Alessia Bardi

CNR

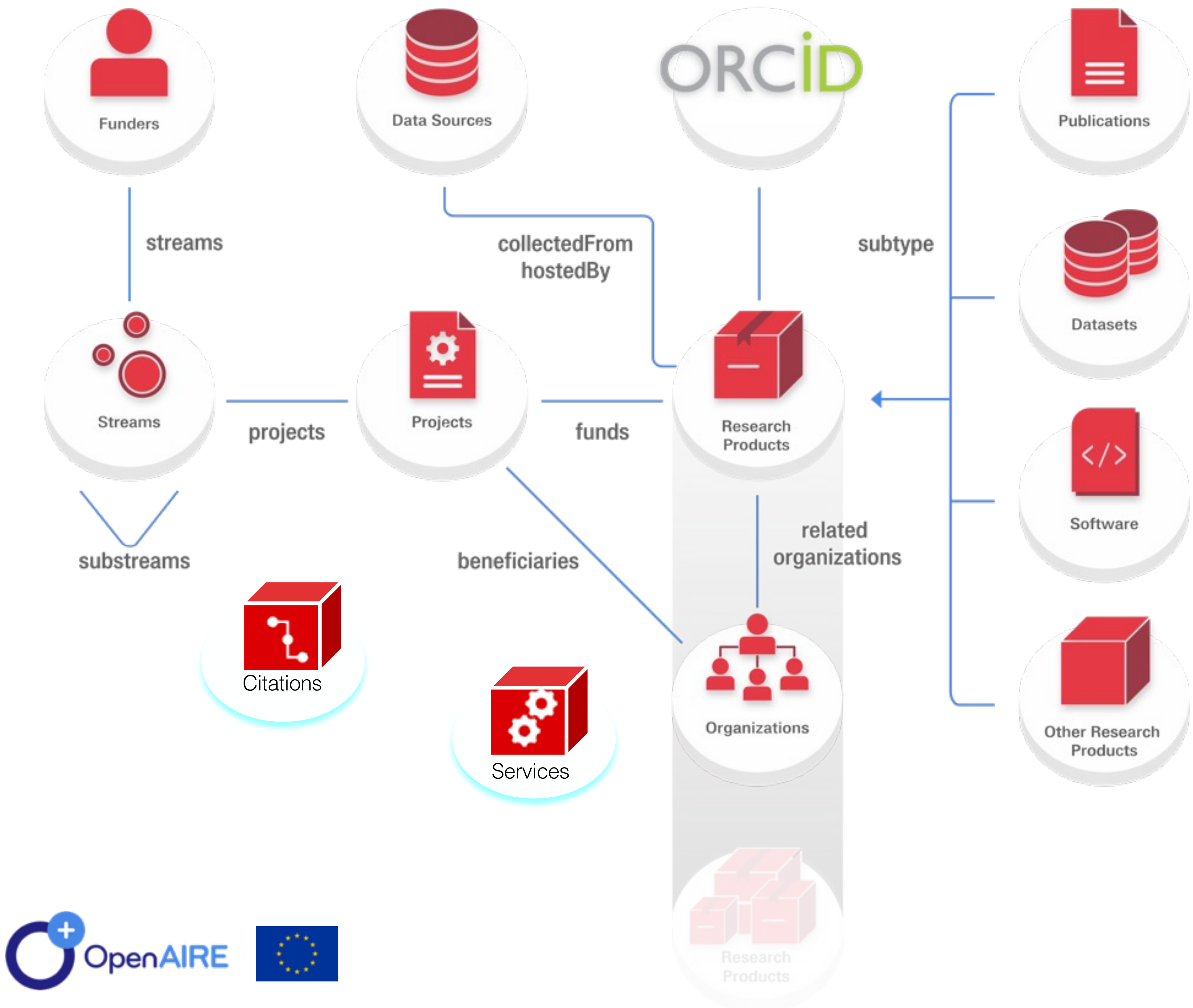
Vocabularies in the OpenAIRE Graph

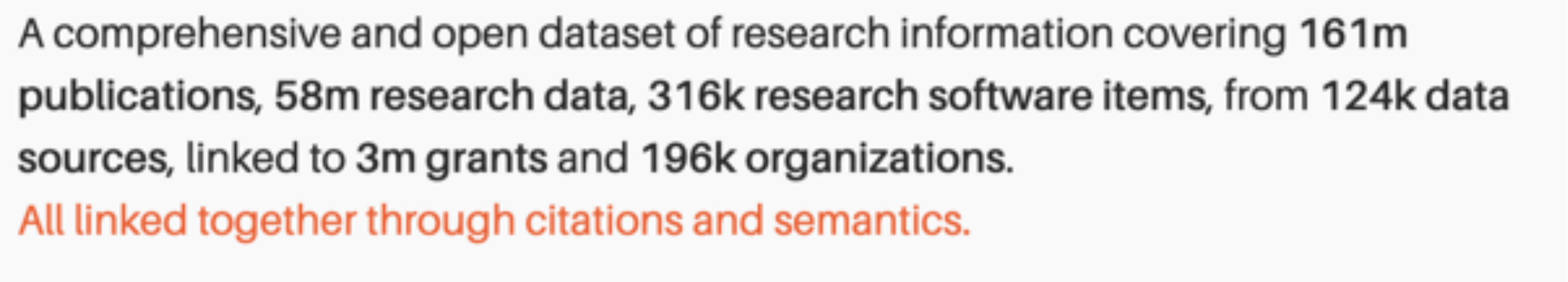
Overview

- The OpenAIRE Graph in a nutshell
- Controlled vocabularies
- Subject classification
- Usage of subjects in OpenAIRE
  - EXPLORE
  - CONNECT (& collaboration with GOTRIPLE)
  - MONITOR

OPENAIRE GRAPH – KEY ASSET IN OPEN SCIENCE INFRASTRUCTURE

INCLUSIVE, TRANSPARENT, REPRODUCIBLE





A comprehensive and open dataset of research information covering **161m publications**, **58m research data**, **316k research software items**, from **124k data sources**, linked to **3m grants** and **196k organizations**.

All linked together through citations and semantics.

Global coverage

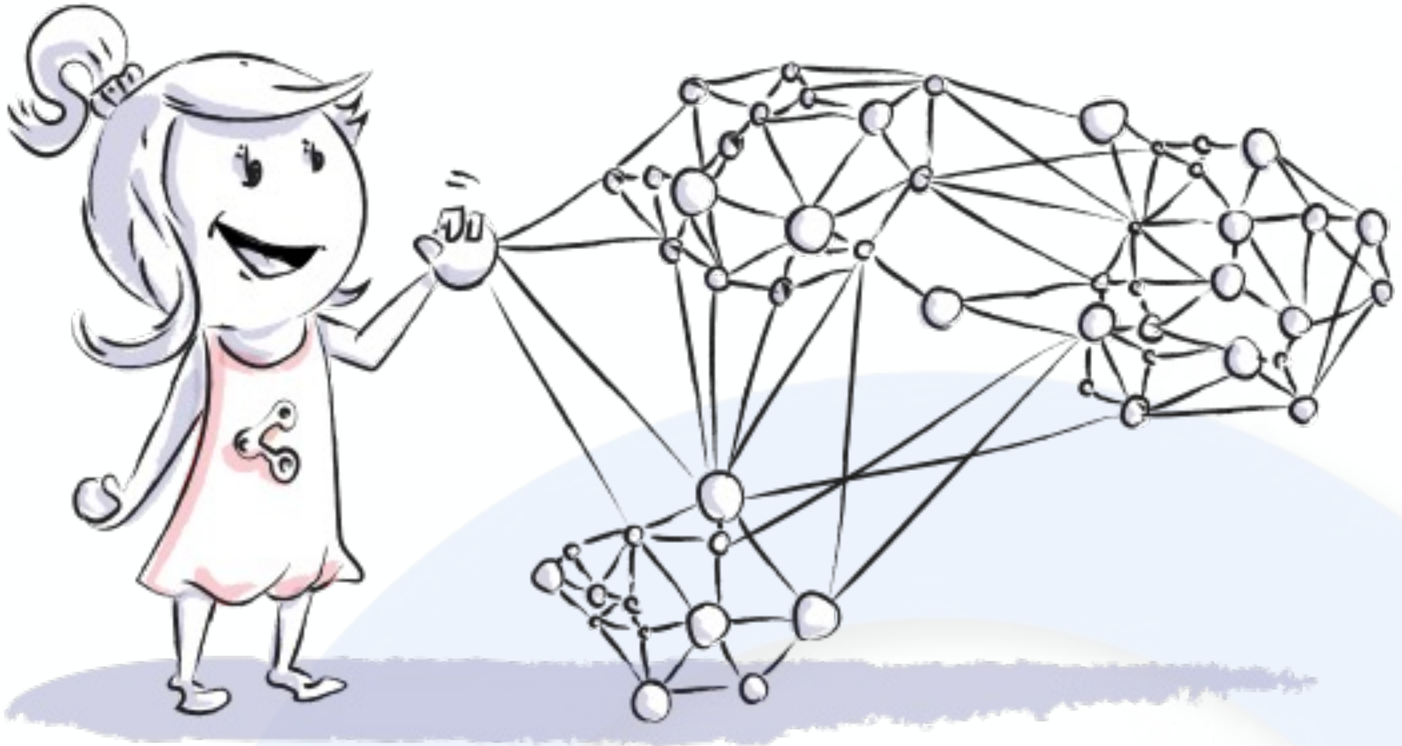
Open &authoritative sources, open APIs, well established metrics

Indicators & visualisations for all needs

OpenAIRE Graph in numbers

- **3Mi** project grants of **25** funders
- **160Mi** publications
- **300k** software
- **57Mi** research data
- **6.3Mi** other research products

Harvested from **+2k** data sources



OpenAIRE Graph

Sources contributing to the Graph

GitHub
Applications Database Supporting
Software Heritage PRESERVING TECHNICAL KNOWLEDGE
elixir bio.tools

… and more **tools and software sources**

frontiers
PLOS
Springer
F1000 Faculty of 1000

… and more **publishers**

CORDIS
NSF
DFG Deutsche Forschungsgemeinschaft

… and more **European and international funders**

OpenAIRE Graph

unpaywall
Microsoft
OpenCitations

… and more **research graphs**

DataCite
re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES
ROR
ORCID
Crossref
OpenDOAR
FAIRsharing.org standards, databases, policies

