The Ecosystem for Data Discovery

Alessia Bardi

ISTI-CNR / OpenAIRE / GoFAIR Discovery Implementation Network

alessia.bardi@isti.cnr.it

Overview

- The GOFAIR Discovery Implementation Network
  - The ecosystem for data discovery
  - Gaps and barriers to data discovery
- OpenAIRE
  - OpenAIRE's actions for data discovery
  - Gateways for research communities (OpenAIRE CONNECT)

GO FAIR

DISCOVERY

OPEN KNOWLEDGE MAPS

AfricArXiv BASE Bioschemas cessda CORE

DataCite DARIAH-EU elixir FRANCE EUDAT gesis HZI HELMHOLTZ Zentrum für Infektionsforschung

Alexander von Humboldt INSTITUTE FOR INTERNET AND SOCIETY hypothes.is Impactstory IGB KNOW Center net7 NIOO

OpenAIRE OPERAS ORCID Scholia TIB ZBW ZB MED

Motivation

Up to 85% of datasets are not reused (Peters et al. 2016)

□ Discoverability is a key challenge when it comes to research data

Lack of adequate user interfaces for data discovery

- Simple reuse of existing interface concepts for publications
- Design from the system‘s rather than the user‘s perspective

New market entrants following a closed/proprietary model

- Not suitable for the Internet of FAIR Data and Services
- Creates new (pay)walls and prevents innovation

Peters I. et al. Research data explored: an extended analysis of citations and altmetrics. Scientometrics. 2016;107:723-744. doi: 10.1007/s11192-016-1887-4

A descriptive framework of the open ecosystem for research data discovery

GOALS

- Supporting the realisation of data discovery tools
- Understand opportunities for innovative solutions contributing to the evolution of the open ecosystem

HOW

- Define the building blocks of the ecosystem, their interactions, and the discovery needs they cover
- Identify gaps of current practices of research data discovery
  - Gaps of the infrastructure
  - Gaps in users' search strategies

Bardi, A. et al. (2022). The Open Ecosystem of e-Infrastructures for Data Discovery: A Review. Zenodo. https://doi.org/10.5281/zenodo.7468089

Open e-infrastructures for data discovery

The e-infrastructures for data discovery with clear and established **open policies**, **open APIs**, and **open licenses** for data, metadata and source code, allowing for **community governance.**

Open e-infrastructures r**emove paywalls**, **avoid lock-in** effects and enable **community participation and outreach.**

