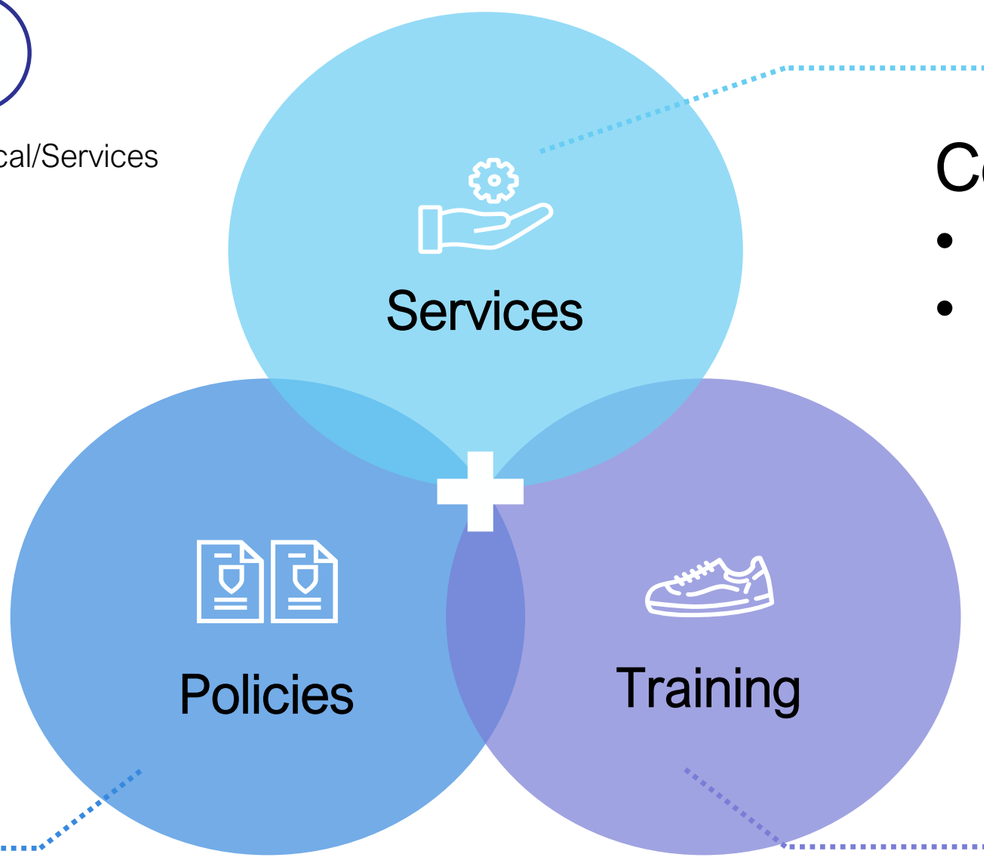
Integration of organizations in the OpenAIRE Research Graph

Alessia Bardi

National Research Council, Italy

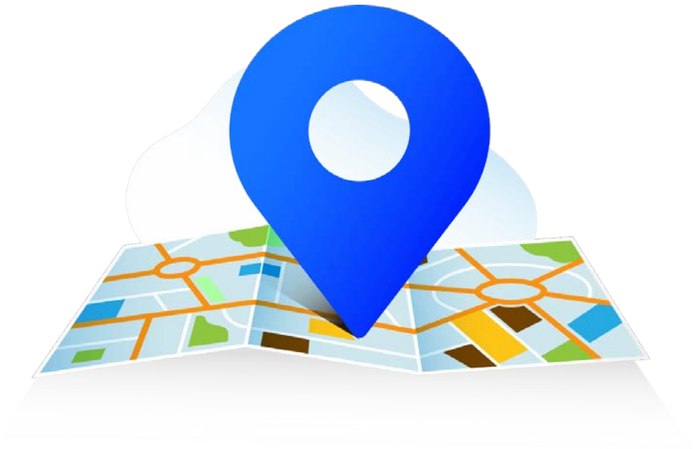
OpenAIRE: three pillars of action
www.openaire.eu



Connecting
- *Establishing infrastructure*
- *Added value services*

Aligning
- *Standard*
- *Guidelines*
- *Practices*
- *Workflows*

Empowering:
- *Open Science*
- *Open Access*
- *Policies*
- *Services*



Tracking Open Science

Reproduciblity and transparency require tracking of all outcomes of science and related ”context”

Monitoring Open Science

Monitoring quality, impact, and ”open scienceness” of science should be a transparent, reproducible process for all, inclusive of research “context”

Discovery Open Science

Discovery of reproducibe science outcomes must find new ways, driven by “scientific intentions” that go beyond the “find articles related to a research topic”.

