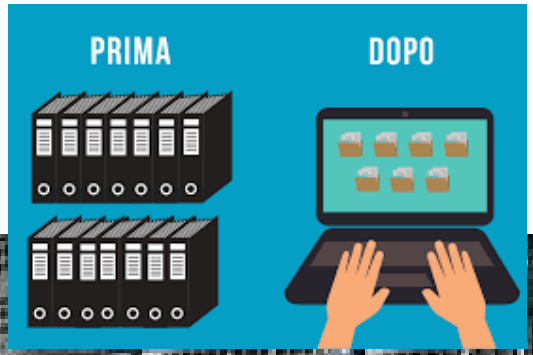
What are?

How we use them?

Examples

Analogy

Analogy



Repository – What is it?

- Data storage / Data base entity/ies into which data has been specifically partitioned for an analytical or reporting purpose
  - Data warehouse
    - a large data repository that aggregates data
    - usually from multiple sources or segments of a business
    - without the data being necessarily related
  - Data lake (needs disambiguation)
    - a large data repository that stores unstructured data
    - classified and tagged with metadata
  - Data mart
    - subsets of the data repository
    - more targeted to what the data user needs and easier to use
  - Data library
  - Data archive
- Particular kind of setup within an overall IT structure, such as a group of databases, where an enterprise or organization has chosen to keep various kinds of data
- Metadata repositories store data about data and data bases
  - Where the data source is, how it was captured, and what it represents

Repository – Some disambiguation

- Repository has additional functionalities compared with registry
  - Registries stores metadata
  - Repositories add relationships with related metadata types
  - Distinction is really loose / hardly enforced
- Disambigua to collections
  - Grouping of digital objects, within a scope
  - Collections are themselves resources in repositories
- What is the difference between archive and repository?

  https://wikidiff.com/archive/repository

  https://www.usgs.gov/data-management/archive-vs-repository-there-difference

  https://www.familytreemagazine.com/libraries-archives/library-repository-archives/

  https://dictionary.archivists.org/entry/repository.html

Repository – Goals

- FAIR principles: Findable
  - *“F4. (meta) data are registered or indexed in a searchable resource.”*
- Keep a certain population of data isolated so that it can be mined
    - isolated: living in its one fragmented/distributed source
  - For greater insight
  - For business intelligence
  - To be used for a specific (reporting) need
- Isolation allows for easier and faster data reporting or analysis because the data is clustered together
  - Not necessarily by location
- Data is preserved and archived

Metadata Repository

- Data repositories
  - Index data
  - Provide access to
    - Data collections
    - Datasets
  - (usually) keep meta information dedicated to filtering
- Metadata repositories
  - (usually) don’t provide storage space
  - Use linking solutions to point or access data
  - Provide rich metadata documents
    - (usually) standardized
    - More general in scope
      - discovery/filtering
      - relationships

Repository – Usage

- There are more and more data open and available through data repositories: it becomes ever more challenging for researchers to find relevant data.
- Example
  - The Open Access Directory (OAD) is a compendium of simple factual lists about open access (OA) to science and scholarship, maintained by the OA community at large. By bringing many OA-related lists together in one place, OAD makes it easier for everyone to discover them, use them for reference, and update them.

Repository – Usage

- There are more and more data open and available through data

Data repositories

This list is part of the Open Access Directory.

- This is a list of repositories and databases for open data.
- Please annotate the entries to indicate the hosting organization, scope, licensing, and usage restrictions (if any). If a repository is open in some respects but not others, please include it with an annotation rather than exclude it.
- If you're not sure whether a given dataset or data collection is open, post your query to Is It Open Data?
- Related lists in OAD: Disciplinary repositories (primarily for texts, not data).
- For news about data repositories, including some newly launched repositories not yet listed here, follow the oa.repositories.data tag of the Open Access Tracking Project.
- See also: re3data.org. The re3data.org project intends to create a global registry of research data repositories.

Contents [hide]

Archaeology

- *Also see* Social sciences.
- Archaeology Data Service.
- Fasti Online. Subdivided in Excavation, Restauration and Survey.
- Open Context. From the Alexandria Archive Institute.
- the Digital Archaeological Record. From Digital Antiquity.

Astronomy

- *Also see* Physics.
- Astronomical Data Archives Center. From the National Astronomical Observatory of Japan.
- Astrophysics Data System. From the Smithsonian Astrophysical Observatory (SAO) and National Aeronautics and Space Administration (NASA).
- The Canadian Astronomy Data Centre. From the National Research Council Canada.
- National Space Science Data Center. From the US National Aeronautics and Space Administration (NASA).

Biology

- *Also see* BCO-DMO, Marine Biology data, listed with Marine Sciences repositories.
- *Also see* DataONE, Entrez databases, KNB, and PANGAEA, listed under Multidisciplinary repositories.
- The Arabidopsis Information Resource - The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for the model higher plant *Arabidopsis thaliana*.
- BOND (Biomolecular Object Network Databank). From Unleashed Informatics.
- The Cell: An Image Library Images of all cell types from all organisms, including intracellular structures and movies or animations demonstrating functions. This project relies upon the cell biology community to populate the library. The Cell: An Image Library™ is a freely accessible, easy-to-search, public repository of reviewed and annotated images, videos, and animations of cells from a variety of organisms, showcasing cell architecture, intracellular functionalities, and both normal and abnormal processes. The purpose of this

Repository – Usage

- There are more and more data open and available through data repositories: it becomes ever more challenging for researchers to find relevant data.
- Example
  - The Open Access Directory (OAD) is a compendium of simple factual lists about open access (OA) to science and scholarship, maintained by the OA community at large. By bringing many OA-related lists together in one place, OAD makes it easier for everyone to discover them, use them for reference, and update them.
- Where to start from?
  - Use Google?
  - Use a specific “global” research index?
  - Find your domain starting point?

RDA Recommendations – DDP-IG (1)

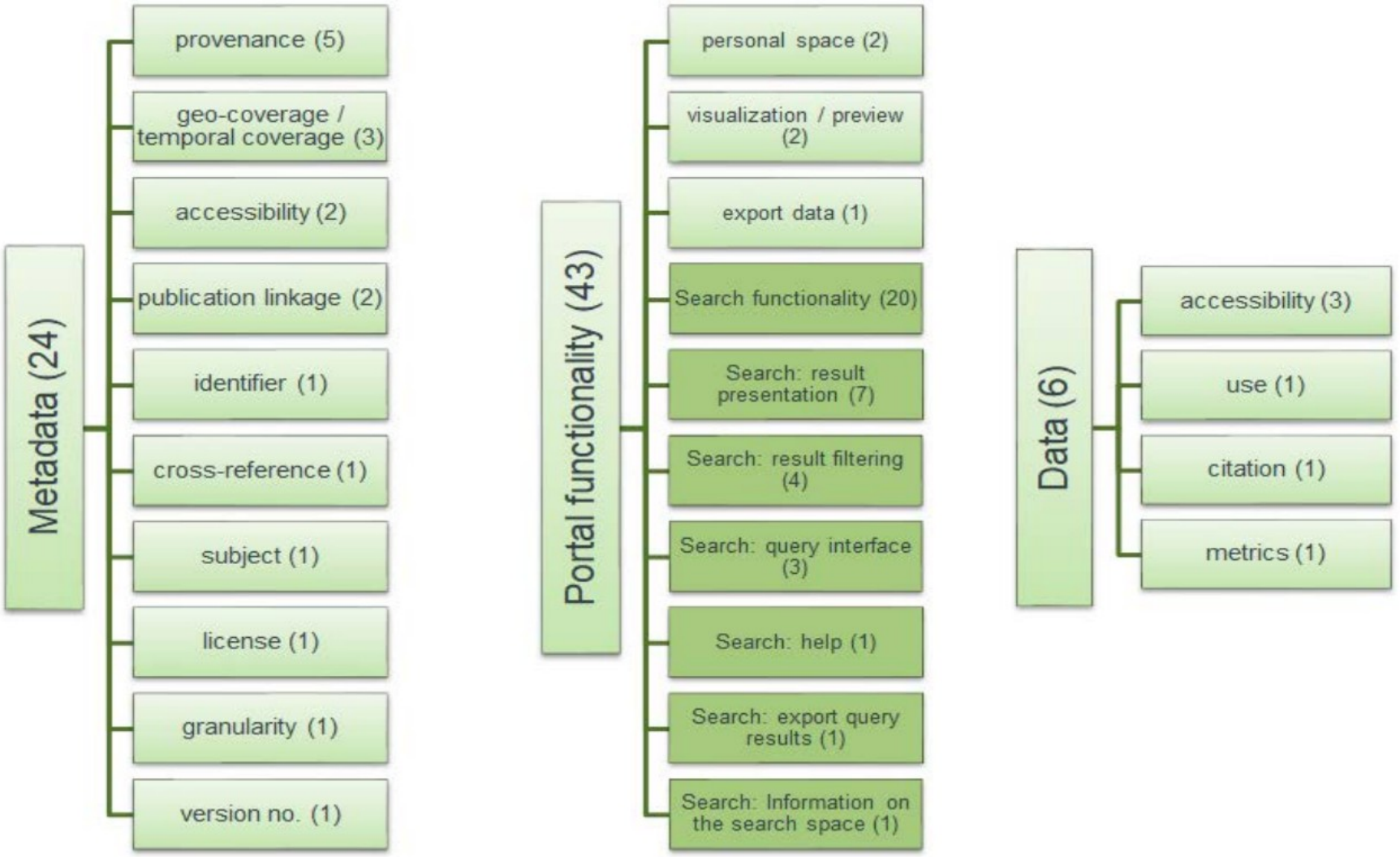
- Data Discovery Paradigms (Interest Group)
  - User Requirements and Recommendations for Data Repositories
    - https://www.rd-alliance.org/group/data-discovery-paradigms-ig/outcomes/data-discovery-paradigms-user-requirements-and
- Purpose
  - Help data repositories improve the findability of data in their repository
- Approach
  - Collected use cases describing users' needs, and the contexts of these needs, when searching for data
- Outcome
  - Identified requirements for data discovery in repositories
  - Proposed a set of recommendations