OpenAIRE

OPEN KNOWLEDGE MAPS

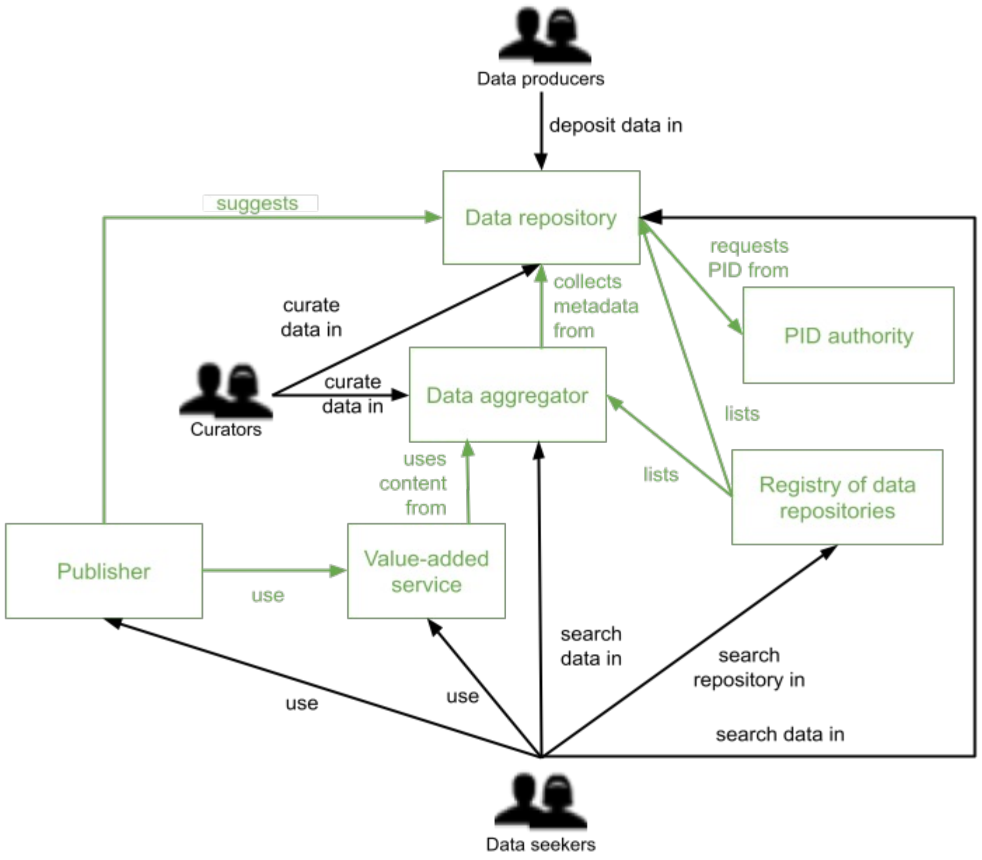
TextGrid Repository

PANGAEA.