… and more **registries**

LA Referencia Red de repositorios de acceso abierto a la ciencia
SciELO
DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

… and more **aggregators**

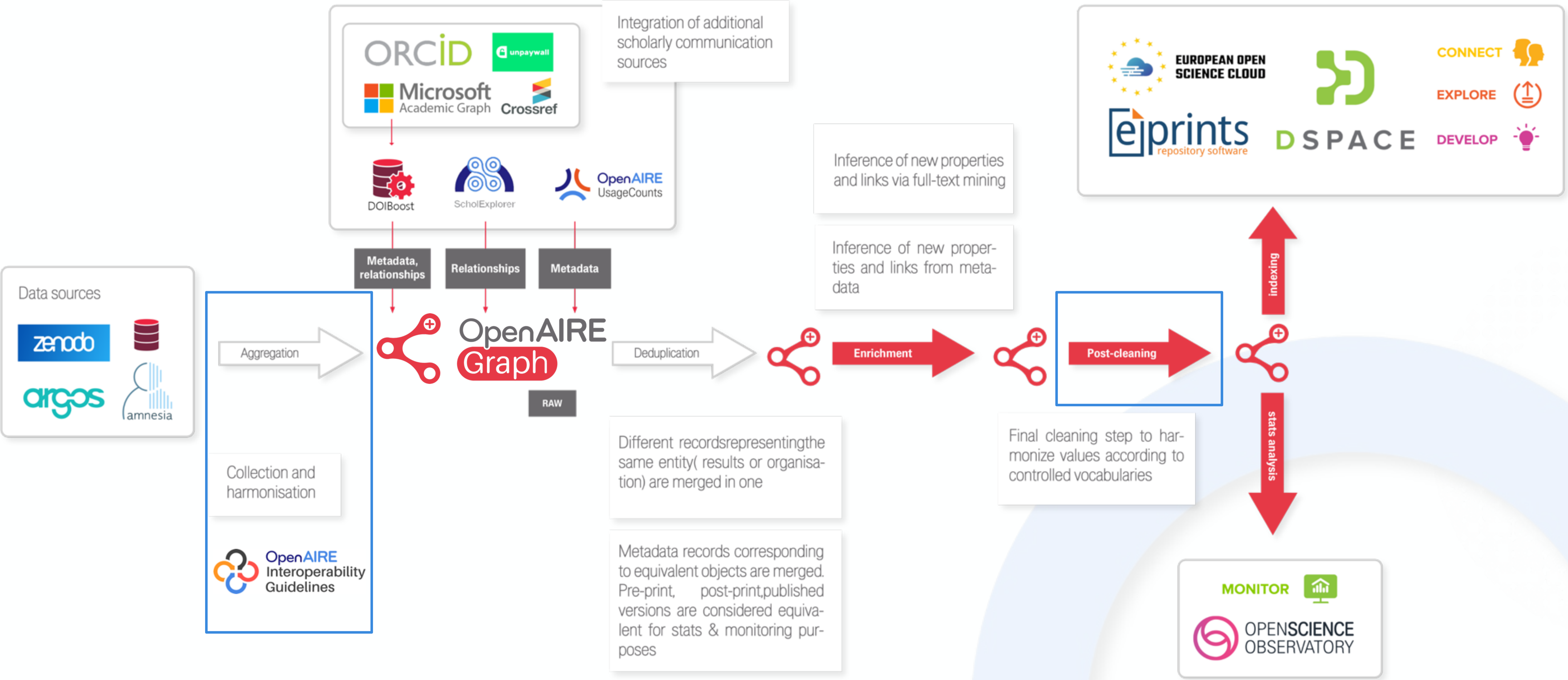
Cornell University Library arXiv.org
RePEc
HAL archives-ouvertes.fr
PubMed
zenodo Research. Shared.

… and more **thematic and institutional repositories**

DARIAH-EU
elixir
EPOS EUROPEAN PLATE OBSERVING SYSTEM
EOSC-Life
ECRIN EUROPEAN CLINICAL RESEARCH INFRASTRUCTURE NETWORK
egi

… and more **einfra and RI sources**

OpenAIRE Graph: the supply chain



Integration of additional scholarly communication sources

Metadata, relationships

Relationships

Metadata

Data sources

Aggregation

Collection and harmonisation

Deduplication

Different recordsrepresentingthe same entity(results or organisation) are merged in one

Metadata records corresponding to equivalent objects are merged. Pre-print, post-print,published versions are considered equivalent for stats & monitoring purposes

Inference of new properties and links via full-text mining

Inference of new properties and links from metadata

Enrichment

Post-cleaning

Final cleaning step to harmonize values according to controlled vocabularies

Controlled vocabularies in the OpenAIRE Interoperability Guidelines

DataCite Vocabularies

Field		
Title type	Author & contributor name type	Funder identifier type
Related identifier type	Alternate identifier type	Relation type

Extensions to Datacite

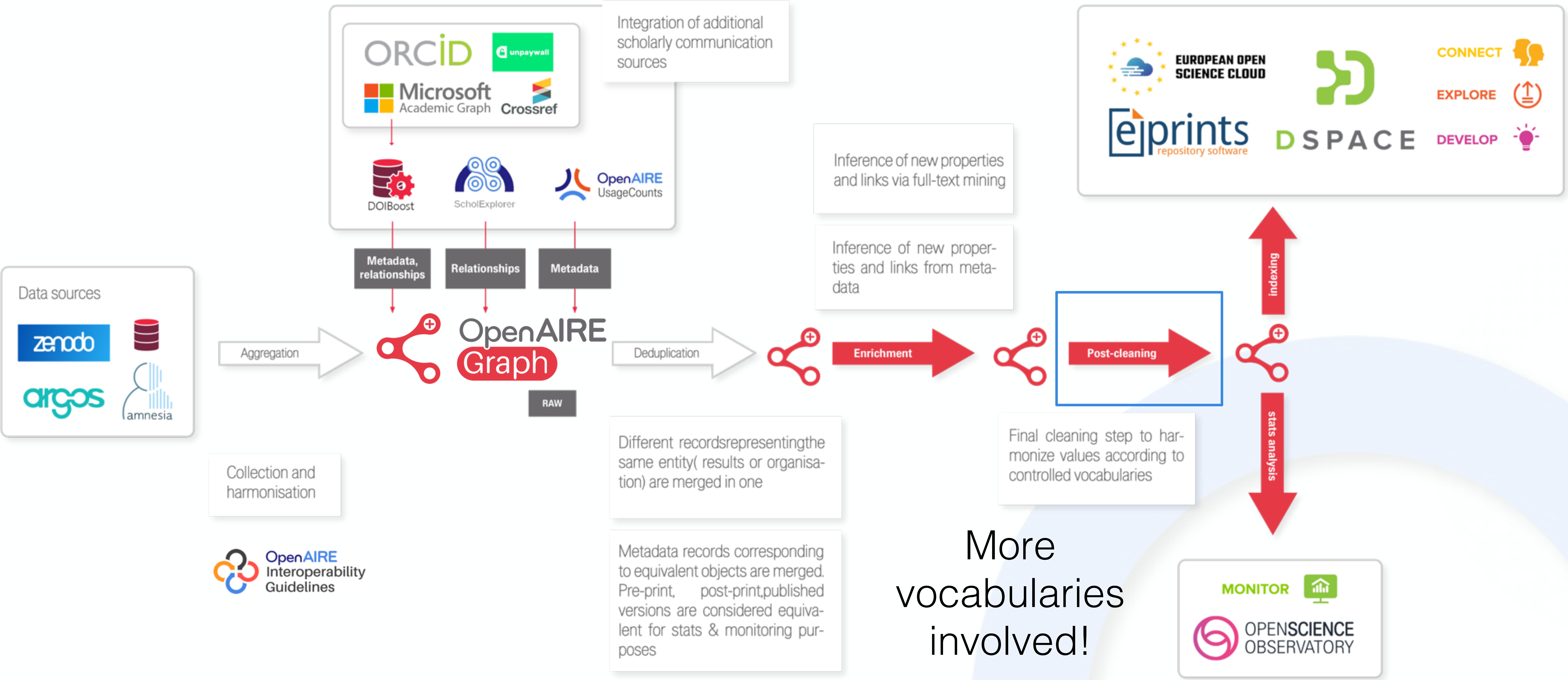
Field	Extension
Identifier type	DOI + other PID types (e.g. Handle)
Contributor type	Datacite + Contributor Roles Taxonomy (CRediT)
Date type	A subset of Datacite vocabulary

COAR Controlled Vocabularies for Repositories (https://vocabularies.coar-repositories.org/)

Field		
Resource type	Access right	Resource version

Based on latest OpenAIRE guidelines for institutional and thematic repositories

OpenAIRE Graph: the supply chain



Integration of additional scholarly communication sources

Collection and harmonisation

Inference of new properties and links via full-text mining

Inference of new properties and links from metadata

Different recordsrepresentingthe same entity(results or organisation) are merged in one

Metadata records corresponding to equivalent objects are merged. Pre-print, post-print,published versions are considered equivalent for stats & monitoring purposes

Final cleaning step to harmonize values according to controlled vocabularies

More vocabularies involved!