RDA – DDP-IG (2) – Rationale

- Data repositories can adapt methodologies and learnt experiences from the design of web-based information systems, and digital library in particular
- Some data repositories have been following the path of user-centred system design principle gathering use cases and requirements
- Some design and evaluation criteria may apply only to a specific repository
- Some trans-repository criteria can be generalised and may serve as guidelines for other data repositories
- DDP-IG study attempts to identify which criteria are of common relevance

RDA – DDP-IG (3) – Use cases

classification scheme summary



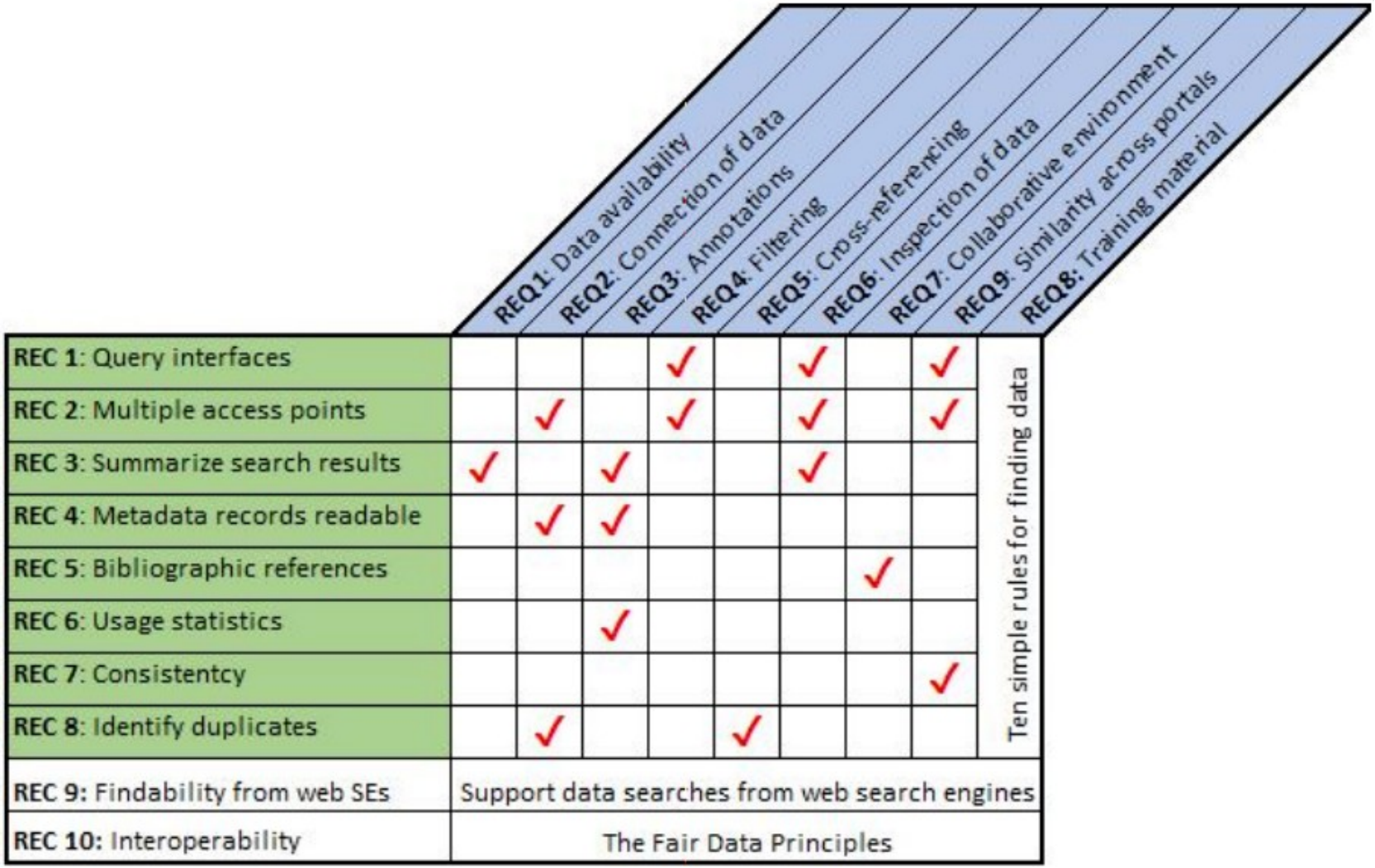
RDA – DDP-IG (4) – Requirements

1) Indication of data availability
   - Search usually dropped if no clear indication of data availability
2) Connection of data with person / institution / paper / citations / grants
   - allows for ranking of datasets, comparative studies, manuscript direct connection
3) Fully annotated data (including granularity, origin, licensing, provenance, ...)
   - validate the use of a dataset in a particular study
4) Filtering of data based on specific criteria on multiple fields at the same time
   - Support targeted studies
5) Cross-referencing of data (same or different repositories)
   - avoid duplication, maximise efficiency and access
6) Visual analytics / inspection of data / thumbnail preview
   - quick visual filtering from a results set to validate dataset use
7) Sharing data (whole dataset/particular records/bibliographic information) in a collaborative environment
   - common space of keeping both data and their versions across time
   - quick check on latest changes
   - share bibliographic information
8) Accompanying educational / training material
9) Portal functionality similar to other established academic portals
   - subject/visual search, free text search, build query functionality, subscription, ...

RDA – DDP-IG (5) – Recommendations

1) Provide a range of query interfaces to accommodate various data search behaviours
2) Provide multiple access points to find data
3) Make it easier to judge relevance, accessibility and reusability of a data collection
4) Make Individual metadata records readable and analysable
5) Be able to share and output bibliographic references
6) Provide feedback about data usage statistics
7) Be consistent with other repositories
8) Identify and aggregate metadata records that describe the same data object
9) Make metadata records easily indexed and searchable by major web search engines
10) Follow API search standards and community adopted vocabularies for interoperability

RDA – DDP-IG (6) – Recommendations



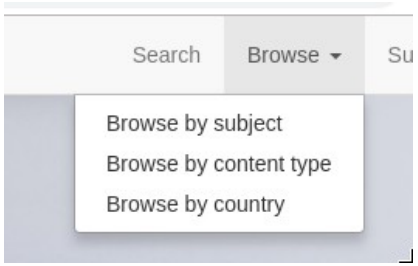
	REQ1: Data availability	REQ2: Connection of data	REQ3: Annotations	REQ4: Filtering	REQ5: Cross-referencing	REQ6: Inspection of data	REQ7: Collaborative environment	REQ9: Similarity across portals	REQ8: Training material
REC 1: Query interfaces				✓		✓		✓	Ten simple rules for finding data
REC 2: Multiple access points		✓		✓		✓		✓	
REC 3: Summarize search results	✓		✓			✓			
REC 4: Metadata records readable		✓	✓						
REC 5: Bibliographic references							✓		
REC 6: Usage statistics			✓						
REC 7: Consistentcy								✓	
REC 8: Identify duplicates		✓			✓				
REC 9: Findability from web SEs	Support data searches from web search engines								
REC 10: Interoperability	The Fair Data Principles								