OpenAIRE Open Science Monitoring

Research admins

Research impact

MONITOR CONNECT

Funders

Research communities

Open Access/Science trends

MONITOR CONNECT

Research Organization

Research Infrastructures

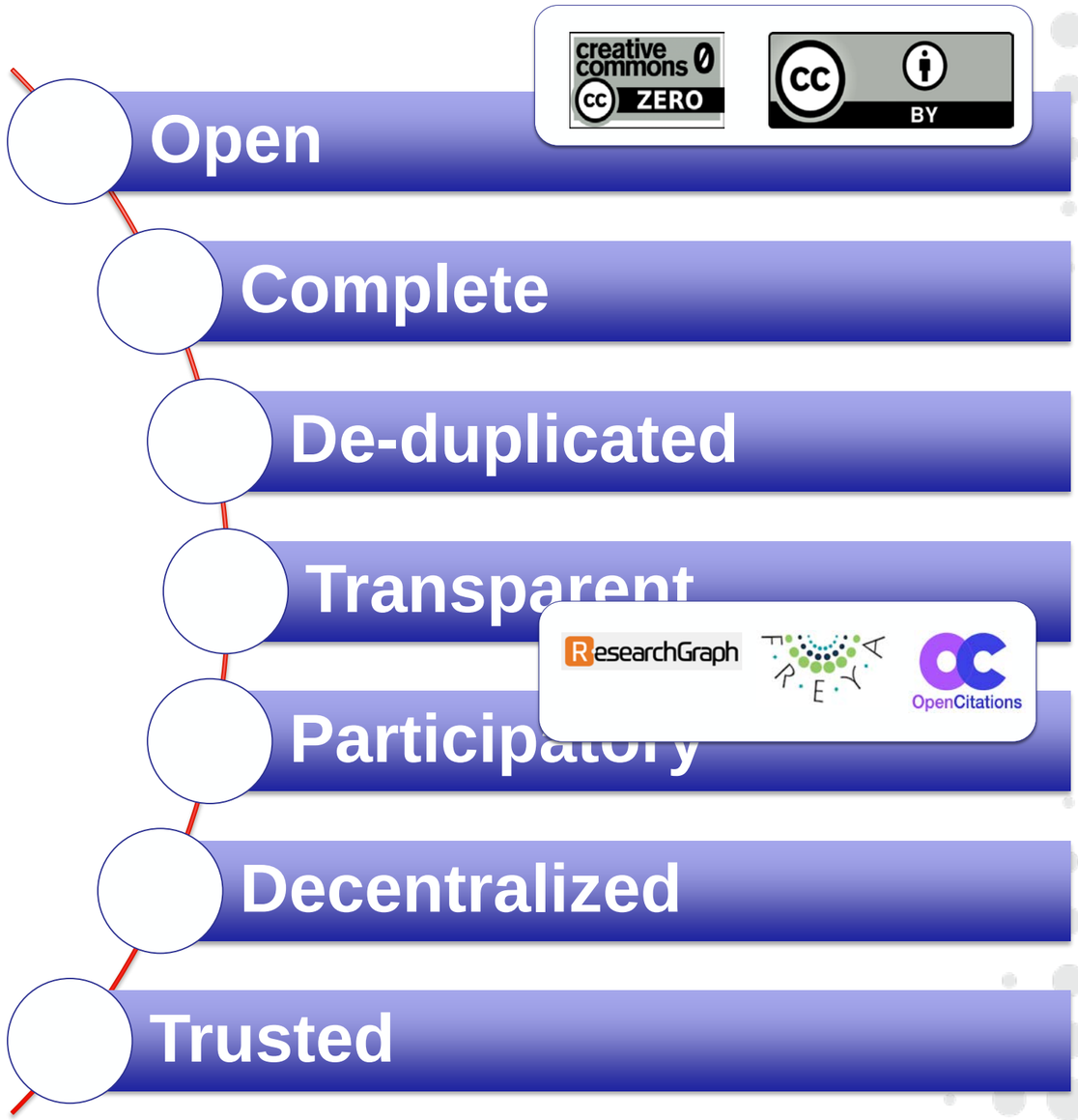
www.openaire.eu

https://monitor.openaire.eu

https://connect.openaire.eu

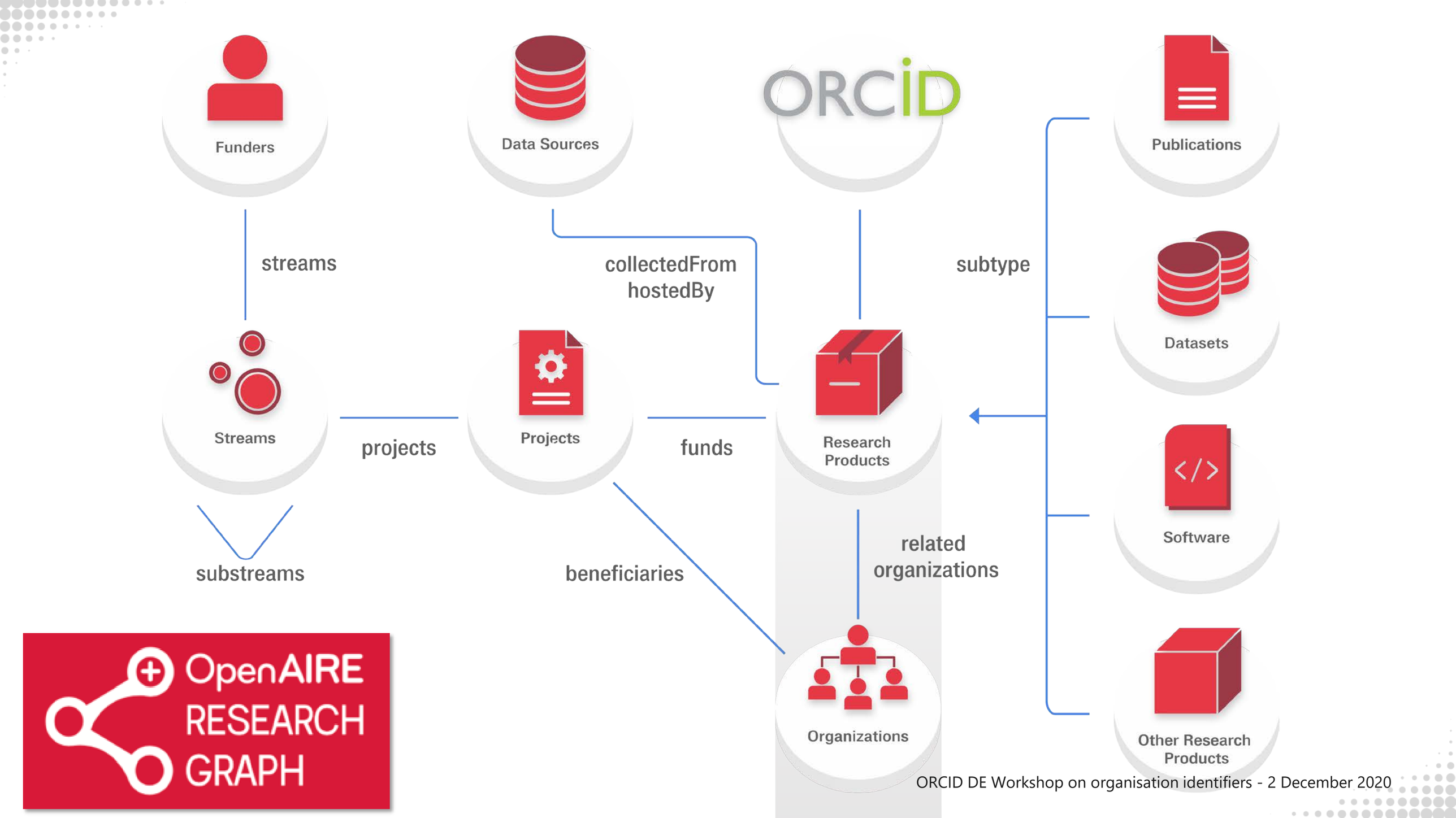
OpenAIRE RESEARCH GRAPH