B2FIND

zenodo

The building blocks of the open ecosystem for data discovery



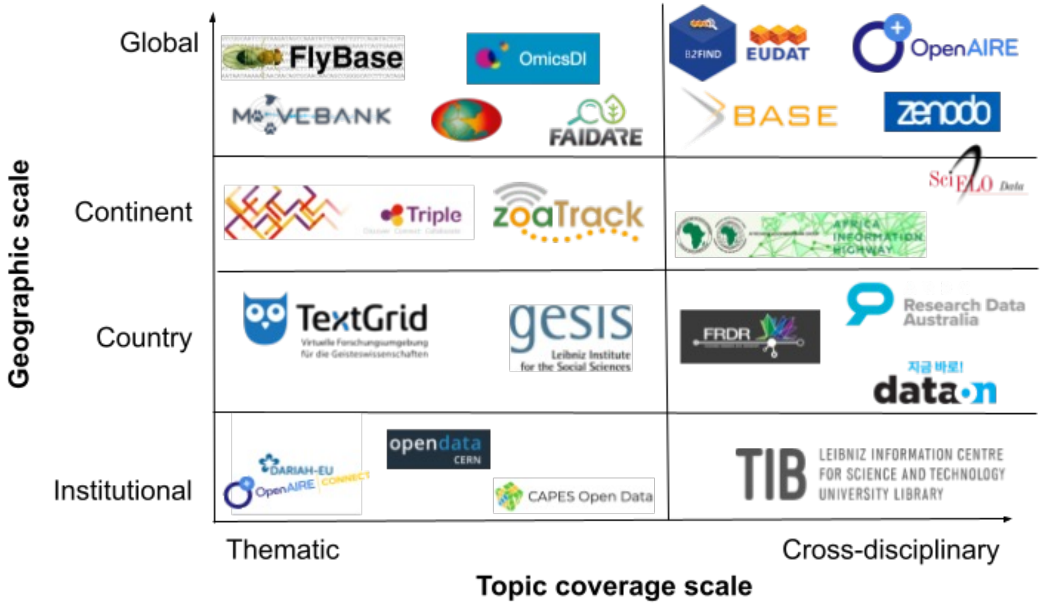
Variety

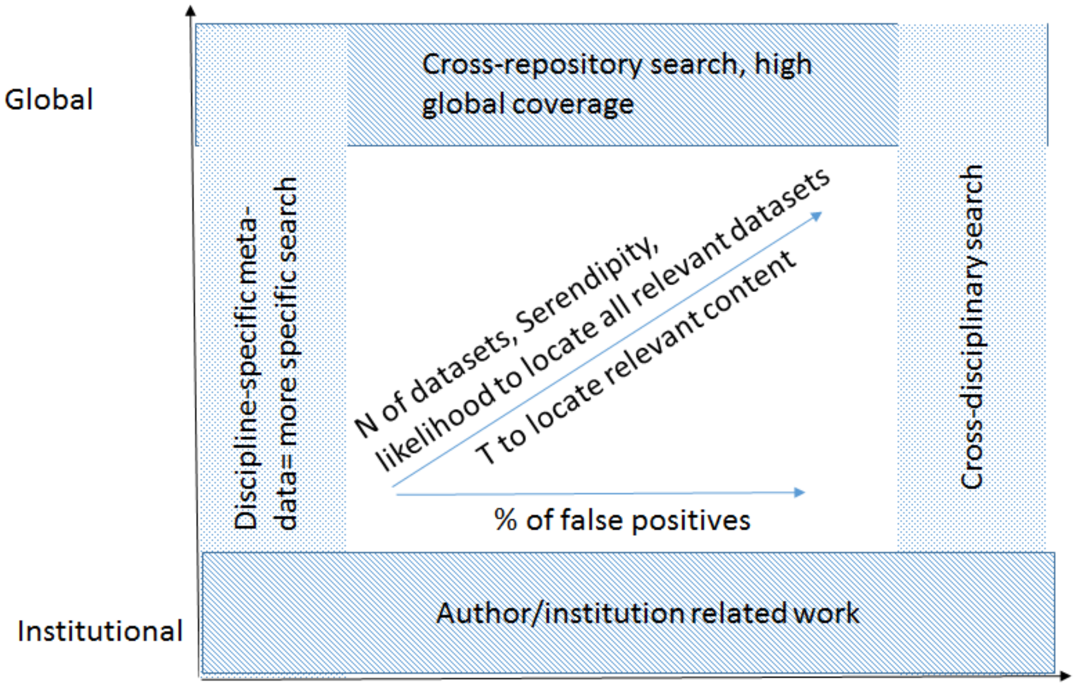
Openness

Dinamicity

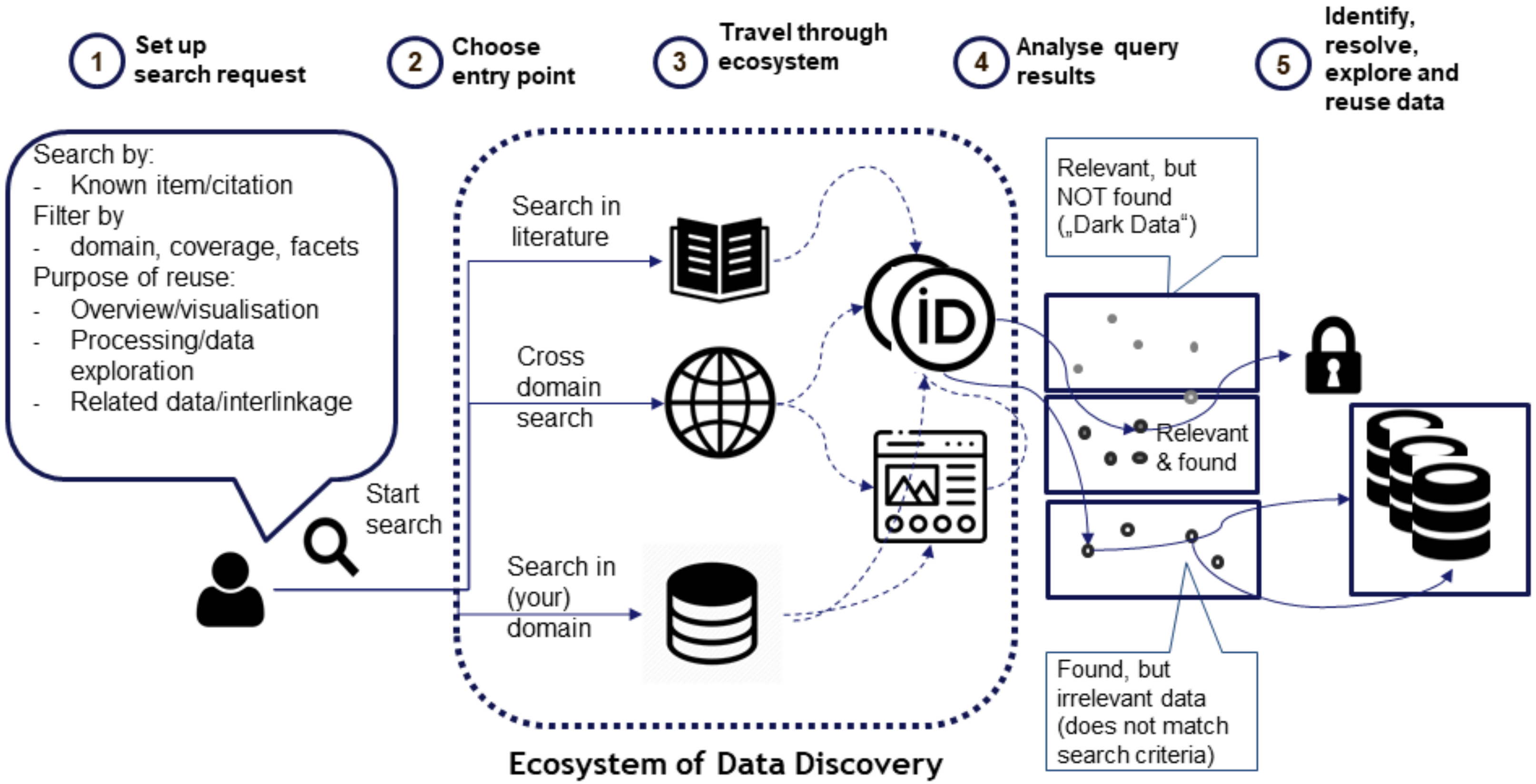
- **Researchers** use the open infrastructure to share their data, make it discoverable and to discover data
- **Publishers** suggest trusted repositories for the deposition of data that support published articles
- **Data repositories** host metadata and files
- **Data aggregators** harvest data or metadata collected from data repositories
- **Registries of data repositories** are directories intended to provide an organised, up-to-date and searchable collection of data repositories
- **PID authorities** offer services for registering persistent and resolvable identifiers to entities
- **Discovery services** offer a front-end (e.g. portal) for the discovery functionality over a set of data.
- **Value-added services** re-use content (files and metadata) hosted elsewhere, expand the ecosystem with innovative discovery services that go beyond the traditional keyword based or browse searches

Data dicovery services: benefits and limitations




Global
Institutional
Cross-repository search, high global coverage
Discipline-specific meta-data= more specific search
Cross-disciplinary search
N of datasets, Serendipity,
likelihood to locate all relevant datasets
T to locate relevant content
% of false positives
Author/institution related work

The Data discovery journey



1 Set up search request
2 Choose entry point