Repositories - examples

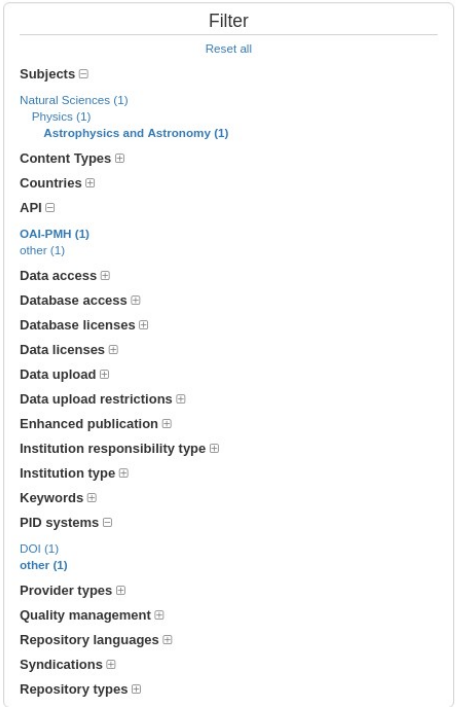
- Repositories can be
  - Domain specific
  - General purpose
- Following slides will show examples of
  - Community driven efforts
  - Project based solutions
  - Super-national driven entities
  - Document based repositories
  - Linked data solutions

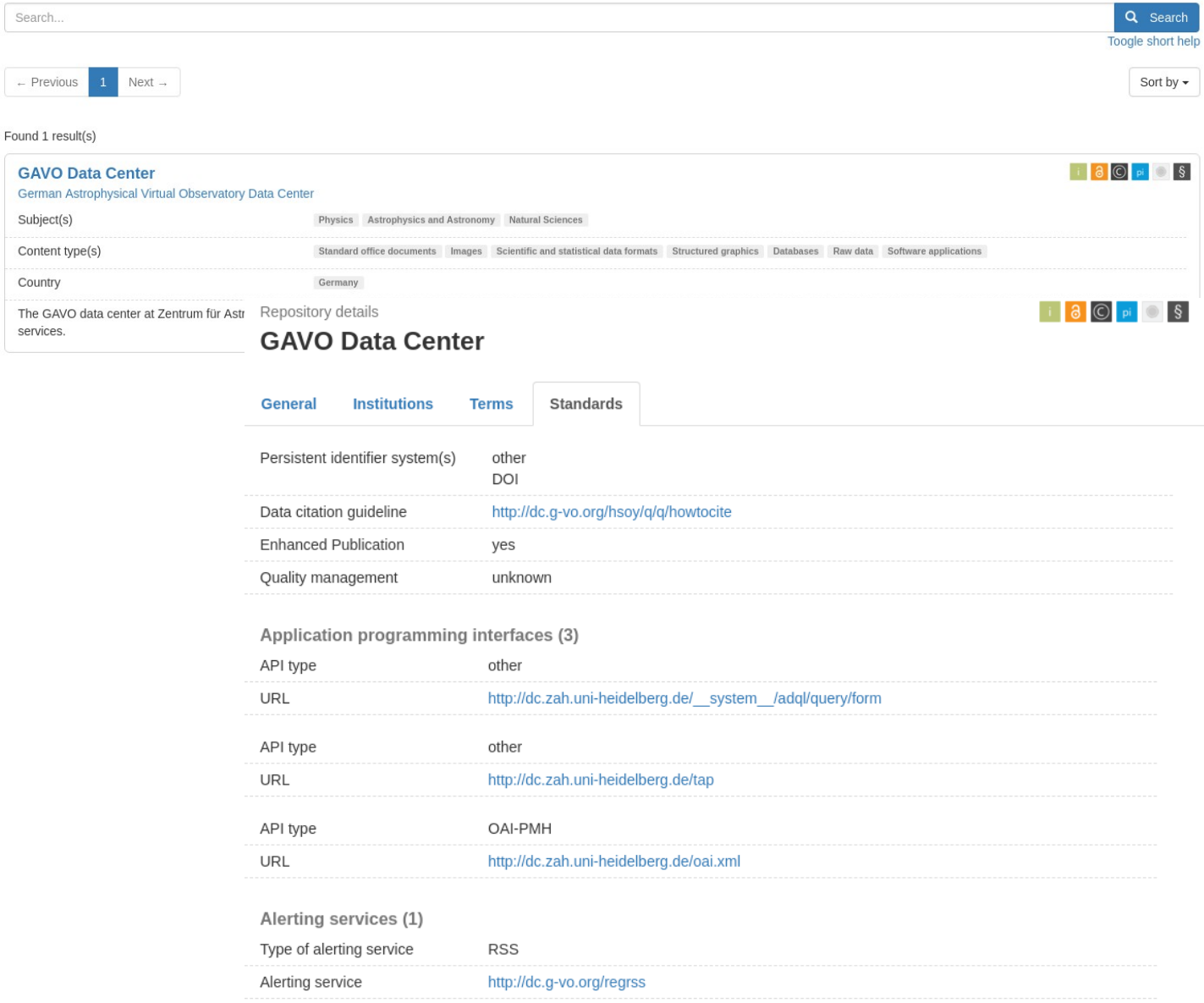
re3data

- REgistry of REsearch REpositories (re3data.org)
- “By offering detailed information on more than 2,000 research data repositories, re3data has become the most comprehensive source of reference for research data infrastructures globally”

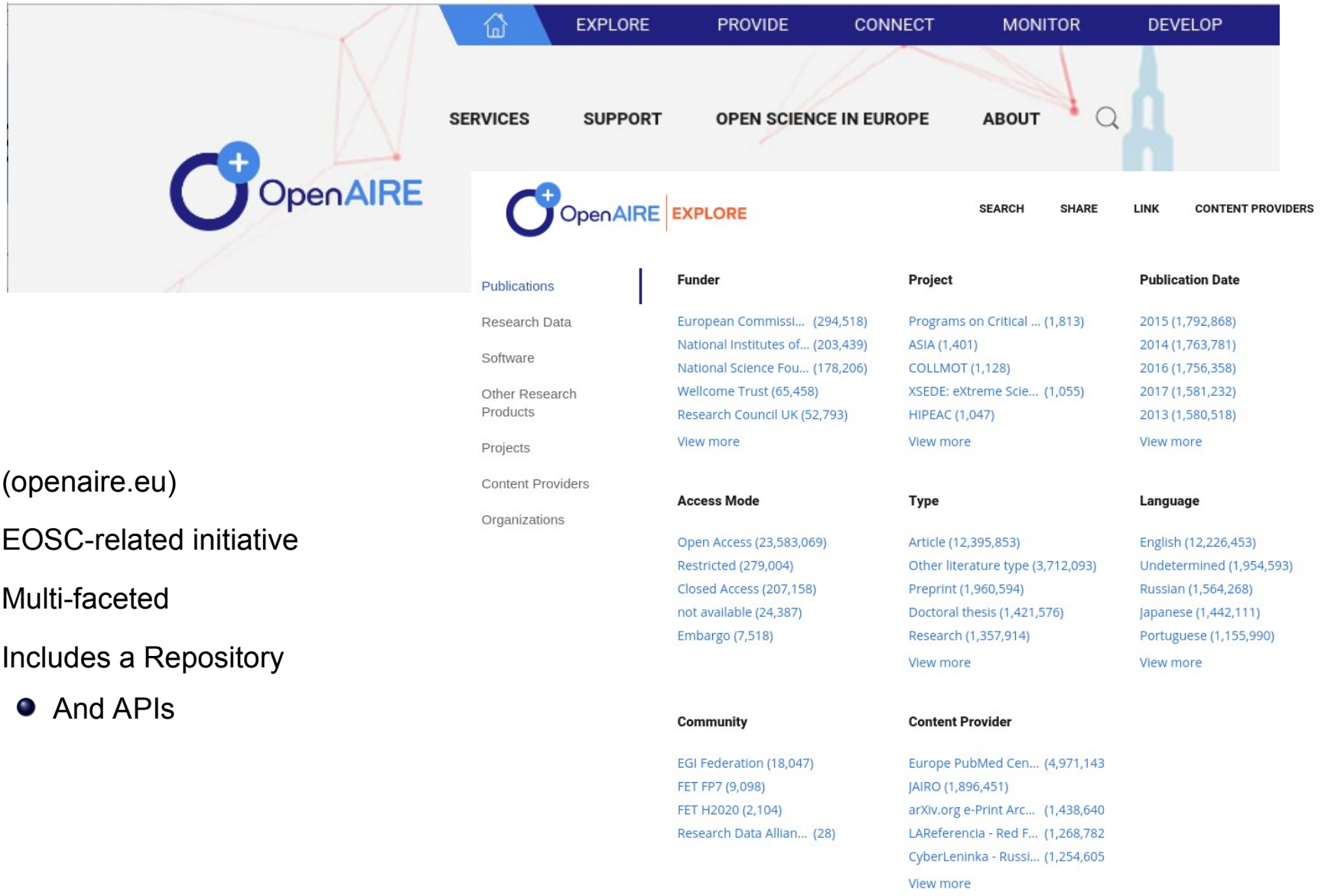


re3data


Filter
Reset all
Subjects
Natural Sciences (1)
Physics (1)
Astrophysics and Astronomy (1)
Content Types
Countries
API
OAI-PMH (1)
other (1)
Data access
Database access
Database licenses
Data licenses
Data upload
Data upload restrictions
Enhanced publication
Institution responsibility type
Institution type
Keywords
PID systems
DOI (1)
other (1)
Provider types
Quality management
Repository languages
Syndications
Repository types