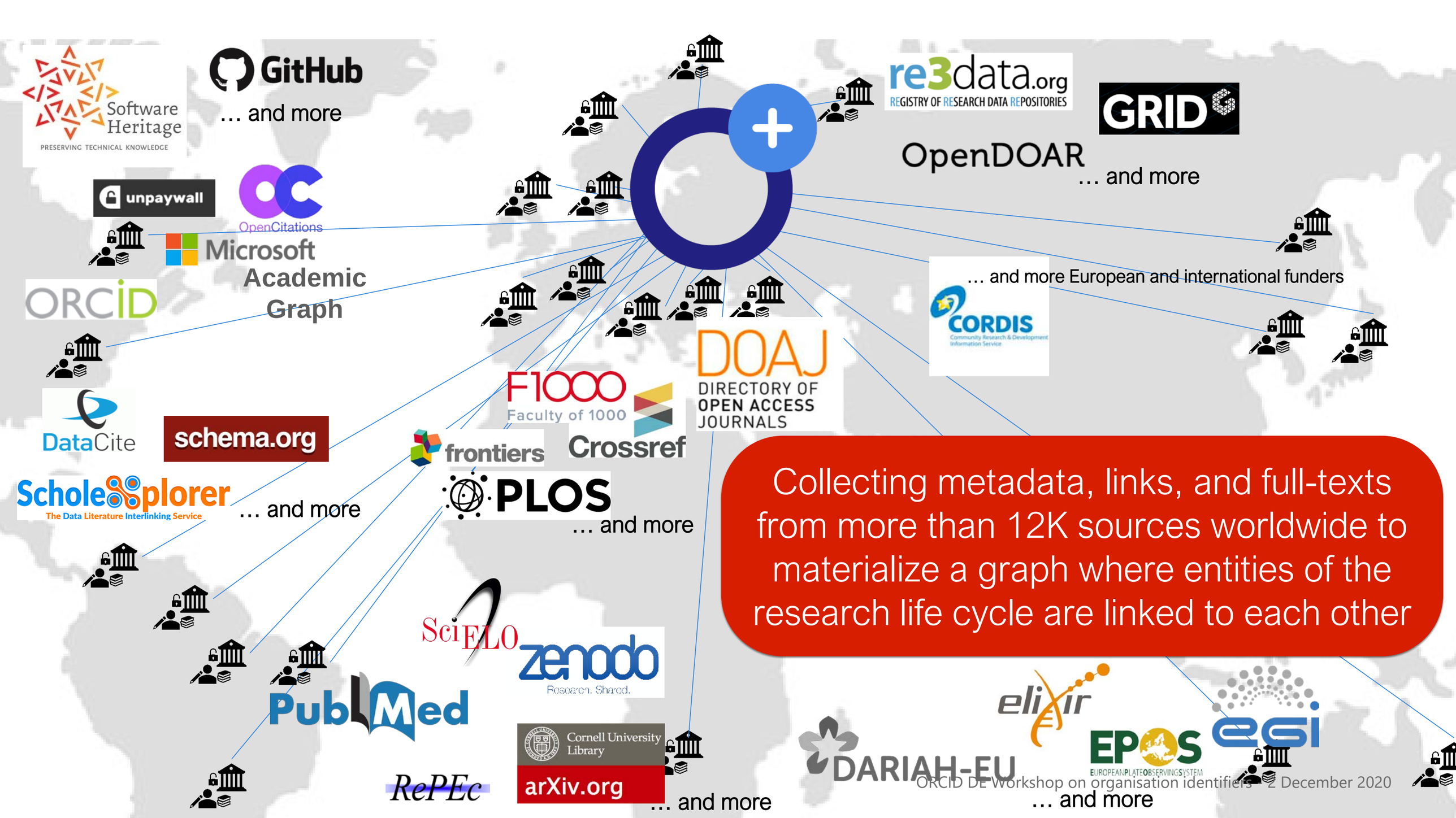
Providing an open metadata research graph of interlinked scientific products, with Open Access information, linked to funding information and research communities