3 Travel through ecosystem
4 Analyse query results
5 Identify, resolve, explore and reuse data
Search by:
- Known item/citation
Filter by
- domain, coverage, facets
Purpose of reuse:
- Overview/visualisation
- Processing/data exploration
- Related data/interlinkage
Start search
Search in literature
Cross domain search
Search in (your) domain
iD
Ecosystem of Data Discovery
Relevant, but NOT found („Dark Data")
Relevant & found
Found, but irrelevant data (does not match search criteria)

Gaps in research data discovery

Gap	Search strategy deficiency	Data infrastructure deficiency
1 - Unstructured or missing search strategies	Missing overview of data discovery ecosystem, lack of data search literacy, missing search strategy, imprecise search terms on researcher's side.	Missing metadata, coarse data granularity, missing search facets/filter possibilities.
2 - Inadequate user interfaces		Missing user involvement, lack of innovative features, proprietary licences prohibiting reuse of software and services
3 - Lack of interoperable and interconnected discovery ecosystem		Lack of technical interoperability, organisational cooperation and metadata interoperability between repositories and aggregators, closed and proprietary indexes
4 - Low recall or low precision	Unfitting repository/search engine, imprecise search terms	Suboptimally configured search engine, missing metadata, missing filter options
5 - Problems with identification, access and reuse of data	Misinterpretation or misuse of data found	Missing provenance, missing information of licences and restrictions