Search...
Search
Toogle short help
← Previous
1
Next →
Sort by
Found 1 result(s)
GAVO Data Center
German Astrophysical Virtual Observatory Data Center
Subject(s)
Physics
Astrophysics and Astronomy
Natural Sciences
Content type(s)
Standard office documents
Images
Scientific and statistical data formats
Structured graphics
Databases
Raw data
Software applications
Country
Germany
The GAVO data center at Zentrum für Astr
services.
Repository details
GAVO Data Center
General
Institutions
Terms
Standards
Persistent identifier system(s)
other
DOI
Data citation guideline
http://dc.g-vo.org/hsoy/q/q/howtocite
Enhanced Publication
yes
Quality management
unknown
Application programming interfaces (3)
API type
other
URL
http://dc.zah.uni-heidelberg.de/__system__/adql/query/form
API type
other
URL
http://dc.zah.uni-heidelberg.de/tap
API type
OAI-PMH
URL
http://dc.zah.uni-heidelberg.de/oai.xml
Alerting services (1)
Type of alerting service
RSS
Alerting service
http://dc.g-vo.org/regrss

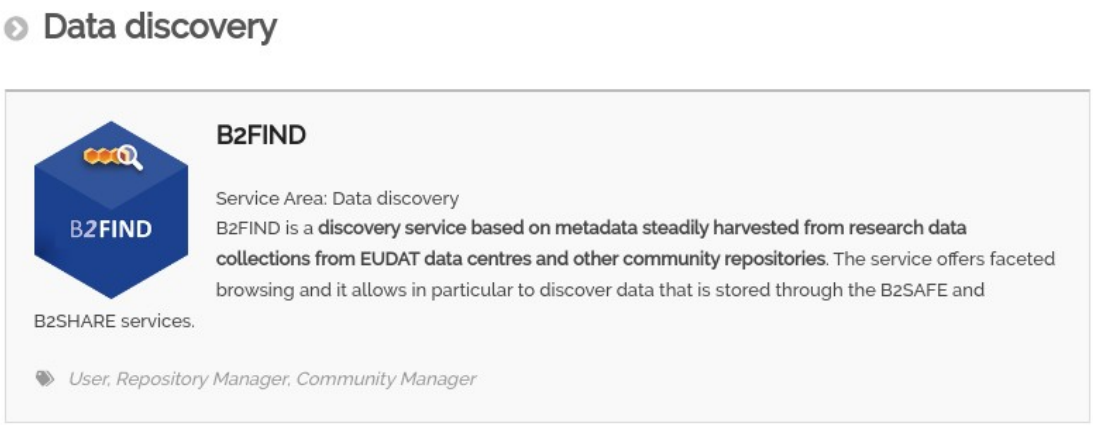
OpenAIRE



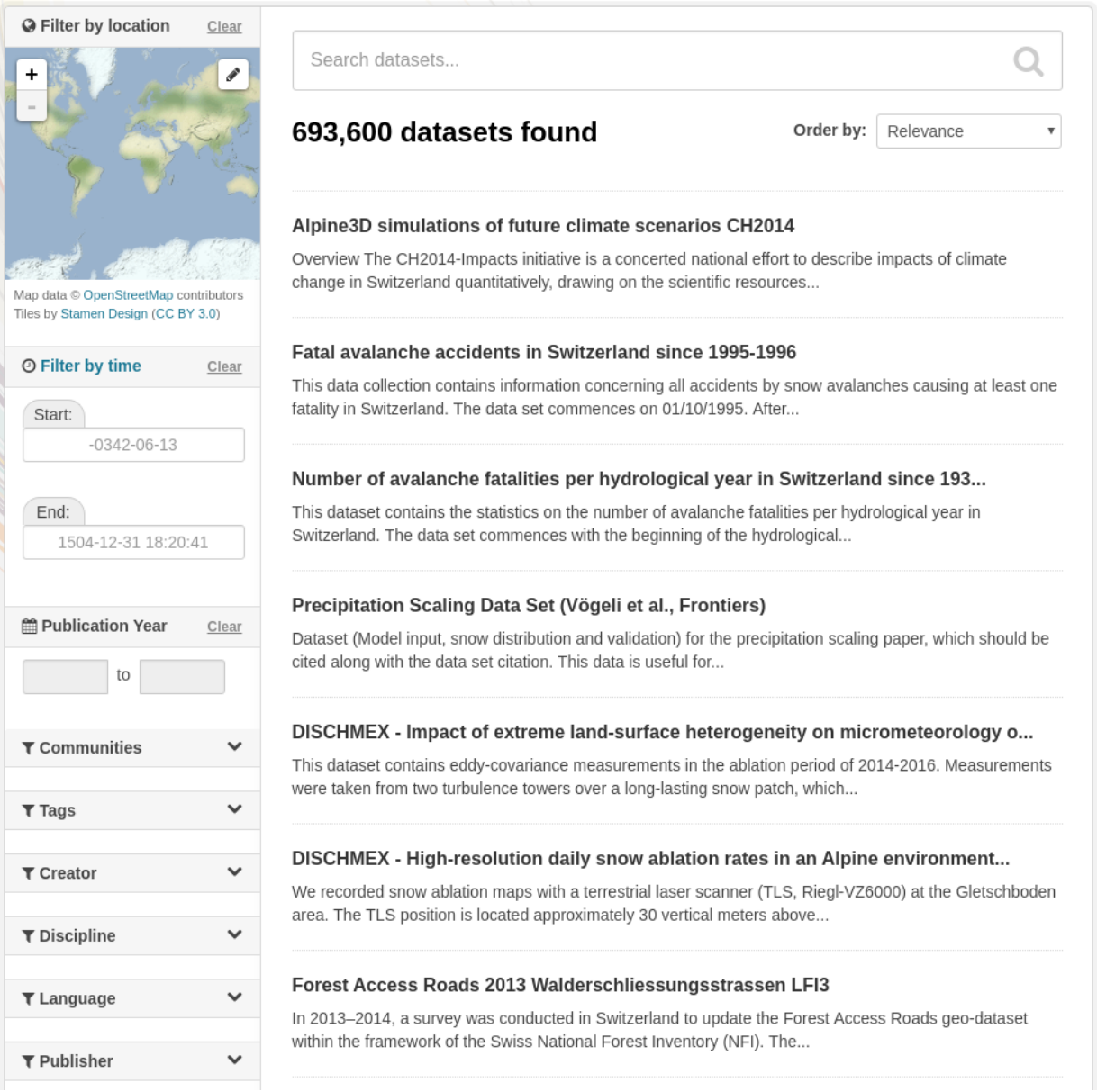
- (openaire.eu)
- EOSC-related initiative
- Multi-faceted
- Includes a Repository
  - And APIs

EUDAT

- (eudat.eu)
- EU funded initiative
  - Connected to EOSC



EUDAT


Filter by location
Clear
Map data © OpenStreetMap contributors
Tiles by Stamen Design (CC BY 3.0)
Filter by time
Clear
Start:
-0342-06-13
End:
1504-12-31 18:20:41
Publication Year
Clear
to
Communities
Tags
Creator
Discipline
Language
Publisher
Search datasets...
693,600 datasets found
Order by:
Relevance
Alpine3D simulations of future climate scenarios CH2014
Overview The CH2014-Impacts initiative is a concerted national effort to describe impacts of climate change in Switzerland quantitatively, drawing on the scientific resources...
Fatal avalanche accidents in Switzerland since 1995-1996
This data collection contains information concerning all accidents by snow avalanches causing at least one fatality in Switzerland. The data set commences on 01/10/1995. After...
Number of avalanche fatalities per hydrological year in Switzerland since 193...
This dataset contains the statistics on the number of avalanche fatalities per hydrological year in Switzerland. The data set commences with the beginning of the hydrological...
Precipitation Scaling Data Set (Vögeli et al., Frontiers)
Dataset (Model input, snow distribution and validation) for the precipitation scaling paper, which should be cited along with the data set citation. This data is useful for...
DISCHMEX - Impact of extreme land-surface heterogeneity on micrometeorology o...
This dataset contains eddy-covariance measurements in the ablation period of 2014-2016. Measurements were taken from two turbulence towers over a long-lasting snow patch, which...
DISCHMEX - High-resolution daily snow ablation rates in an Alpine environment...
We recorded snow ablation maps with a terrestrial laser scanner (TLS, Riegl-VZ6000) at the Gletschboden area. The TLS position is located approximately 30 vertical meters above...
Forest Access Roads 2013 Walderschliessungsstrassen LFI3
In 2013–2014, a survey was conducted in Switzerland to update the Forest Access Roads geo-dataset within the framework of the Swiss National Forest Inventory (NFI). The...