Controlled vocabularies for

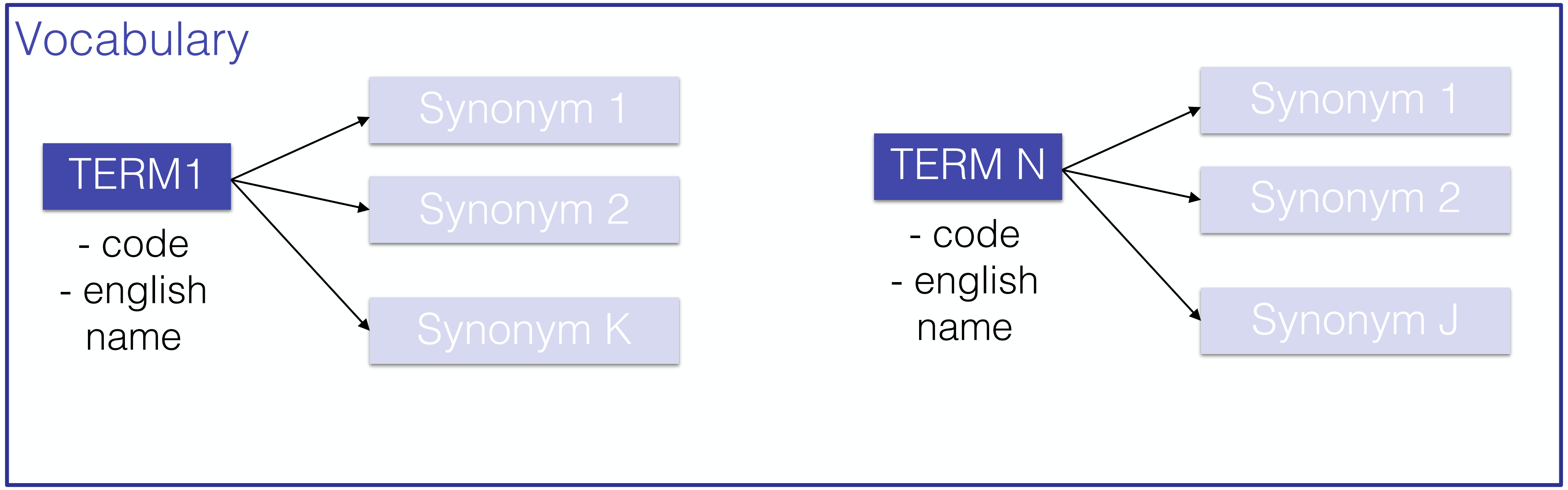
- Types:
  - Of research results (publication, data, software, other)
  - Of persistent identifiers (e.g. DOI, handle, ORCID, ror)
  - Of datasources (e.g. repository, aggregator, journal archive)

  → Used as input to the EOSC Vocabularies for data sources and research products
- Classification schemes
  - The list of subject classification vocabularies OpenAIRE supports
    - e.g. ACM, JEL, LCSH, MeSH, rXiv
- Fields of Science
- Sustainable Development Goals

Controlled vocabularies

Associations
metadata field → controlled vocabulary

1 field 1 vocabulary
1 vocabulary N fields (e.g. one vocabulary for PID types, used for result.pid and organisation.pid)



Vocabulary

TERM1
- code
- english name

Synonym 1
Synonym 2
Synonym K

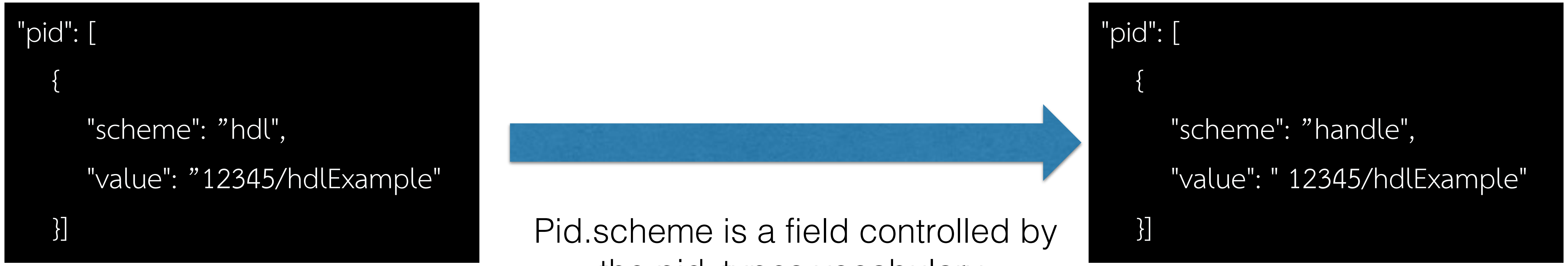
TERM N
- code
- english name

Synonym 1
Synonym 2
Synonym J

- Vocabularies are managed by the OpenAIRE’s aggregation system, with integrated GUI for update
- All vocabularies exported via API https://api.openaire.eu/vocabularies/

Controlled fields in OpenAIRE

- Controlled fields are metadata properties whose values should comply to a vocabulary
- For each vocabulary, OpenAIRE maintains a “vocabulary of synonyms”
    - Each term of the vocabulary is associated with the list of synonyms that are present in the harvested records
    - When a synonym is found, it is “cleaned” i.e. changed with the corresponding term



```
"pid": [
   {
       "scheme": ”hdl",
       "value": ”12345/hdlExample"
   }]
```

```
"pid": [
   {
       "scheme": ”handle",
       "value": " 12345/hdlExample"
   }]
```

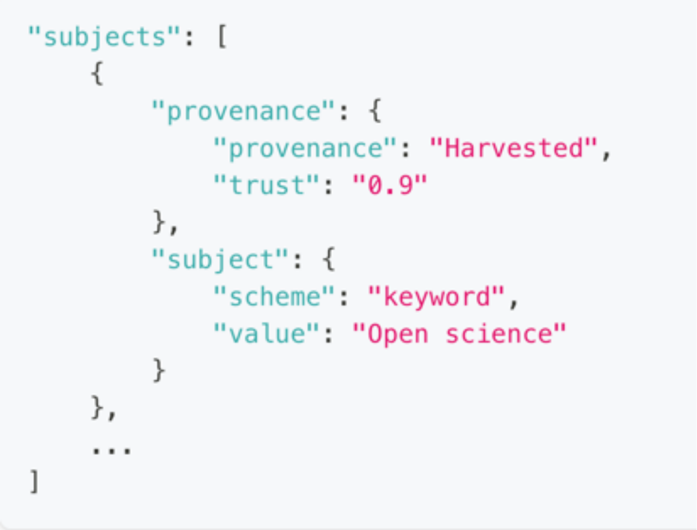
Pid.scheme is a field controlled by the pid_types vocabulary.

Subject classification: Standard schemes, SDGs, FoS

Subjects in OpenAIRE

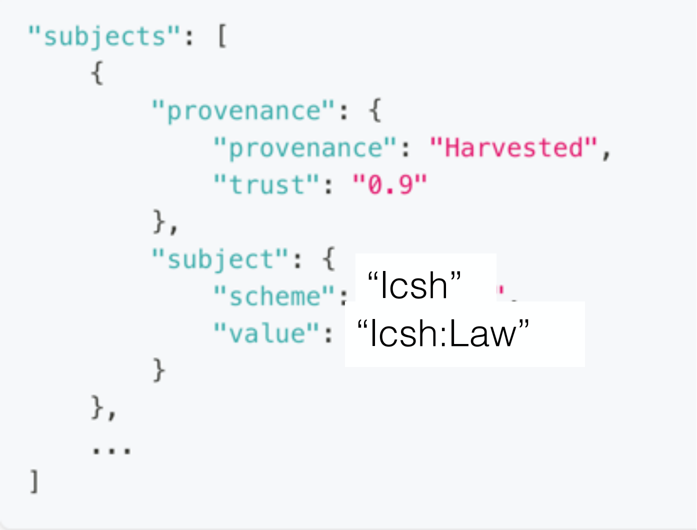
Harvested from sources

<dc:subject>Open Science</dc:subject>



```
"subjects": [
    {
        "provenance": {
            "provenance": "Harvested",
            "trust": "0.9"
        },
        "subject": {
            "scheme": "keyword",
            "value": "Open science"
        }
    },
    ...
]
```

<dc:subject>lcsh:Law</dc:subject>



```
"subjects": [
    {
        "provenance": {
            "provenance": "Harvested",
            "trust": "0.9"
        },
        "subject": { 
            "scheme": “lcsh”
            "value": “lcsh:Law”
        }
    },
    ...
]
```

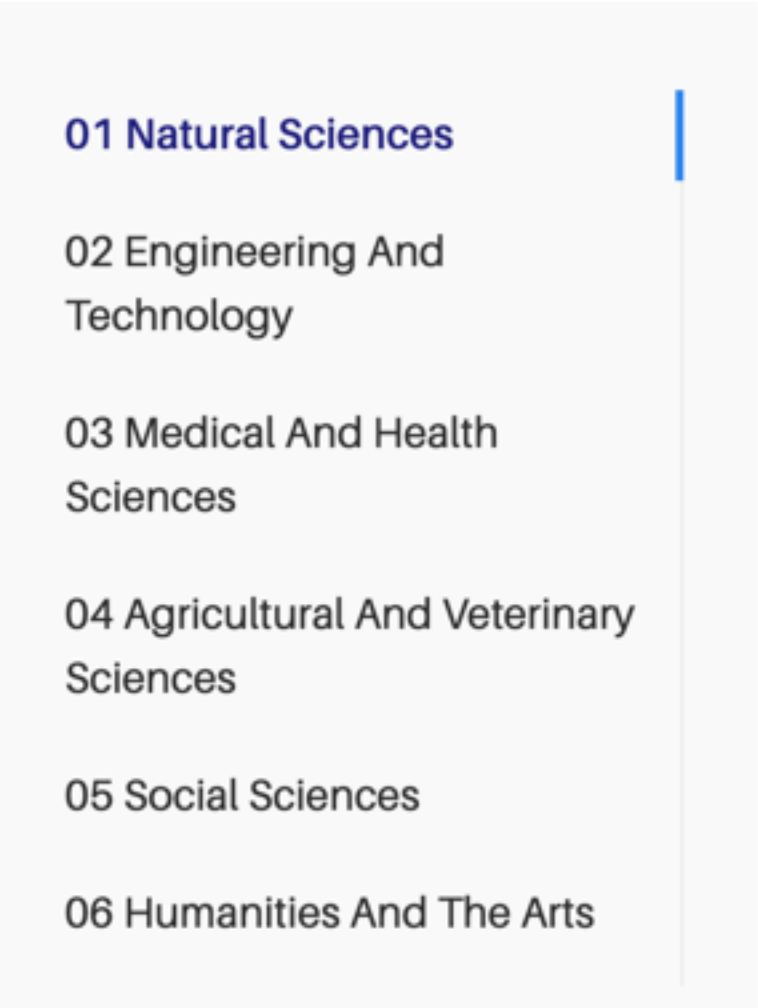
Mostly keywords (40M), but also LCSH (5.5M), DDC (1M) and others. Total results with an harvested subject: 42M