Supporting research data discovery in OpenAIRE

EXPLORE

Global cross-disciplinary discovery service for metadata available in the OpenAIRE Graph

https://explore.openaire.eu

CONNECT

Service for the realisation of customisable discovery portals for research communities

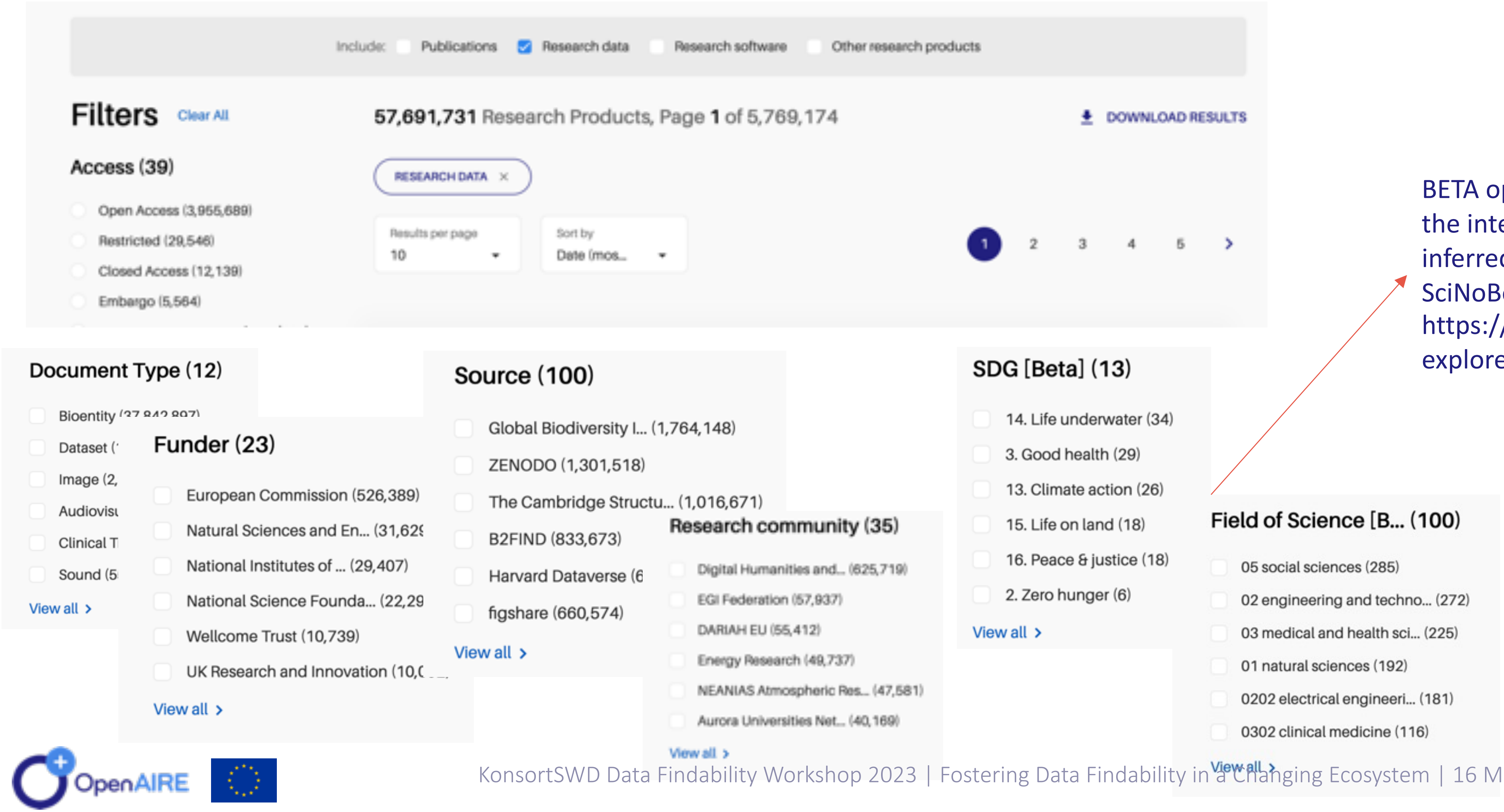
https://connect.openaire.eu

Include research literature, software, data and other entities

Searching research data in OpenAIRE EXPLORE and CONNECT

- Simple search (keyword search on any metadata field)
- Advanced search (specific terms in specific fields, specific terms NOT in specific fields, also in AND/OR)
- Find research data related to a publication or software
- Faceted search by:
  - Access right
  - Publication year
  - Type of data (e.g. image, bioentity, sound)
  - Funding
  - Country (based on authors' affiliations and/or institutional repositories of provenance)
  - Language (of the metadata)
  - Hosting repository
  - Fields of science and sustainable development goals of reference
  - Related research communities

Research Data in the OpenAIRE Graph



Include: □ Publications ☑ Research data □ Research software □ Other research products

Filters Clear All

57,691,731 Research Products, Page **1** of 5,769,174

DOWNLOAD RESULTS

RESEARCH DATA ×

Results per page: 10

Sort by: Date (mos...

Access (39)

- ○ Open Access (3,955,689)
- ○ Restricted (29,546)
- ○ Closed Access (12,139)
- ○ Embargo (5,564)

Document Type (12)

- □ Bioentity (37,842,897)
- □ Dataset (
- □ Image (2,
- □ Audiovisu
- □ Clinical T
- □ Sound (5

View all ›

Funder (23)

- □ European Commission (526,389)
- □ Natural Sciences and En... (31,62
- □ National Institutes of ... (29,407)
- □ National Science Founda... (22,29
- □ Wellcome Trust (10,739)
- □ UK Research and Innovation (10,(

View all ›

Source (100)

- □ Global Biodiversity I... (1,764,148)