OpenAIRE

European Commission


Funders
Data Sources
ORCID
Publications
streams
collectedFrom
hostedBy
subtype
Datasets
Streams
projects
Projects
funds
Research
Products
Software
substreams
beneficiaries
related
organizations
Organizations
Other Research
Products
OpenAIRE
RESEARCH
GRAPH


Software Heritage
PRESERVING TECHNICAL KNOWLEDGE
GitHub
… and more
re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES
GRID
OpenDOAR
… and more
unpaywall
OpenCitations
Microsoft Academic Graph
ORCID
… and more European and international funders
CORDIS
Community Research & Development Information Service
DataCite
schema.org
F1000
Faculty of 1000
DOAJ
DIRECTORY OF OPEN ACCESS JOURNALS
frontiers
Crossref
ScholeXplorer
The Data Literature Interlinking Service
… and more
PLOS
… and more
Collecting metadata, links, and full-texts from more than 12K sources worldwide to materialize a graph where entities of the research life cycle are linked to each other
SciELO
zenodo
Research. Shared.
PubMed
Cornell University Library
arXiv.org
RePEc
… and more
elixir
DARIAH-EU
EPOS
EUROPEAN PLATE OBSERVING SYSTEM
egi
… and more

PIDs in OpenAIRE

Authors

ORCID

Microsoft Academic

Organizations

GRID

ROR

CORDIS Community Research & Development Information Service

and other funders

Products

DataCite FIND, ACCESS, AND REUSE DATA

EMBL-EBI

arXiv.org

Identifiers.org

Crossref

Funders/Projects

FCT Fundação para a Ciência e a Tecnologia

CORDIS Community Research & Development Information Service

NWO

Science Foundation Ireland sfi For what's next

ACADEMY OF FINLAND

and other funders

Data sources

re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES

OpenDOAR

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

\+ journal lists from publishers

FAIRSFAIR Fostering Fair Data Practices in Europe

Services (fortcoming…)