Inferred by OpenAIRE’s mining algorithms

- Terms from standard vocabularies: ACM, rXiv, MeSH. 21M publications
- SDG classification. 700K publications Based on UN SDG classification
- FoS classification. 15M publications. Based on SciNoBo : A Hierarchical Multi-Label Classifier of Scientific Publications by Athena Research Center

About the FoS classification

- Based on the EOCD 2015 Frascati Manual and EuroSciVoc
- The algorithm is based on citations and references and the FoS of their venues (from science-metrix)



01 Natural Sciences

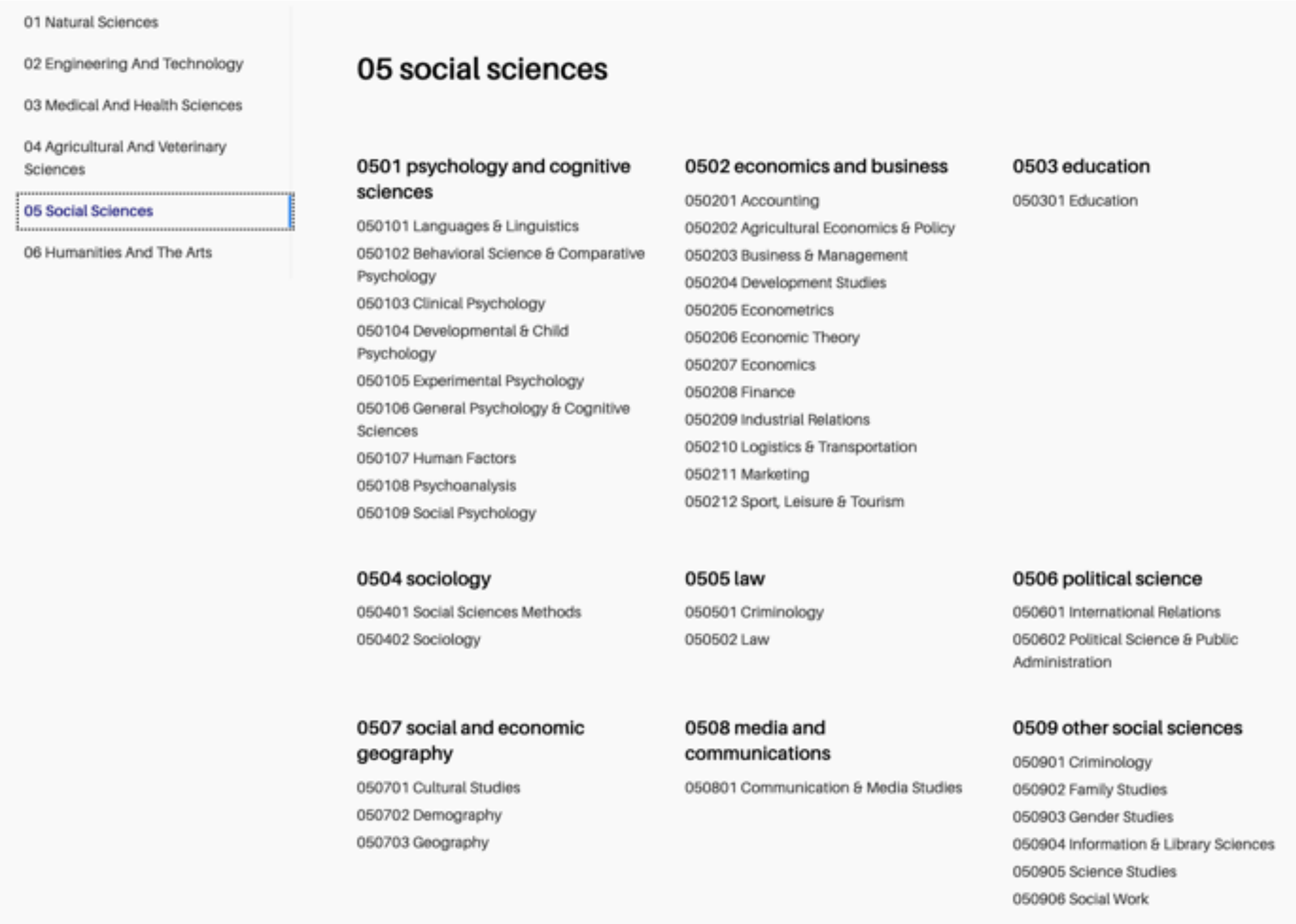
02 Engineering And Technology

03 Medical And Health Sciences

04 Agricultural And Veterinary Sciences

05 Social Sciences

06 Humanities And The Arts



01 Natural Sciences

02 Engineering And Technology

03 Medical And Health Sciences

04 Agricultural And Veterinary Sciences

05 Social Sciences

06 Humanities And The Arts

05 social sciences

0501 psychology and cognitive sciences

050101 Languages & Linguistics
050102 Behavioral Science & Comparative Psychology
050103 Clinical Psychology
050104 Developmental & Child Psychology
050105 Experimental Psychology
050106 General Psychology & Cognitive Sciences
050107 Human Factors
050108 Psychoanalysis
050109 Social Psychology

0502 economics and business

050201 Accounting
050202 Agricultural Economics & Policy
050203 Business & Management
050204 Development Studies
050205 Econometrics
050206 Economic Theory
050207 Economics
050208 Finance
050209 Industrial Relations
050210 Logistics & Transportation
050211 Marketing
050212 Sport, Leisure & Tourism

0503 education

050301 Education

0504 sociology

050401 Social Sciences Methods
050402 Sociology

0505 law

050501 Criminology
050502 Law

0506 political science

050601 International Relations
050602 Political Science & Public Administration

0507 social and economic geography

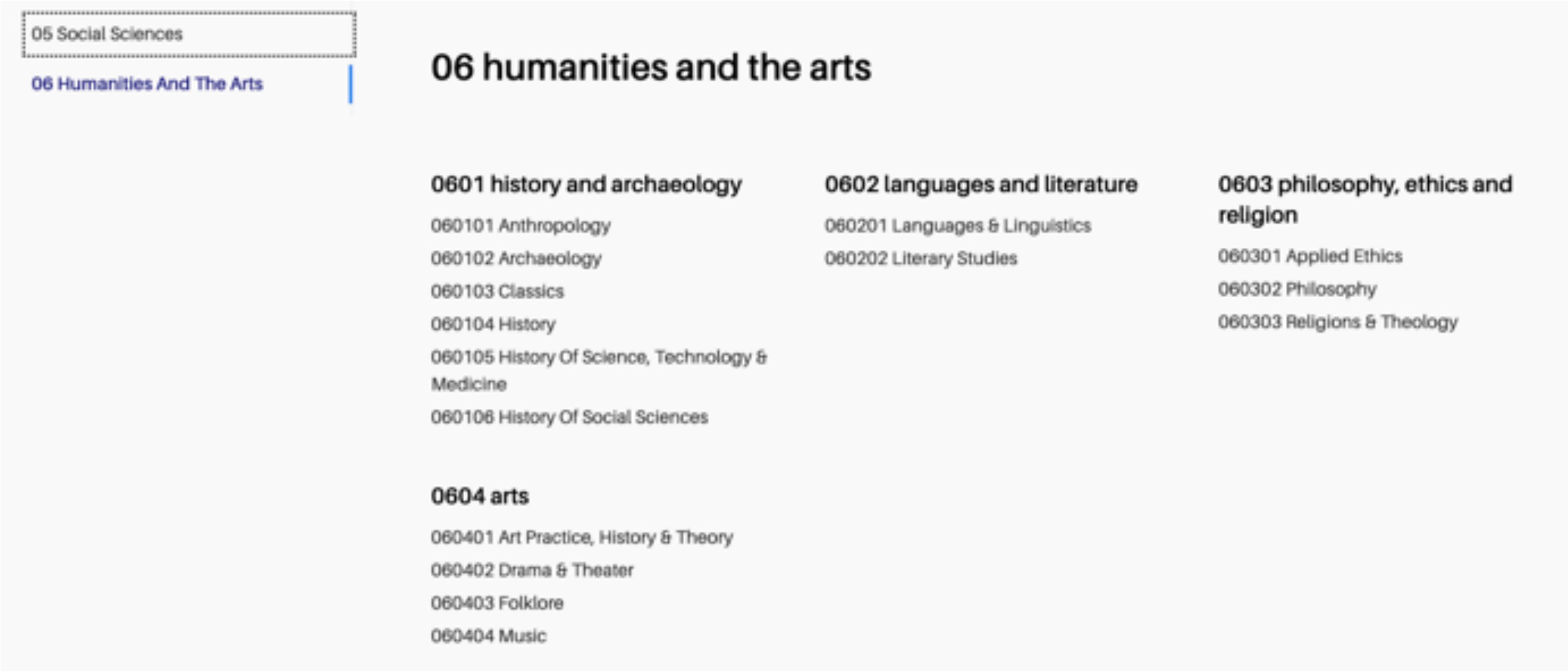
050701 Cultural Studies
050702 Demography
050703 Geography

0508 media and communications

050801 Communication & Media Studies

0509 other social sciences

050901 Criminology
050902 Family Studies
050903 Gender Studies
050904 Information & Library Sciences
050905 Science Studies
050906 Social Work



05 Social Sciences

06 Humanities And The Arts

06 humanities and the arts

0601 history and archaeology

060101 Anthropology
060102 Archaeology
060103 Classics
060104 History
060105 History Of Science, Technology & Medicine
060106 History Of Social Sciences