- □ ZENODO (1,301,518)
- □ The Cambridge Structu... (1,016,671)
- □ B2FIND (833,673)
- □ Harvard Dataverse (6
- □ figshare (660,574)

View all ›

Research community (35)

- □ Digital Humanities and... (625,719)
- □ EGI Federation (57,937)
- □ DARIAH EU (55,412)
- □ Energy Research (49,737)
- □ NEANIAS Atmospheric Res... (47,581)
- □ Aurora Universities Net... (40,169)

View all ›

SDG [Beta] (13)

- □ 14. Life underwater (34)
- □ 3. Good health (29)
- □ 13. Climate action (26)
- □ 15. Life on land (18)
- □ 16. Peace & justice (18)
- □ 2. Zero hunger (6)

View all ›

Field of Science [B... (100)

- □ 05 social sciences (285)
- □ 02 engineering and techno... (272)
- □ 03 medical and health sci... (225)
- □ 01 natural sciences (192)
- □ 0202 electrical engineeri... (181)
- □ 0302 clinical medicine (116)

View all ›

BETA options: we are experimenting the integration of FoS and SDG inferred by the AI algorithms of SciNoBo (Athena Research Center) https://www.openaire.eu/openaire-explore-introducing-sdgs-and-fos

Challenges and approaches for data discovery

- Visibility of data repositories
  - Search for data sources from different registries
- Poor metadata of research data
  - Putting research in context: inference of links between data, publications, software and other types of research products
  - Enriching dataset metadata based on related publications (with typically richer metadata)
- Continously monitoring portals’ usage with Matomo and Google Analytics
- Involving users in the loop
  - Community calls with repository managers and research communities
- With OpenAIRE CONNECT we create customisable portal where only the relevant subset of the OpenAIRE Graph is searchable
  - What does ”relevant” mean? Experts of the community decide

Supporting communities in the data deluge with OpenAIRE CONNECT?