FAIRsharing

FAIRsharing.org
standards, databases, policies

Search all of FAIRsharing | Standards | Databases | Policies | Collections | Add/Claim Content | Stats | Log in or Register

A curated, informative and educational resource on data and metadata *standards*, inter-related to *databases* and data *policies*.

(fairsharing.org)

View as Table | View as Grid

Sort by: Best Match

Recommended Records: Recommended

Associated Publication?: No Publication | Has Publication

Claimed?: No Maintainer | Has Maintainer

Record Status: Uncertain | Deprecated | In developm | Ready

Standard Type

- Model/Format 21
- Reporting Guideline 19
- Terminology Artifact 3
- Identifier Schema 1

Domains

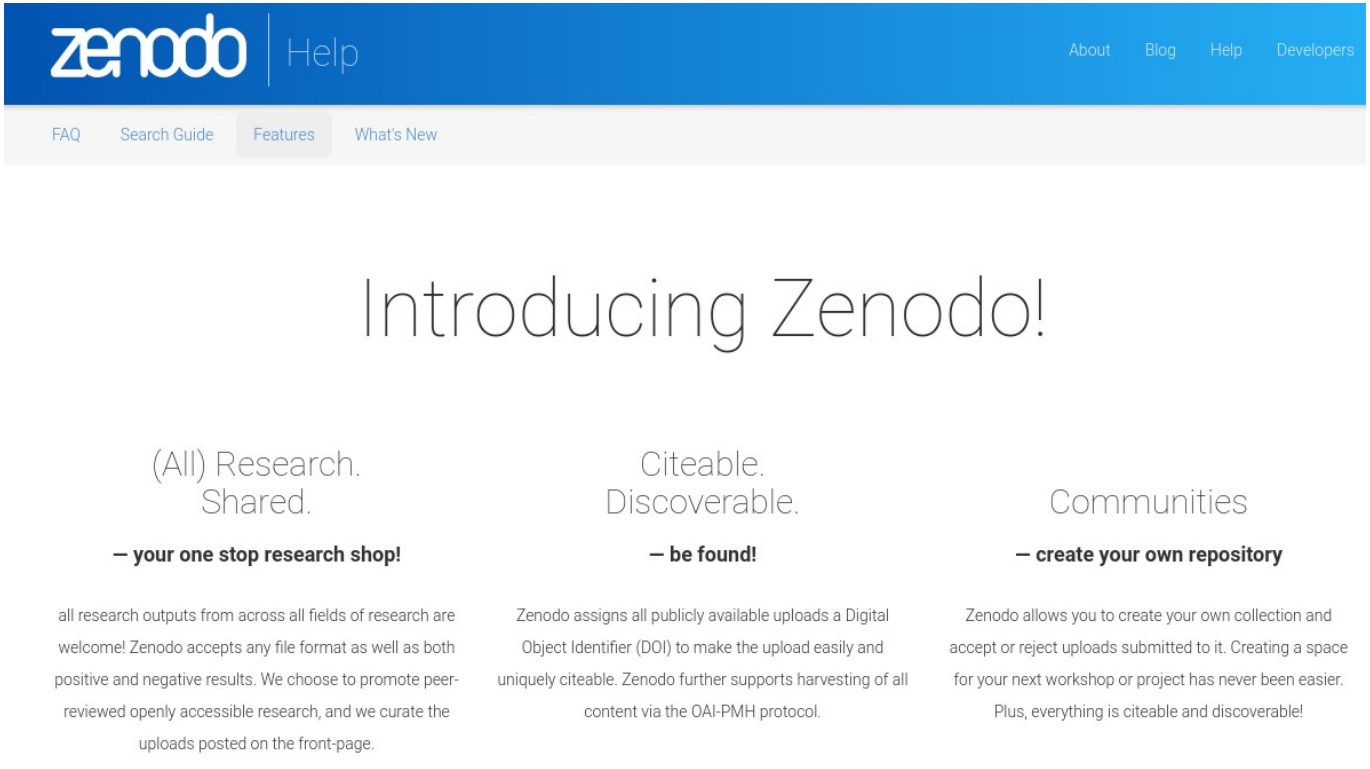
- Web Service 10
- Data Model 7
- Report 4
- Resource Metadata 4

44 records in view

Registry	Name	Abbreviation	Type	Subject	Domain	Taxonomy	Related Database	Related Standard	Related Policy	In Collection/Recommendation	Status
	IVOA Identifier	IVOA Identifier	Standard	Astrophysics And Astronomy	Centrally Registered Identifier	Not applicable	None	IVOA Registry Interfaces StandardsRegExt VOResource VODataService RegTAP Plus 2 more...	None	International Virtual Observatory Alliance (IVOA)	R
	IVOA Registry Interfaces	IVOA Registry Interfaces	Standard	Astrophysics And Astronomy	None	Not applicable	None	StandardsRegExt VOResource IVOA RM SimpleDALRegExt VODataService Plus 2 more...	None	International Virtual Observatory Alliance (IVOA)	R
	IVOA Standard for Unified Content Descriptors	IVOA UCD	Standard	Astrophysics And Astronomy	None	Not applicable	None	UCD1+ Vocabulary Vocabularies in the Virtual Observatory VOUnits	None	International Virtual Observatory Alliance (IVOA)	R
	IVOA Credential Delegation Protocol	IVOA Credential Delegation Protocol	Standard	Astrophysics And Astronomy	Web Service	Not applicable	None	IVOA Web Services Basic Profile VOSI PDL VOSpace UWS Plus 1 more...	None	International Virtual Observatory Alliance (IVOA)	R
	~~IVOA Web Services Basic Profile~~	~~IVOA Web Services Basic Profile~~	Standard	Astrophysics And Astronomy	Web Service	Not applicable	None	PDL SSO - Authentication VOSpace IVOA Credential Delegation Protocol UWS Plus 1 more...	None	International Virtual Observatory Alliance (IVOA)	D
	IVOA Support Interfaces	VOSI	Standard	Astrophysics And Astronomy	Web Service	Not applicable	None	PDL VOSpace UWS IVOA Web Services Basic Profile SimpleDALRegExt Plus 2 more...	None	International Virtual Observatory Alliance (IVOA)	R