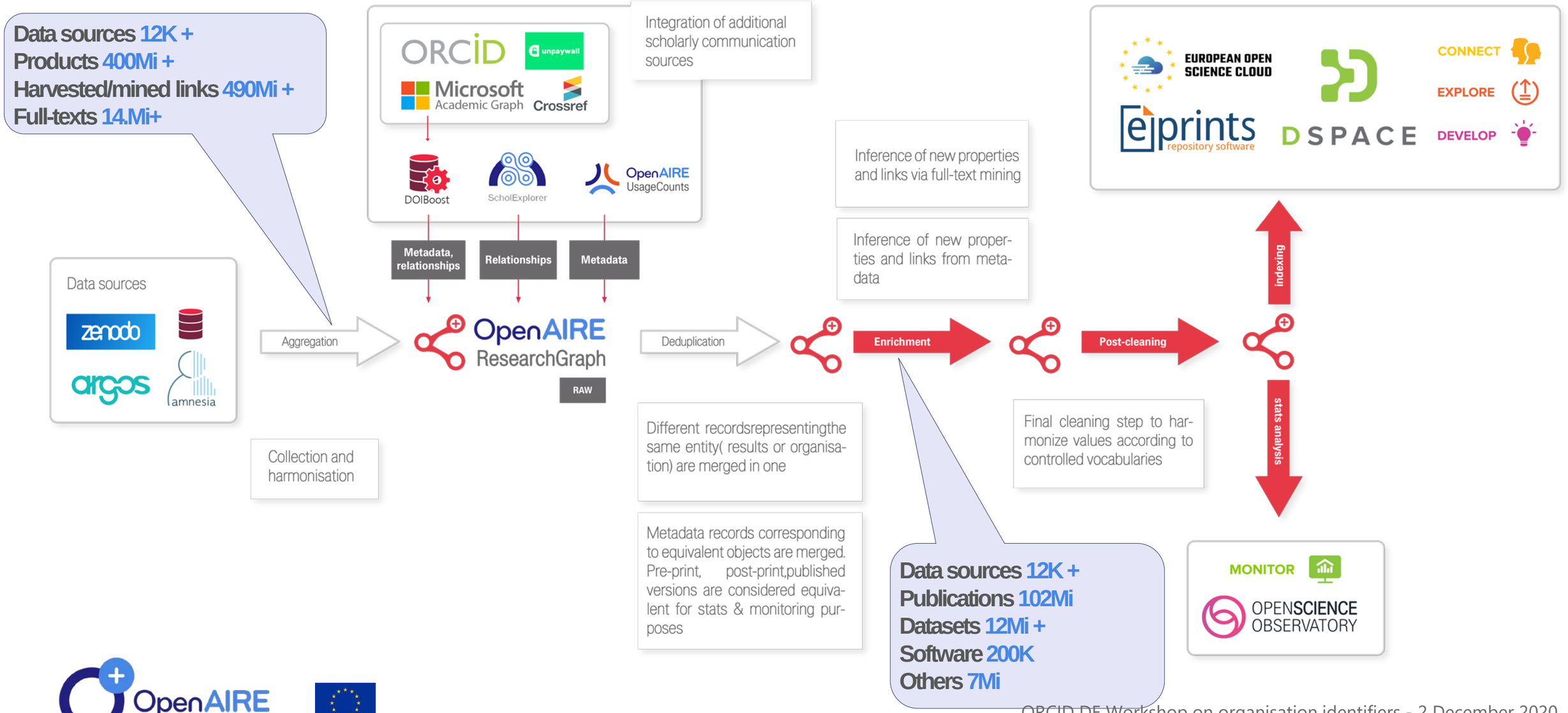
EUROPEAN OPEN SCIENCE CLOUD

…Clusters and RIs service catalogues

OpenAIRE

European Commission

OpenAIRE Research Graph: the supply chain



APIs integration with third-party services

https://develop.openaire.eu

EUROPEAN OPEN SCIENCE CLOUD

The Graph as EOSC Resource Catalogue

2030 CONNECT
a United Nations online technology platform for the SDGs

DARIAH-EU · EGI · EPOS EUROPEAN PLATE OBSERVING SYSTEM · elixir

Research infrastructures

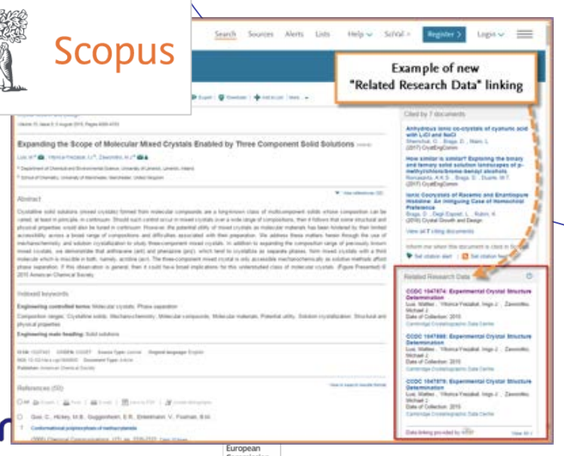
AGINFRA · instruct ERIC · SDSN GREECE · CLARIN Common Language Resources and Technology Infrastructure