A portal with:

- A customised view of the Open Research Graph
- Complete branding capabilities
- Discovery and monitoring functionalities

Facilitator of OS practices

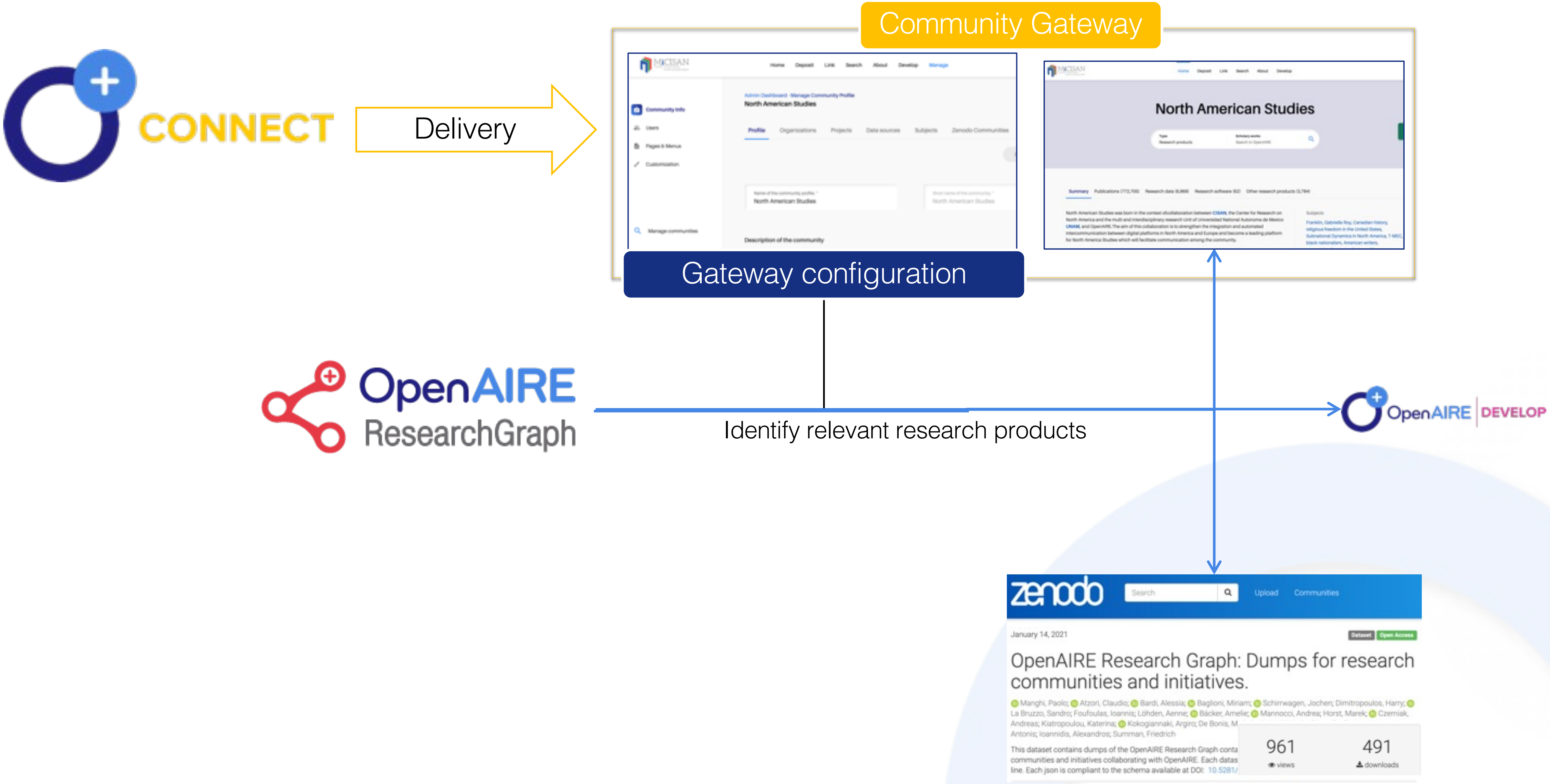
- Linking & claiming research
- Finding repositories
- Sharing