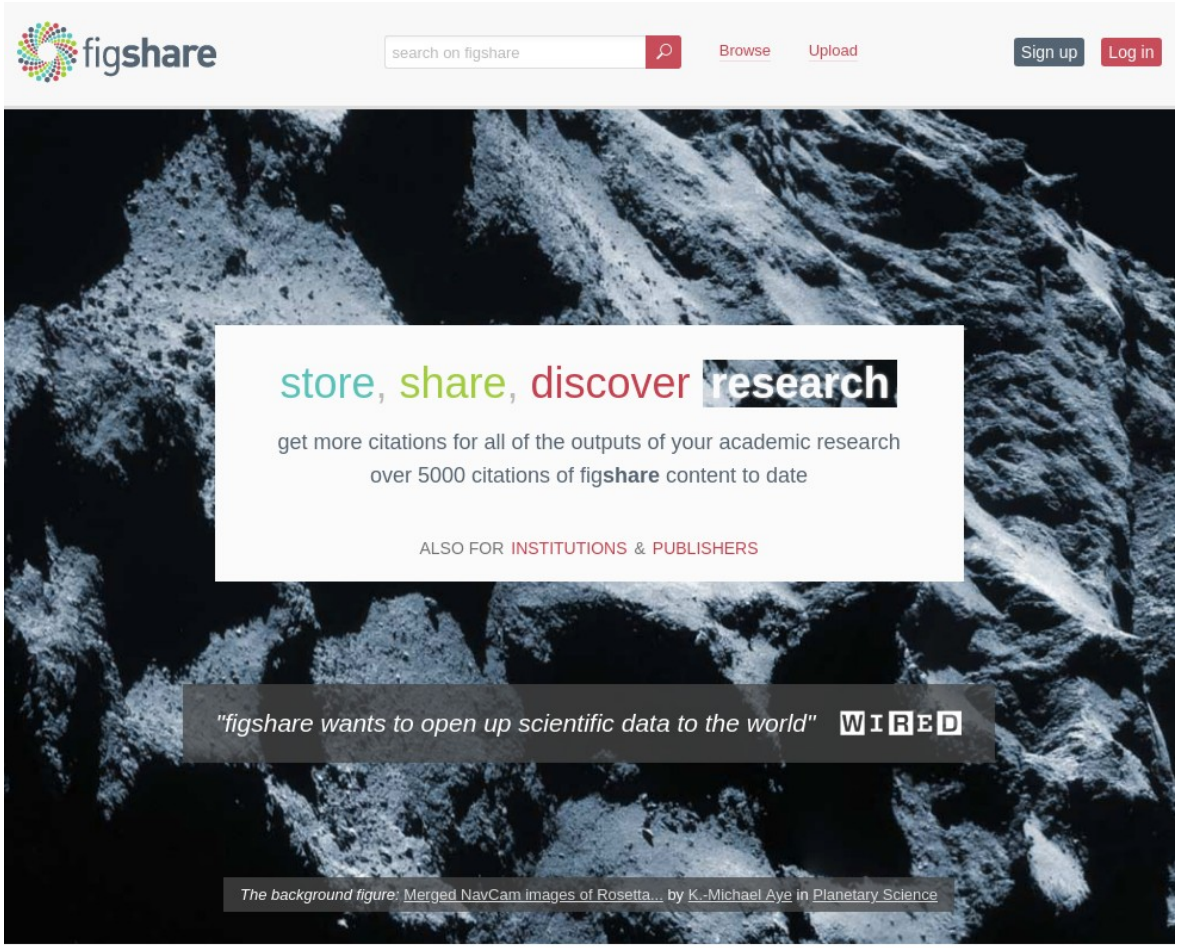
Zenodo

- (zenodo.org and about.zenodo.org/)
  - https://zenodo.org/record/802100#.YbshJ9so_RY
  - CERN based
  - OpenAIRE connected
  - General content repository, mainly papers/proceedings/presentations...



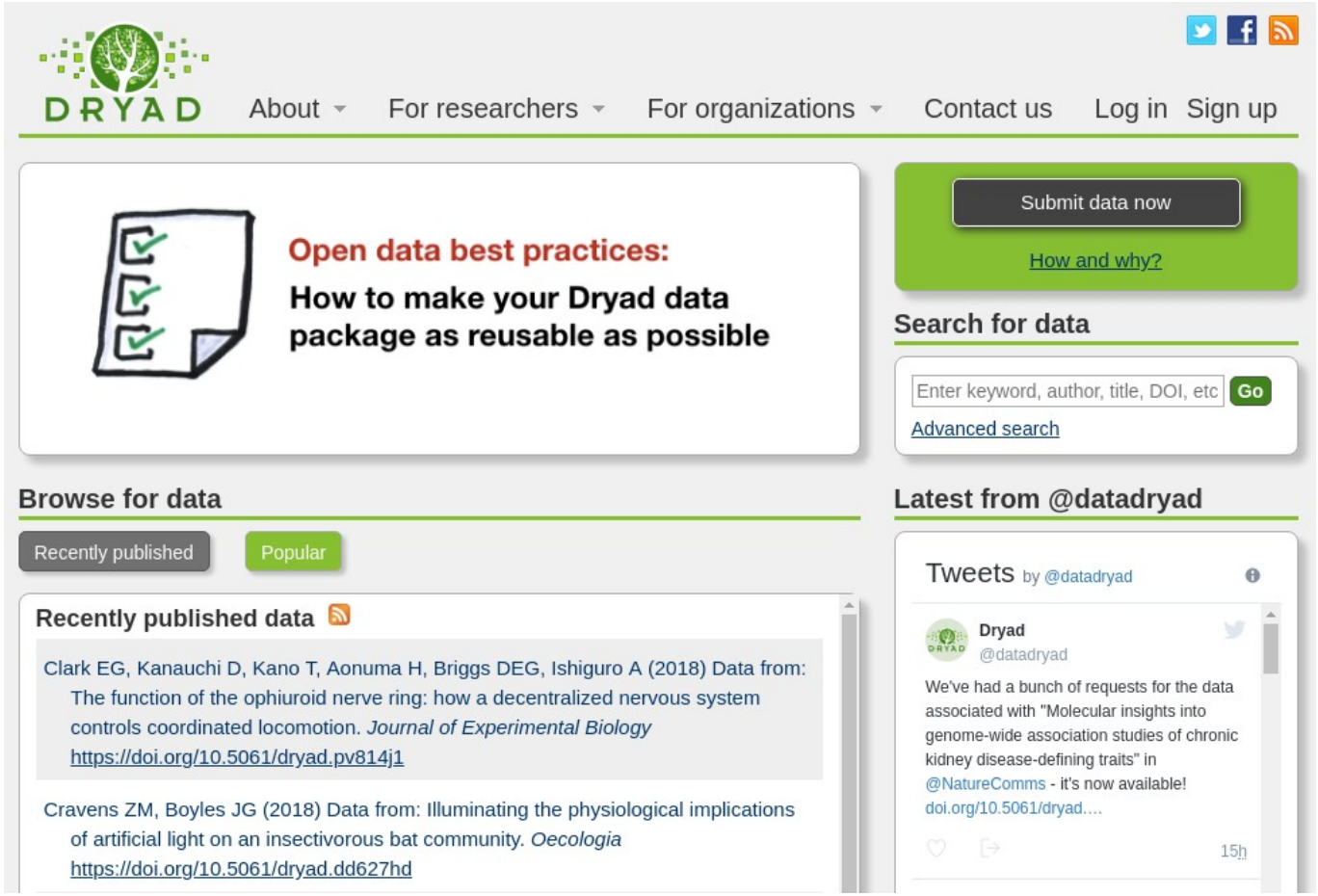
figshare

- (figshare.com)
- Publication driven
- Support service
- [upload limits]



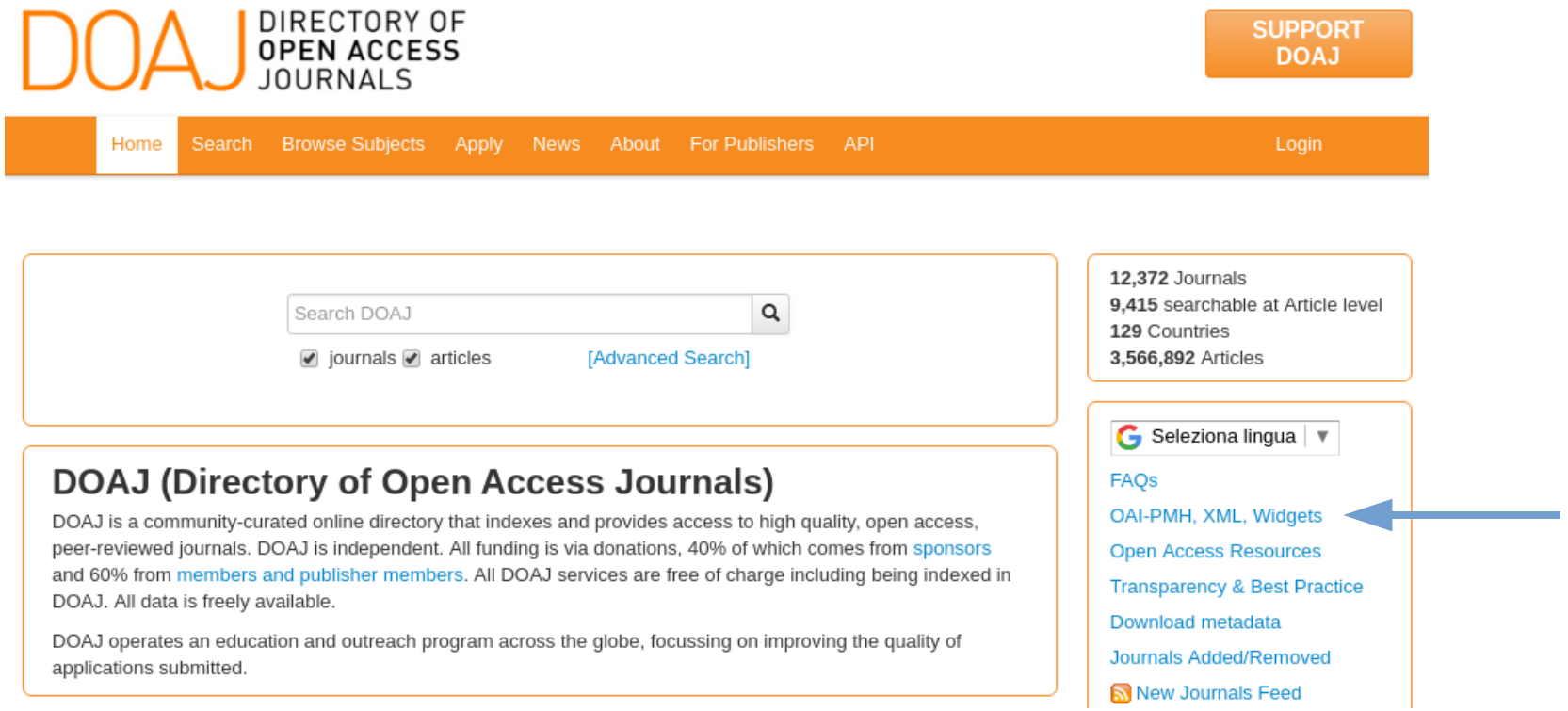
Dryad

- (datadryad.org)
  - The Dryad Digital Repository is a curated resource that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad provides a general-purpose home for a wide diversity of datatypes.
    - Nonprofit