ORCID Early 2021
Connecting Research and Researchers

SciVal

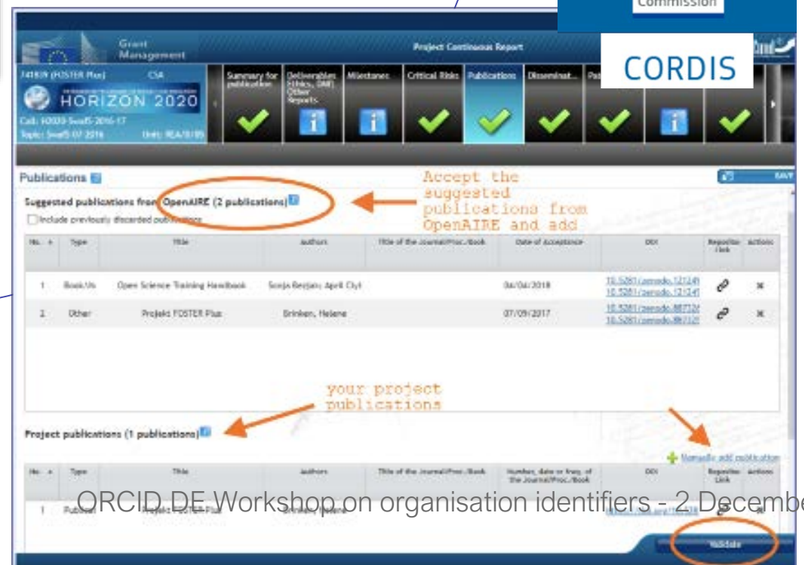
OpenAIRE RESEARCH GRAPH

Scopus



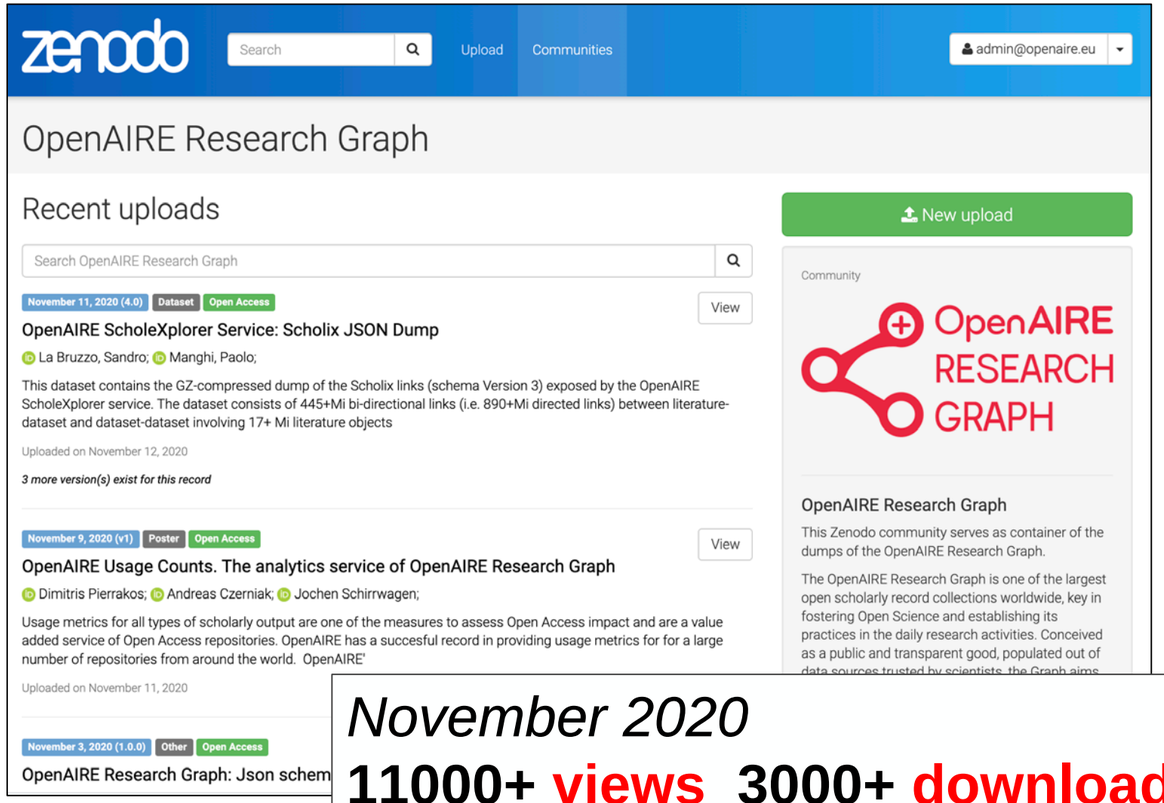
Mendeley

European Commission
CORDIS



OpenAIRE Research Graph: Dumps

- OpenAIRE Research Graph
- OpenAIRE COVID-19
- OpenAIRE communities
- DOIBoost
- Scholexplorer