Integrated with:

- Other OpenAIRE services: Zenodo, UsageCounts, EXPLORE
- ORCID
- Global AAI Standard

A service on demand, operated by OpenAIRE:

- Data updates on a regular basis
- Full IT support of the service: installation, maintenance, upgrade, backups



MiCISAN

Home Deposit Link Search About Develop Manage

Community info
Users
Pages & Menus
Customization
Manage communities

Admin Dashboard · Manage Community Profile

North American Studies

Profile Organizations Projects Data sources Subjects Zenodo Communities

Name of the community profile: *
North American Studies

Short name of the community: *
North American Studies

Description of the community

MiCISAN

Home Deposit Link Search About Develop

North American Studies

Type: Research products
Scholarly works: Search in OpenAIRE

Summary Publications (772,705) Research data (8,868) Research software (82) Other research products (3,784)

North American Studies was born in the context of collaboration between CISAN, the Center for Research on North America and the multi and interdisciplinary research Unit of Universidad National Autonoma de Mexico UNAM, and OpenAIRE. The aim of this collaboration is to strengthen the integration and automated intercommunication between digital platforms in North America and Europe and become a leading platform for North America Studies which will facilitate communication among the community.

Subjects
Franklin, Gabrielle Roy, Canadian history, religious freedom in the United States, Subnational Dynamics in North America, T-MEC, black nationalism, American writers,

zenodo Search Upload Communities

January 14, 2021 Dataset Open Access

OpenAIRE Research Graph: Dumps for research communities and initiatives.

Manghi, Paolo; Atzori, Claudio; Bardi, Alessia; Baglioni, Miriam; Schirrwagen, Jochen; Dimitropoulos, Harry; La Bruzzo, Sandro; Foufoulas, Ioannis; Löhden, Aenne; Bäcker, Amelie; Mannocci, Andrea; Horst, Marek; Czerniak, Andreas; Kiatropoulou, Katerina; Kokogiannaki, Argiro; De Bonis, M Antonis; Ioannidis, Alexandros; Summan, Friedrich

This dataset contains dumps of the OpenAIRE Research Graph conta communities and initiatives collaborating with OpenAIRE. Each datas line. Each json is compliant to the schema available at DOI: 10.5281/

961 views 491 downloads

https://doi.org/10.5281/zenodo.3974604

Community Gateway content configuration in more detail

Gateway curators

Keywords — Discipline-specific subject terms to be found in the metadata

Projects — From the 22 funders integrated in OpenAIRE

Data sources — E.g. thematic repositories, archives and journals

Zenodo communities — With research products relevant for the field

Organizations — Organizations in the field: we look for them in the affiliations

Researchers

Link — Link products in OpenAIRE, Crossref, Datacite, ORCID to the community, also in bulk

OpenAIRE algorithms

Propagation

If a community result is supplemented by another research product, then the latter is also added in the community gateway

Full-text mining

Links to projects
Affiliations
Document classification
. . .

Lessons learned

- Thanks to the GOFAIR Discovery IN there is a clearer way of the different agents and their interactions in the ecosystem for data discovery
- Involving users is important for better discovery services
- No one size fits all: there is no BEST discovery service because no discovery service can effectivly implement each and every data discovery journey
- Thanks to its openness, the open ecosystem is dynamic and vital for the the realisation of new, advanced services

What’s going on

- OpenAIRE is working on intelligent search (FAIRCORE4EOSC) and methods to enrich dataset metadata
- Open Knowledge Maps is adapting its visual discovery approach to the specific features of datasets
- RDA WG on interoperoperability of Scientific Knowledge Graph
- European Open Science Cloud Interoperability Framework
- . . .

Thank you!

Alessia Bardi

alessia.bardi@isti.cnr.it

OpenAIRE-Nexus has received funding from the
pean Union's Horizon 2020 Research and Innovation
programme under Grant Agreements No. 101017452