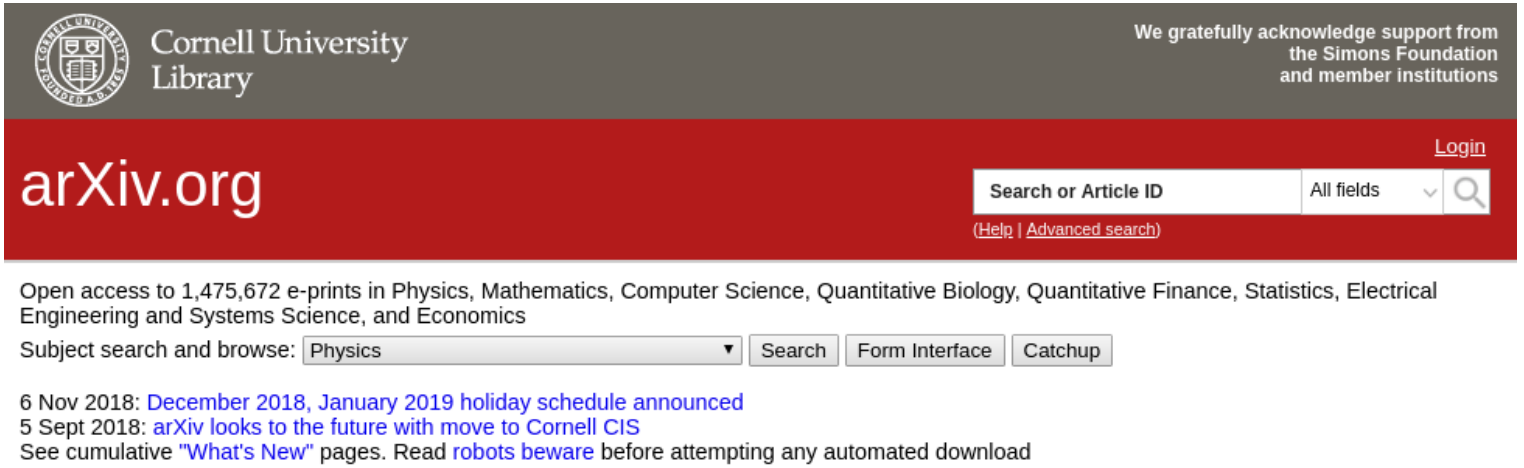
DOAJ

- Directory of Open Access Journals
  - (doaj.org)
  - Journals as primary records
    - Articles out of them



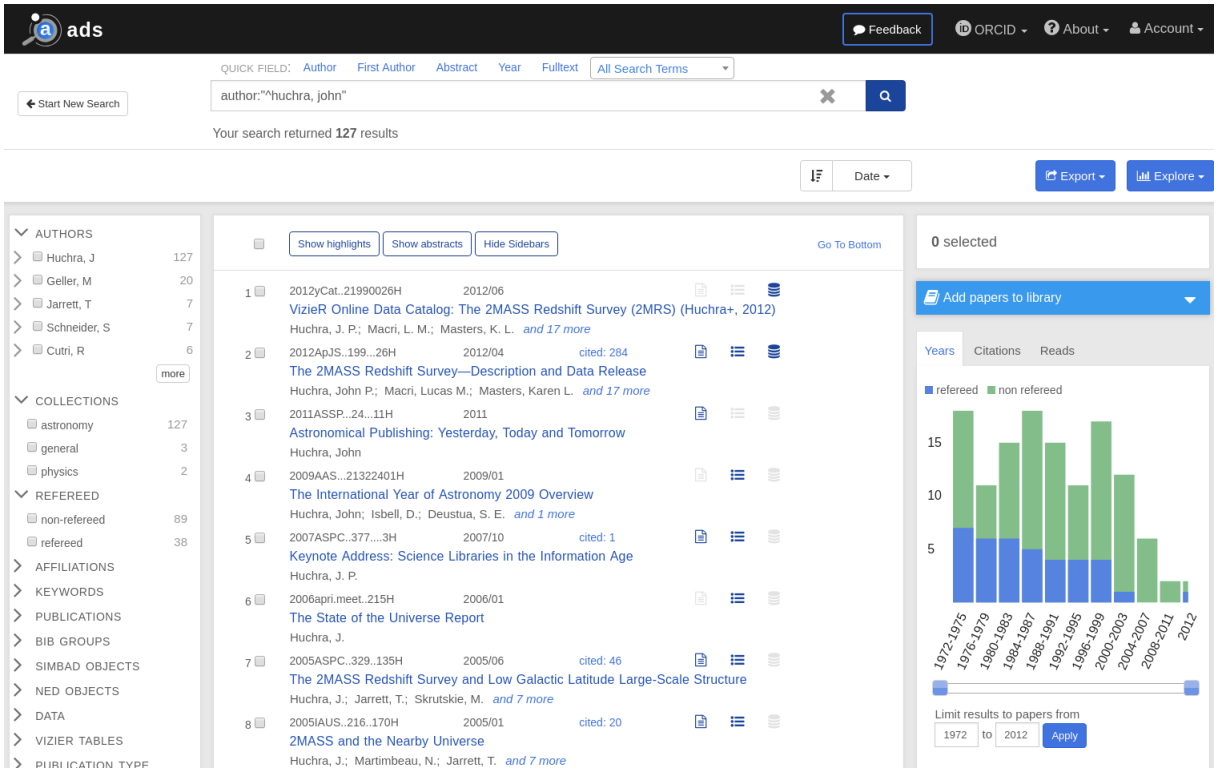
arXiv.org

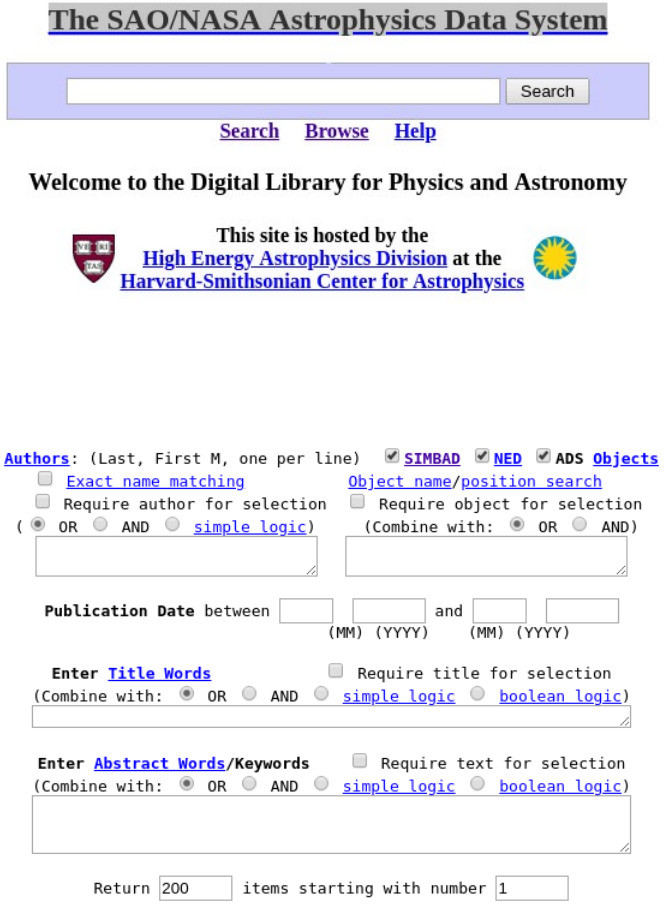
- Open access to 1,475,672 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance, Statistics, Electrical Engineering and Systems Science, and Economics
  - (arxiv.org)
- User driven repository
- Used mainly for pre-prints and ongoing work
  - Author-paper relationship “limited”