0602 languages and literature

060201 Languages & Linguistics
060202 Literary Studies

0603 philosophy, ethics and religion

060301 Applied Ethics
060302 Philosophy
060303 Religions & Theology

0604 arts

060401 Art Practice, History & Theory
060402 Drama & Theater
060403 Folklore
060404 Music

Usage of subjects in OpenAIRE

Use of subjects in OpenAIRE Services

OpenAIRE EXPLORE

OpenAIRE CONNECT

OpenAIRE CONNECT

OpenAIRE MONITOR

Portal search

Simple keyword search

Advanced search

Specific search for FoS and SDGs

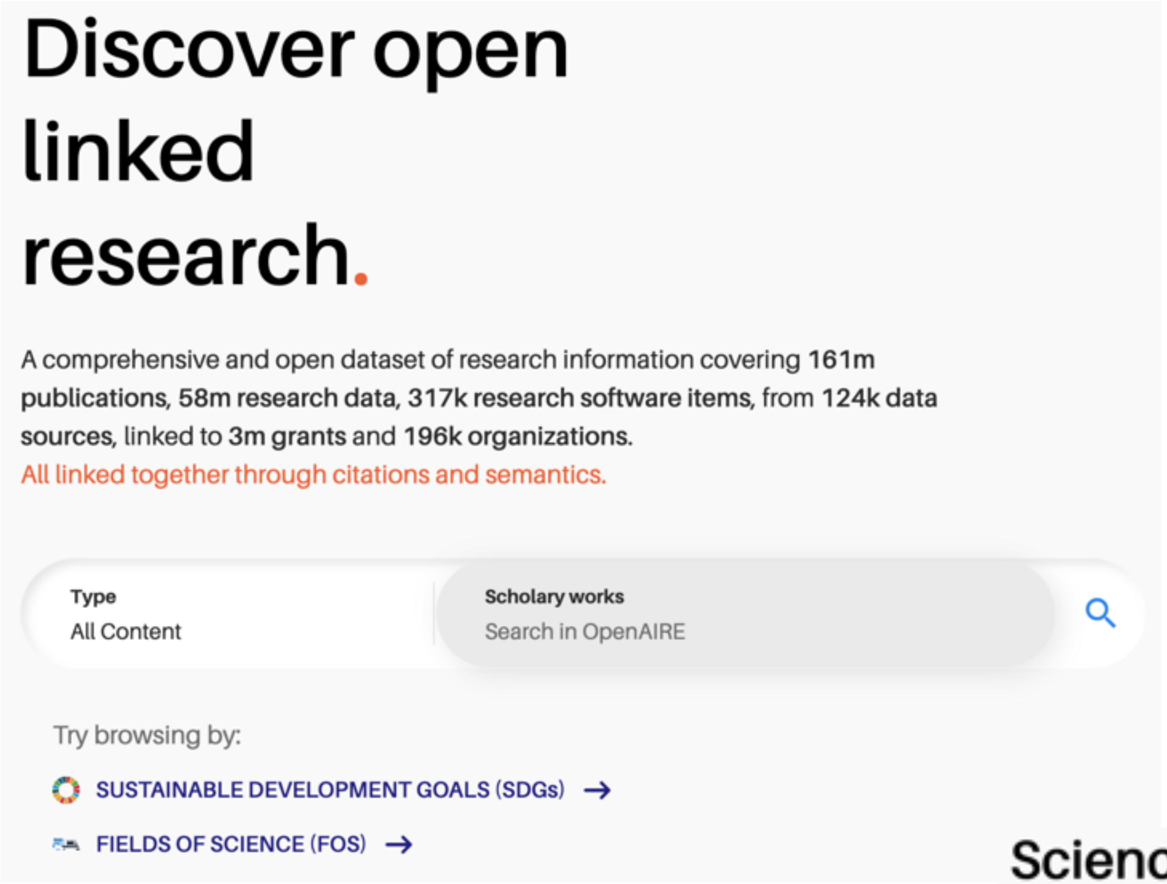
Identify community outputs

Criteria for research communities to identify relevant research products

Statistics

Statistics for funders, institutions, and research initiatives/infrastructures based on SDGs and FoS (WIP)

Portal search

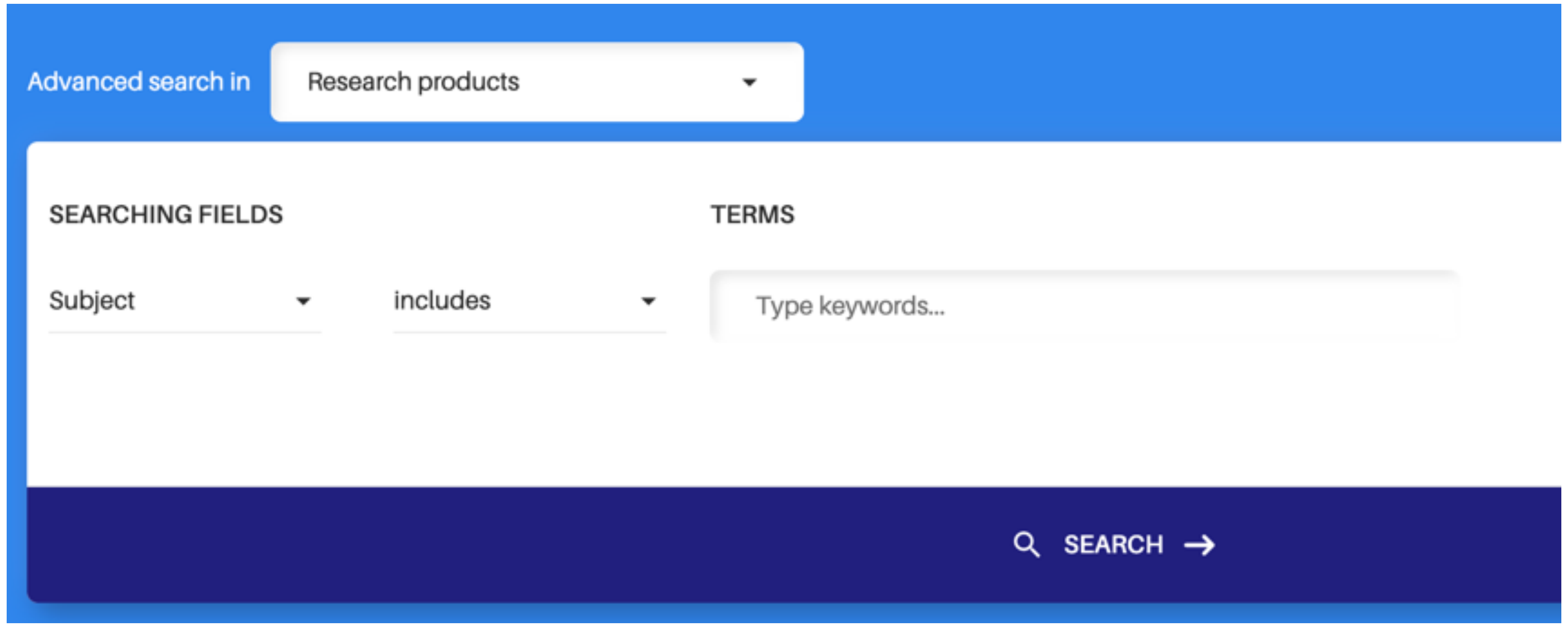


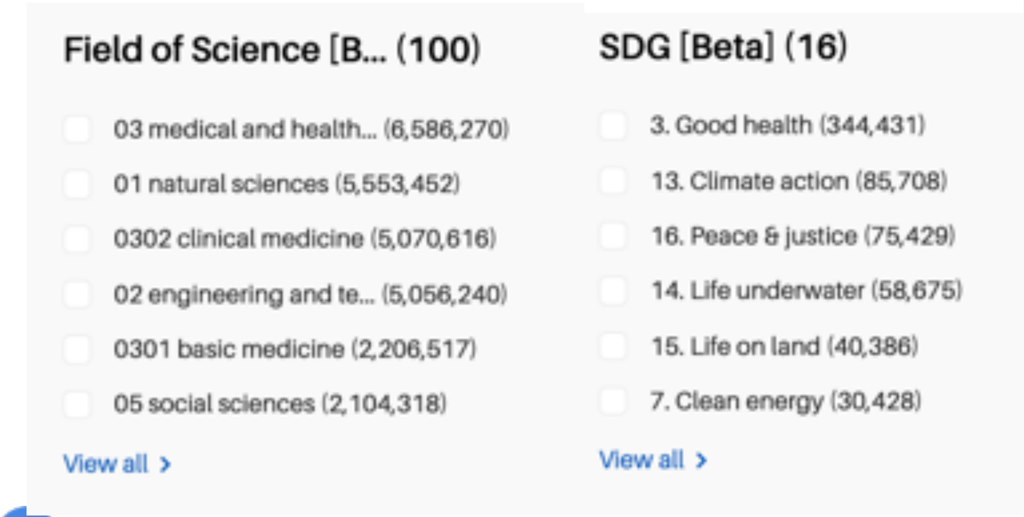
Discover open linked research.

A comprehensive and open dataset of research information covering **161m publications**, **58m research data**, **317k research software items**, from **124k data sources**, linked to **3m grants** and **196k organizations**.
All linked together through citations and semantics.