ADS

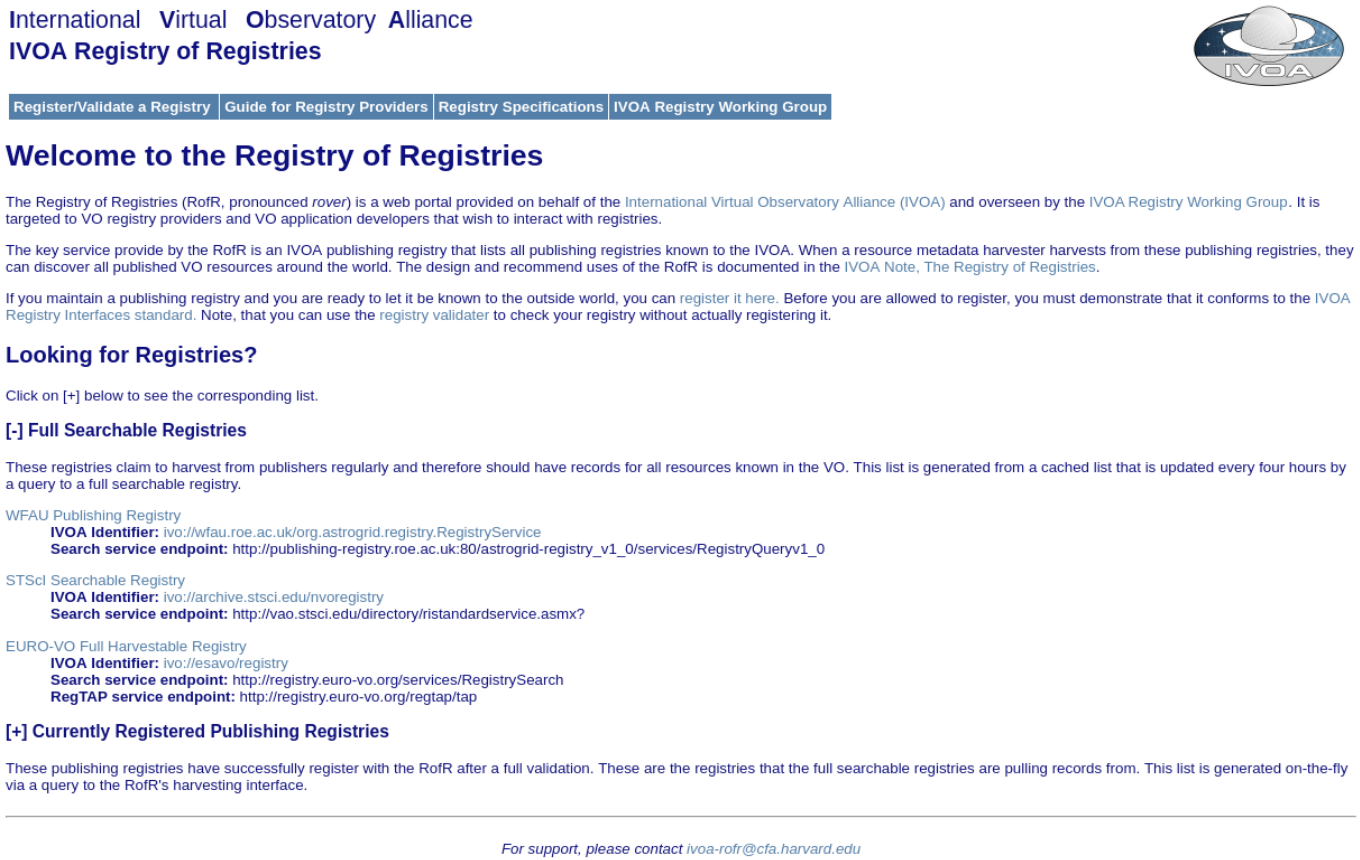
- The SAO/NASA Astrophysics Data System
  - (adsabs.harvard.edu)





RofR (IVOA)

- Registry of Registries
  - (rofr.ivoa.net)
- Distributed
- Domain driven
- API based
  - GUI exist
    - Limited



International Virtual Observatory Alliance
IVOA Registry of Registries

Register/Validate a Registry | Guide for Registry Providers | Registry Specifications | IVOA Registry Working Group

Welcome to the Registry of Registries

The Registry of Registries (RofR, pronounced *rover*) is a web portal provided on behalf of the International Virtual Observatory Alliance (IVOA) and overseen by the IVOA Registry Working Group. It is targeted to VO registry providers and VO application developers that wish to interact with registries.

The key service provide by the RofR is an IVOA publishing registry that lists all publishing registries known to the IVOA. When a resource metadata harvester harvests from these publishing registries, they can discover all published VO resources around the world. The design and recommend uses of the RofR is documented in the IVOA Note, The Registry of Registries.

If you maintain a publishing registry and you are ready to let it be known to the outside world, you can register it here. Before you are allowed to register, you must demonstrate that it conforms to the IVOA Registry Interfaces standard. Note, that you can use the registry validater to check your registry without actually registering it.

Looking for Registries?

Click on [+] below to see the corresponding list.

[-] Full Searchable Registries

These registries claim to harvest from publishers regularly and therefore should have records for all resources known in the VO. This list is generated from a cached list that is updated every four hours by a query to a full searchable registry.

WFAU Publishing Registry
IVOA Identifier: ivo://wfau.roe.ac.uk/org.astrogrid.registry.RegistryService
Search service endpoint: http://publishing-registry.roe.ac.uk:80/astrogrid-registry_v1_0/services/RegistryQueryv1_0

STScI Searchable Registry
IVOA Identifier: ivo://archive.stsci.edu/nvoregistry
Search service endpoint: http://vao.stsci.edu/directory/ristandardservice.asmx?

EURO-VO Full Harvestable Registry
IVOA Identifier: ivo://esavo/registry
Search service endpoint: http://registry.euro-vo.org/services/RegistrySearch
RegTAP service endpoint: http://registry.euro-vo.org/regtap/tap

[+] Currently Registered Publishing Registries

These publishing registries have successfully register with the RofR after a full validation. These are the registries that the full searchable registries are pulling records from. This list is generated on-the-fly via a query to the RofR's harvesting interface.

For support, please contact ivoa-rofr@cfa.harvard.edu

Do still exist single “archives”?

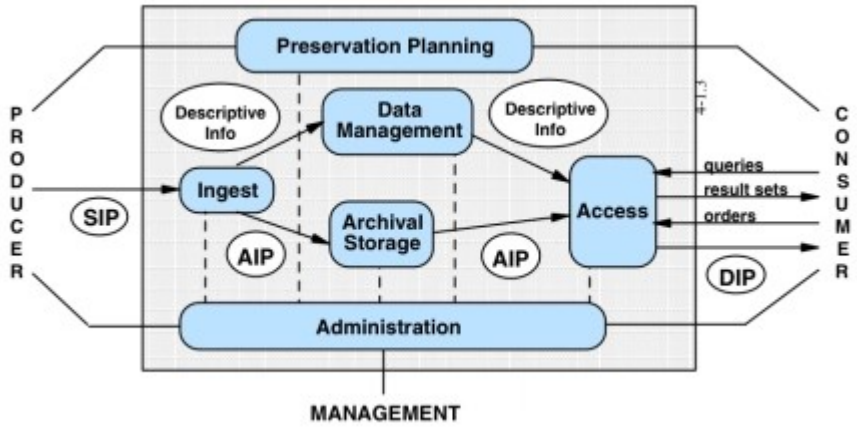
- A lot
  - Domain specific
  - Custom based
  - Some trying to reach open interoperation
    - Depending on the domain (usually)
- They’re usually listed as items inside more general repositories
- They can provide quite powerful analysis/discovery interfaces
  - Because of the specific scenario they expose
- Some astrophysics VO examples
  - ESA (http://sky.esa.int/)
  - ESO (http://archive.eso.org/scienceportal/home)
  - MAST (https://archive.stsci.edu/)
  - CADC (http://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/)

Repository Certification

- National and international funders will (likely) mandate
  - Open data
  - Data management policies for the long-term storage and accessibility of data
  - Access to funded data products and proper data management plans (DMP)
- Need to store shared data in a trustworthy data repository
  - managed, curated, and archived to preserve the initial investment in collecting them
- Sustainability of repositories raises a number of challenging issues in different areas
  - Organizational, technical, financial, legal, etc.
- Certification can be an important contribution to ensuring the reliability and durability of data repositories
- By becoming certified, repositories can demonstrate to both their users and their funders that an independent authority has evaluated them and endorsed their trustworthiness.

Core Trust Seal

- (www.coretrustseal.org)
  - Core certification involves a minimally intensive process whereby data repositories supply evidence that they are sustainable and trustworthy
  - A repository first conducts an internal self-assessment, which is then reviewed by community peers
  - Such assessments help data communities to improve the quality and transparency of their processes, and to increase awareness of and compliance with established standards
  - This community approach guarantees an inclusive atmosphere in which the candidate repository and the reviewers closely interact
- https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf



Open Archive Information System