Try browsing by:

- SUSTAINABLE DEVELOPMENT GOALS (SDGs) →
- FIELDS OF SCIENCE (FOS) →





Field of Science [B... (100)

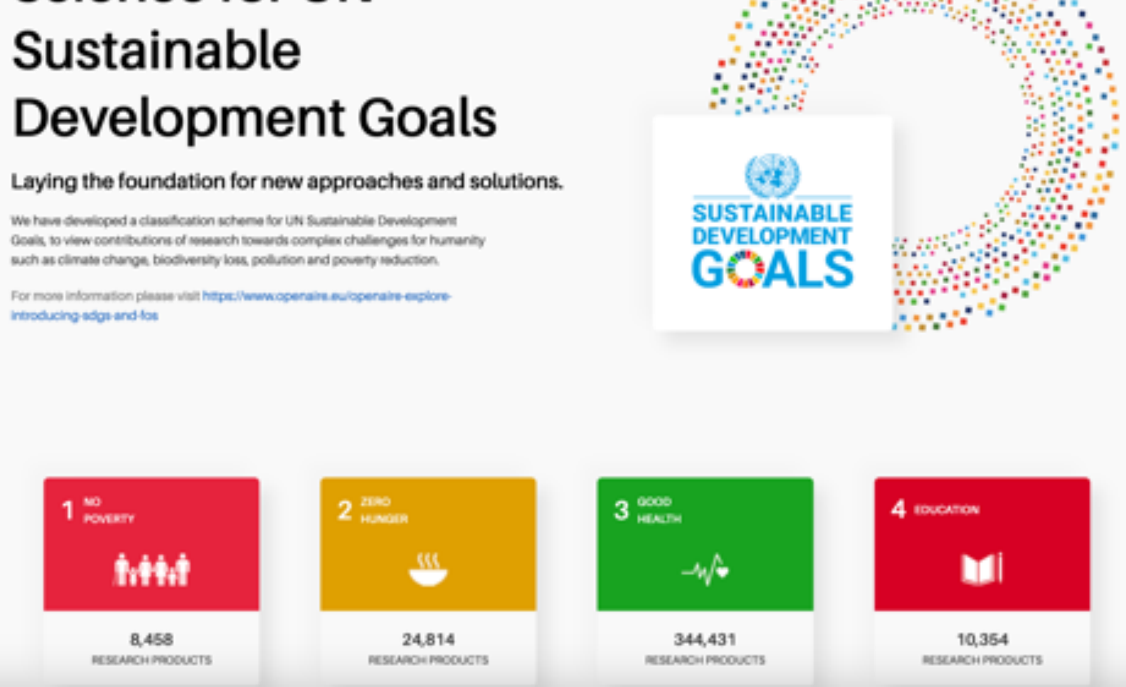
☐ 03 medical and health... (6,586,270)
☐ 01 natural sciences (5,553,452)
☐ 0302 clinical medicine (5,070,616)
☐ 02 engineering and te... (5,056,240)
☐ 0301 basic medicine (2,206,517)
☐ 05 social sciences (2,104,318)

View all ›

SDG [Beta] (16)

☐ 3. Good health (344,431)
☐ 13. Climate action (85,708)
☐ 16. Peace & justice (75,429)
☐ 14. Life underwater (58,675)
☐ 15. Life on land (40,386)
☐ 7. Clean energy (30,428)

View all ›



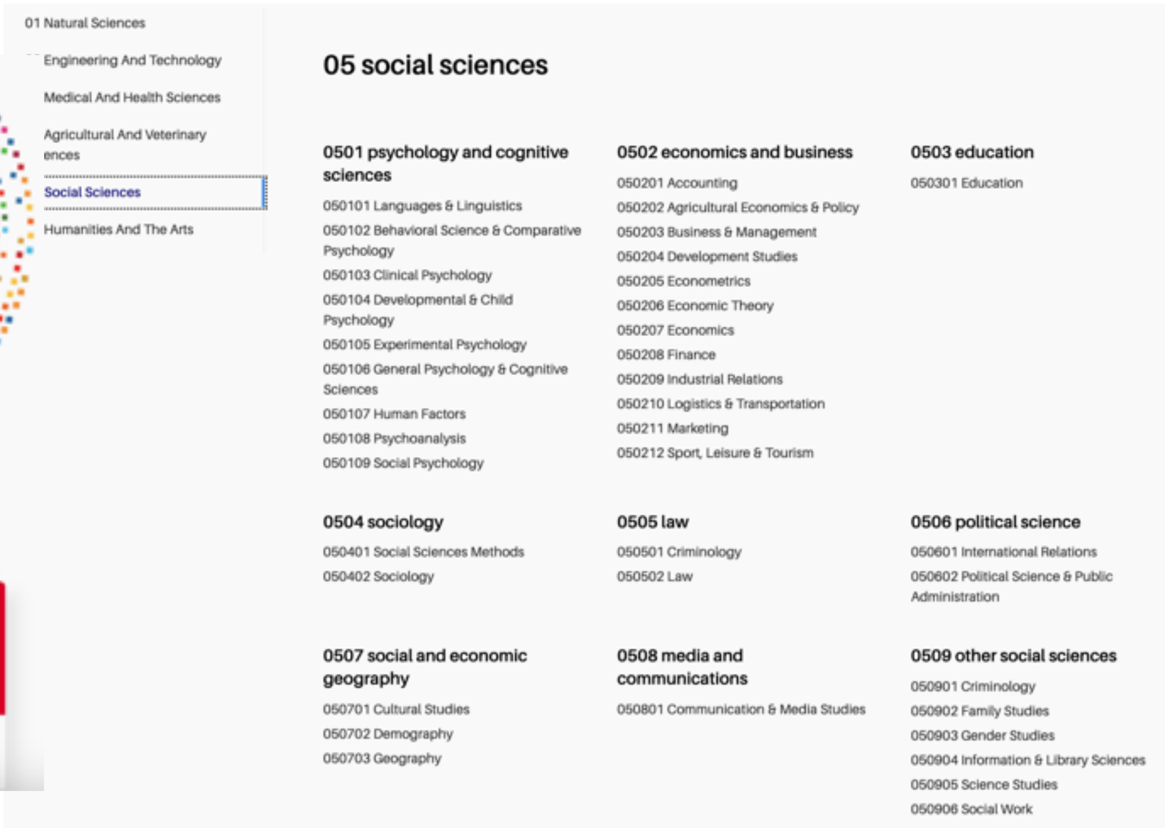
Science for UN Sustainable Development Goals

Laying the foundation for new approaches and solutions.

We have developed a classification scheme for UN Sustainable Development Goals, to view contributions of research towards complex challenges for humanity such as climate change, biodiversity loss, pollution and poverty reduction.

For more information please visit https://www.openaire.eu/openaire-explore-introducing-sdgs-and-fos

1 NO POVERTY	2 ZERO HUNGER	3 GOOD HEALTH	4 EDUCATION
8,458 RESEARCH PRODUCTS	24,814 RESEARCH PRODUCTS	344,431 RESEARCH PRODUCTS	10,354 RESEARCH PRODUCTS



01 Natural Sciences
Engineering And Technology
Medical And Health Sciences
Agricultural And Veterinary ences
Social Sciences
Humanities And The Arts

05 social sciences

0501 psychology and cognitive sciences
050101 Languages & Linguistics
050102 Behavioral Science & Comparative Psychology
050103 Clinical Psychology
050104 Developmental & Child Psychology
050105 Experimental Psychology
050106 General Psychology & Cognitive Sciences
050107 Human Factors
050108 Psychoanalysis
050109 Social Psychology

0502 economics and business
050201 Accounting
050202 Agricultural Economics & Policy
050203 Business & Management
050204 Development Studies
050205 Econometrics
050206 Economic Theory
050207 Economics
050208 Finance
050209 Industrial Relations
050210 Logistics & Transportation
050211 Marketing
050212 Sport, Leisure & Tourism

0503 education
050301 Education

0504 sociology
050401 Social Sciences Methods
050402 Sociology

0505 law
050501 Criminology
050502 Law

0506 political science
050601 International Relations
050602 Political Science & Public Administration

0507 social and economic geography
050701 Cultural Studies
050702 Demography
050703 Geography

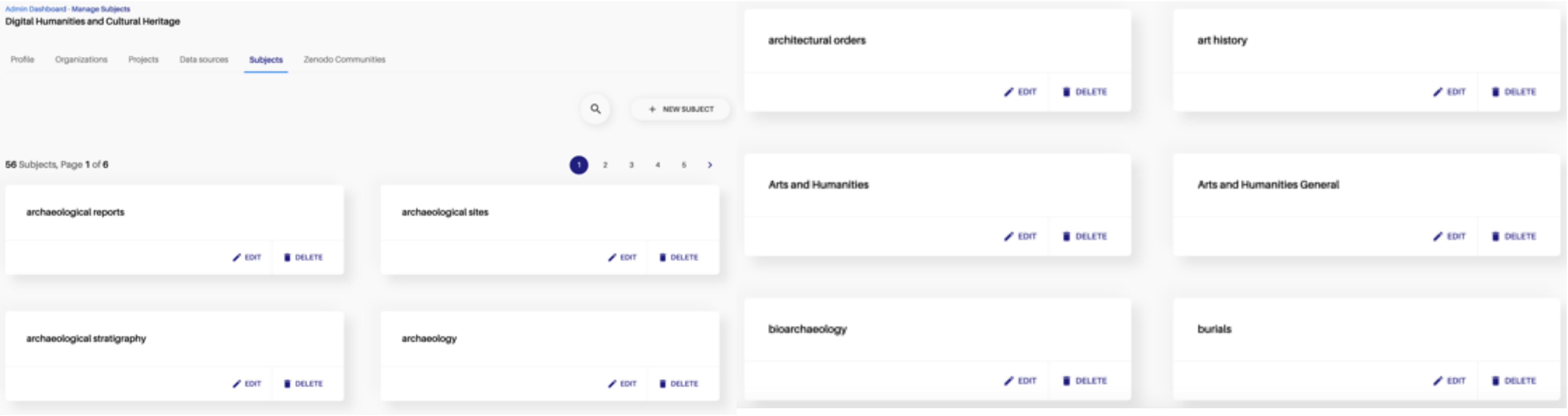
0508 media and communications
050801 Communication & Media Studies

0509 other social sciences
050901 Criminology
050902 Family Studies
050903 Gender Studies
050904 Information & Library Sciences
050905 Science Studies
050906 Social Work

OpenAIRE NEXUS

CONNECT Gateway configuration

With CONNECT, a research community can have a portal where only a subset of the OpenAIRE Graph is available: the subset that is relevant for the research community



Admin Dashboard · Manage Subjects

Digital Humanities and Cultural Heritage

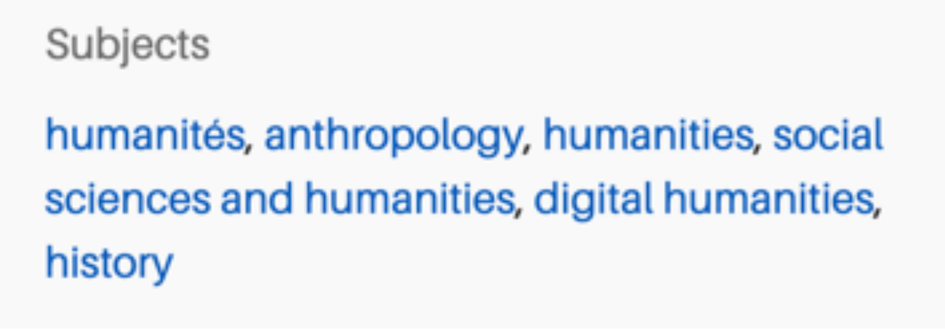
56 Subjects, Page 1 of 6

- archaeological reports
- archaeological sites
- archaeological stratigraphy
- archaeology
- architectural orders
- art history
- Arts and Humanities
- Arts and Humanities General
- bioarchaeology
- burials

OpenAIRE CONNECT & GOTRIPLE

Configuration 1

Using a lot of keywords and terms from other vocabularies

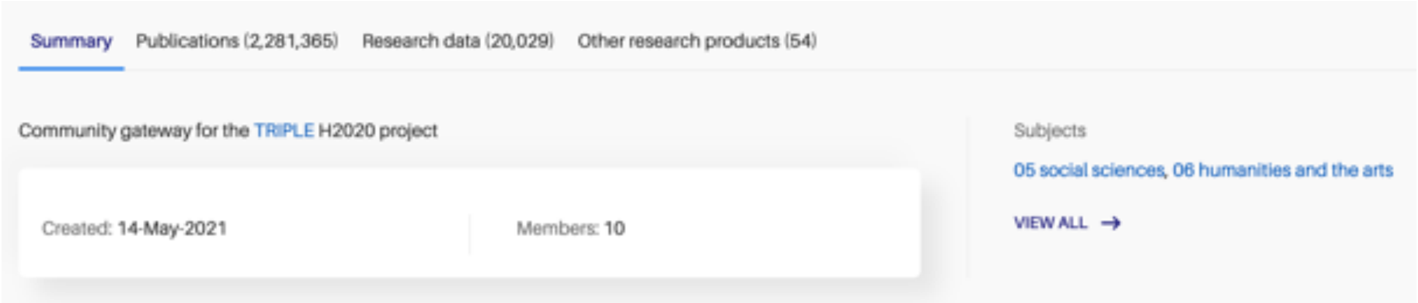


Subjects

humanités, anthropology, humanities, social sciences and humanities, digital humanities, history

Configuration 2

Using FoS: everything from 005 Social Sciences and 006 Humanities and the Arts



Community gateway for the TRIPLE H2020 project

Created: 14-May-2021

Members: 10

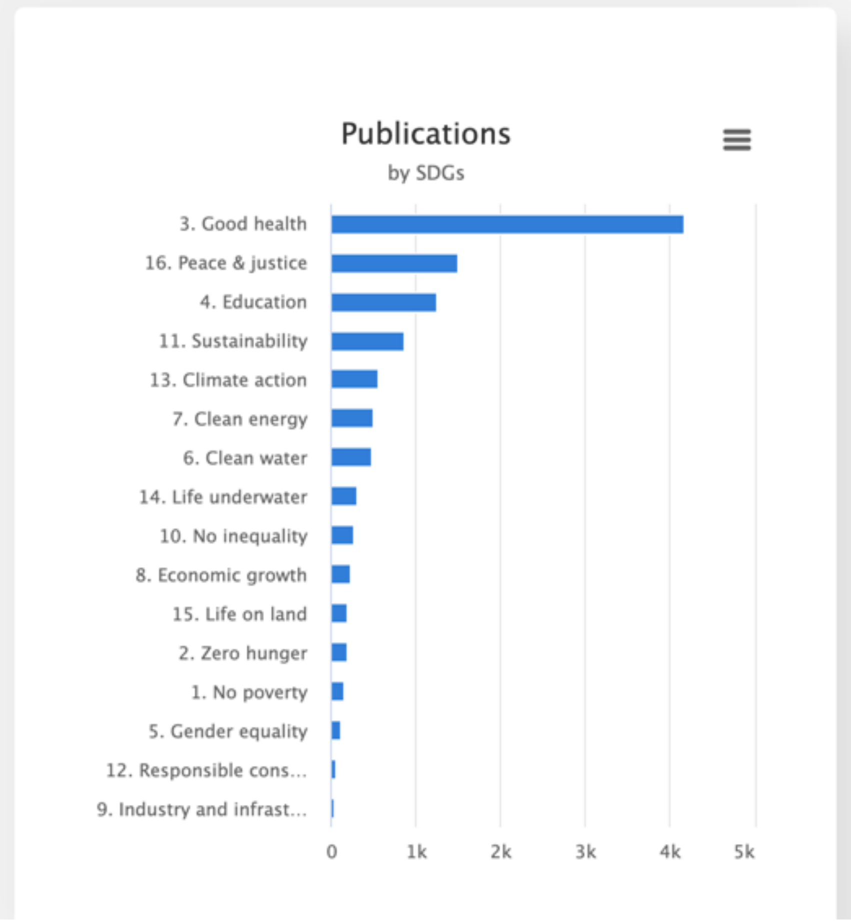
Subjects

05 social sciences, 06 humanities and the arts

The automatic daily backup of the configuration makes it easy to rollback to a previous configuration if the new one does not improve the coverage and precision as planned.

Statistics

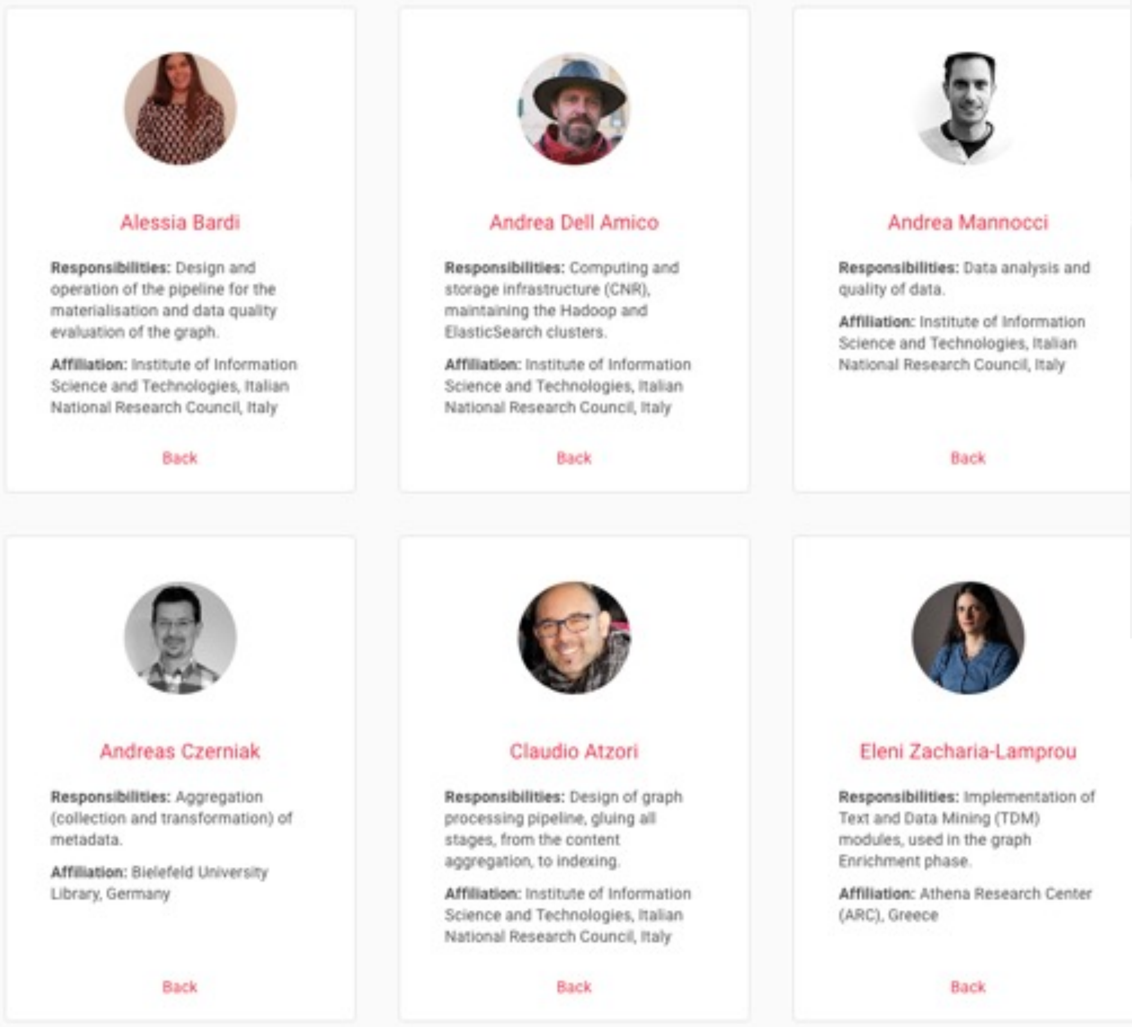
Statistics based on subjects can help institutions at identifying disciplines to be promoted, monitor the growth of cross-disciplinary research, or the fields in which they have more impact.



Conclusion

- The adoption of controlled vocabularies in the OpenAIRE Graph is pervasive
- Harmonisation is not a ”do it once” process:
  - Harmositation is done every time OpenAIRE aggregates a source and every time a new version of the graph is created
  - Vocabularies are living resources: continously updated by the aggregation team to cope with new values
- Topics (subjects) very important to support researchers and communities in the data and literature deluge and to promote alternative ways of discovering research outputs

The team



Alessia Bardi

Responsibilities: Design and operation of the pipeline for the materialisation and data quality evaluation of the graph.

Affiliation: Institute of Information Science and Technologies, Italian National Research Council, Italy

Back

Andrea Dell Amico

Responsibilities: Computing and storage infrastructure (CNR), maintaining the Hadoop and ElasticSearch clusters.

Affiliation: Institute of Information Science and Technologies, Italian National Research Council, Italy

Back

Andrea Mannocci

Responsibilities: Data analysis and quality of data.

Affiliation: Institute of Information Science and Technologies, Italian National Research Council, Italy

Back

Andreas Czerniak

Responsibilities: Aggregation (collection and transformation) of metadata.

Affiliation: Bielefeld University Library, Germany

Back

Claudio Atzori

Responsibilities: Design of graph processing pipeline, gluing all stages, from the content aggregation, to indexing.

Affiliation: Institute of Information Science and Technologies, Italian National Research Council, Italy

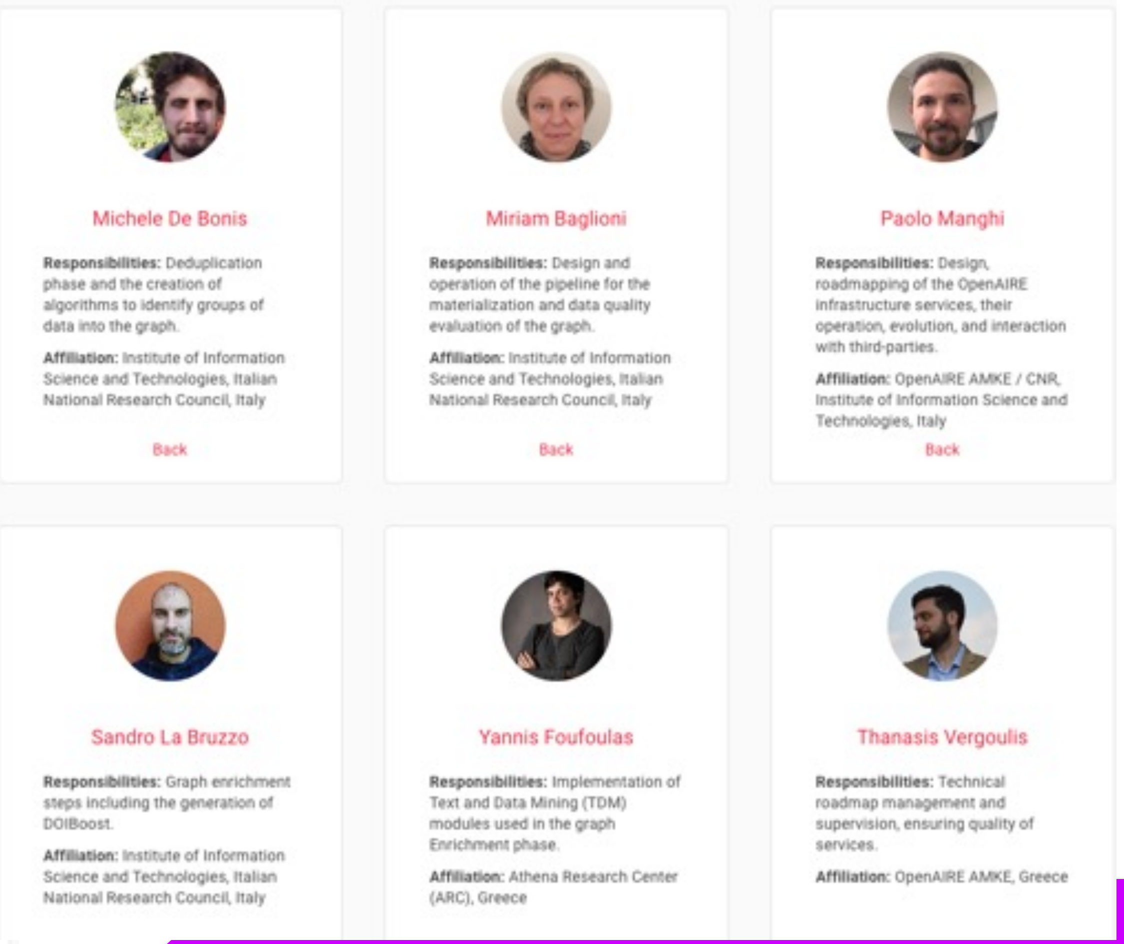
Back

Eleni Zacharia-Lamprou

Responsibilities: Implementation of Text and Data Mining (TDM) modules, used in the graph Enrichment phase.

Affiliation: Athena Research Center (ARC), Greece

Back



Michele De Bonis

Responsibilities: Deduplication phase and the creation of algorithms to identify groups of data into the graph.

Affiliation: Institute of Information Science and Technologies, Italian National Research Council, Italy

Back

Miriam Baglioni

Responsibilities: Design and operation of the pipeline for the materialization and data quality evaluation of the graph.

Affiliation: Institute of Information Science and Technologies, Italian National Research Council, Italy

Back

Paolo Manghi

Responsibilities: Design, roadmapping of the OpenAIRE infrastructure services, their operation, evolution, and interaction with third-parties.

Affiliation: OpenAIRE AMKE / CNR, Institute of Information Science and Technologies, Italy

Back

Sandro La Bruzzo

Responsibilities: Graph enrichment steps including the generation of DOIBoost.

Affiliation: Institute of Information Science and Technologies, Italian National Research Council, Italy

Yannis Foufoulas

Responsibilities: Implementation of Text and Data Mining (TDM) modules used in the graph Enrichment phase.

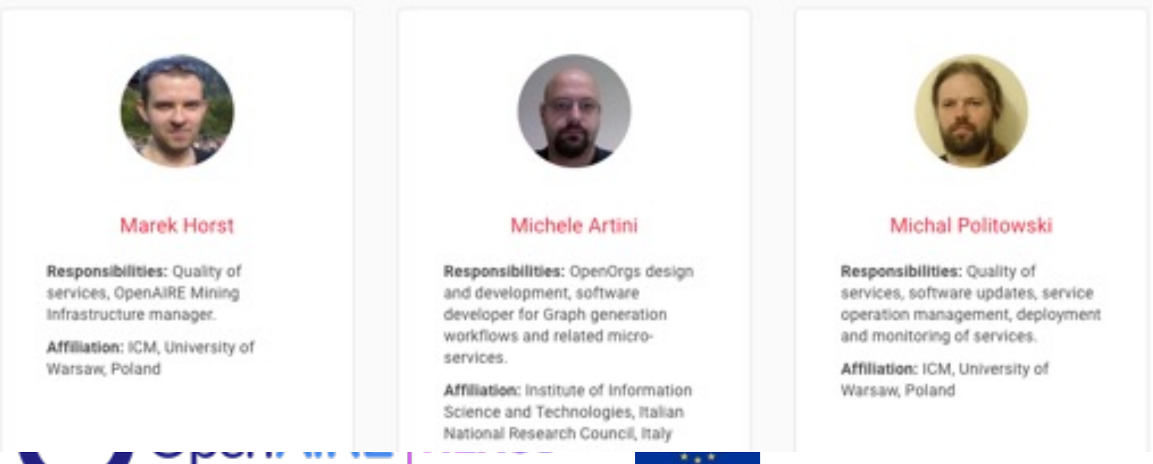
Affiliation: Athena Research Center (ARC), Greece

Thanasis Vergoulis

Responsibilities: Technical roadmap management and supervision, ensuring quality of services.

Affiliation: OpenAIRE AMKE, Greece

Because I am doing this presentation, but it would not have been possible without them



Marek Horst

Responsibilities: Quality of services, OpenAIRE Mining Infrastructure manager.

Affiliation: ICM, University of Warsaw, Poland

Michele Artini

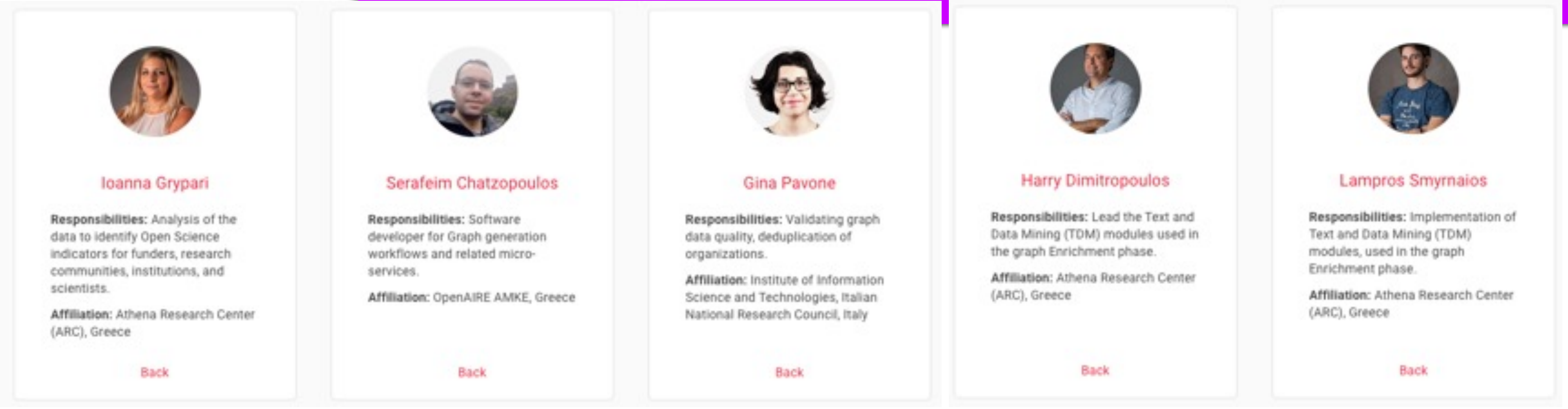
Responsibilities: OpenOrgs design and development, software developer for Graph generation workflows and related micro-services.

Affiliation: Institute of Information Science and Technologies, Italian National Research Council, Italy

Michal Politowski

Responsibilities: Quality of services, software updates, service operation management, deployment and monitoring of services.

Affiliation: ICM, University of Warsaw, Poland



Ioanna Grypari

Responsibilities: Analysis of the data to identify Open Science indicators for funders, research communities, institutions, and scientists.

Affiliation: Athena Research Center (ARC), Greece

Back

Serafeim Chatzopoulos

Responsibilities: Software developer for Graph generation workflows and related micro-services.

Affiliation: OpenAIRE AMKE, Greece

Back

Gina Pavone

Responsibilities: Validating graph data quality, deduplication of organizations.

Affiliation: Institute of Information Science and Technologies, Italian National Research Council, Italy

Back

Harry Dimitropoulos

Responsibilities: Lead the Text and Data Mining (TDM) modules used in the graph Enrichment phase.

Affiliation: Athena Research Center (ARC), Greece

Back

Lampros Smyrnaios

Responsibilities: Implementation of Text and Data Mining (TDM) modules, used in the graph Enrichment phase.

Affiliation: Athena Research Center (ARC), Greece

Back

@openaire_eu

#OpenAIRE-Nexus

THANK YOU

Alessia Bardi

Alessia.bardi@isti.cnr.it