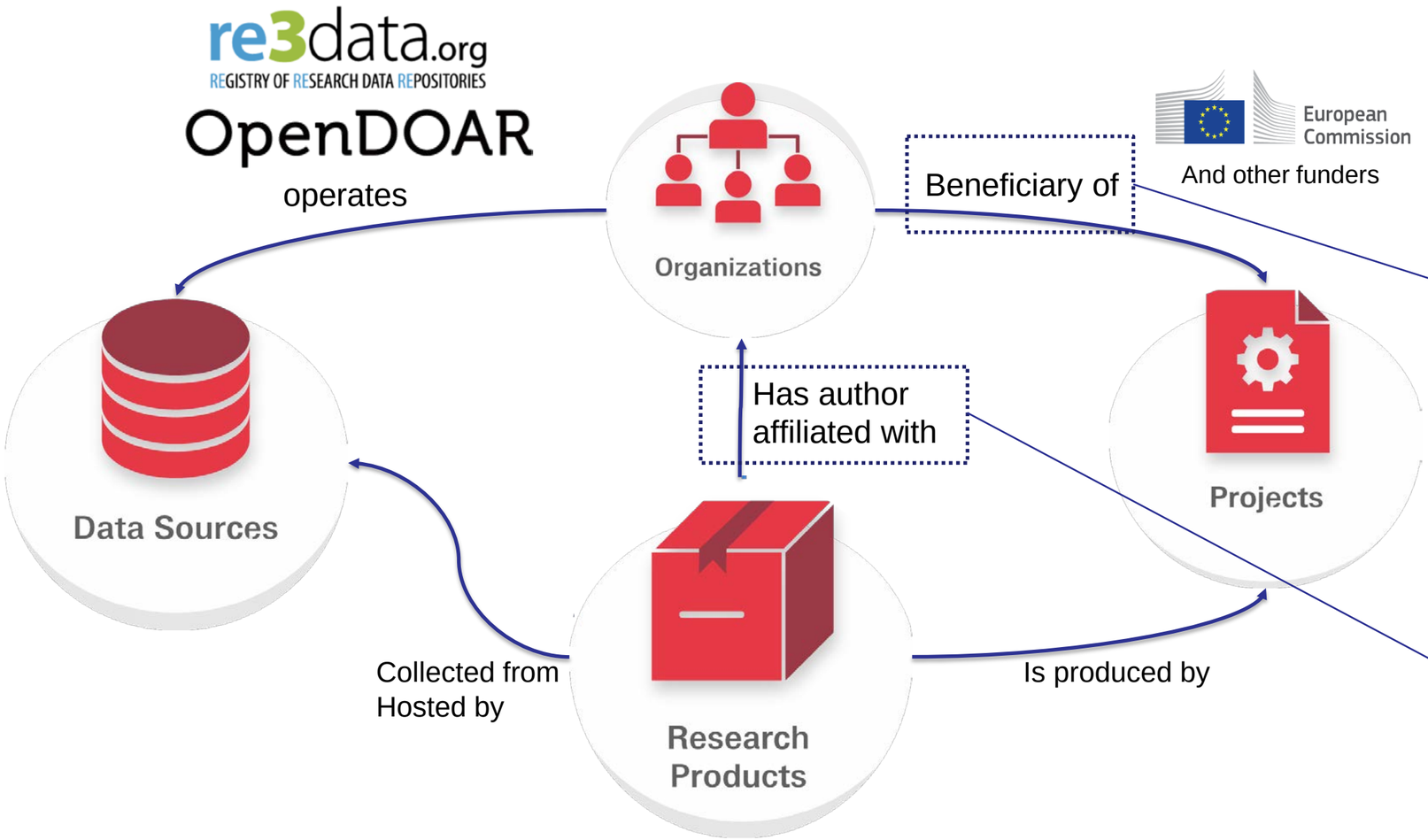
November 2020

11000+ views 3000+ downloads

https://zenodo.org/communities/openaire-research-graph

Organisations in OpenAIRE

Organizations in OpenAIRE



3M relationships involving 140K organisations and 2.5M research projects

25M relationships involving 78K organisations and 18M research products:

- 210K collected from Microsoft Academic Graph
- 7M via full-text mining
- 18M propagated because the organisation operates the institutional repository hosting the research product

Import phase

OpenDOAR → 5K, links to data sources, No PID, OpenAIRE ID based on legal name and country

re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES → 5K, links to data sources, 1K with ROR and/or RRID and/or VIAF PIDs. OpenAIRE ID based on legal name and country

European Commission CORDIS EU research results → 64K, links to projects. With PIC numbers.

GRID → 97K. PIDs: GRID.ac. Other PIDs to be included.

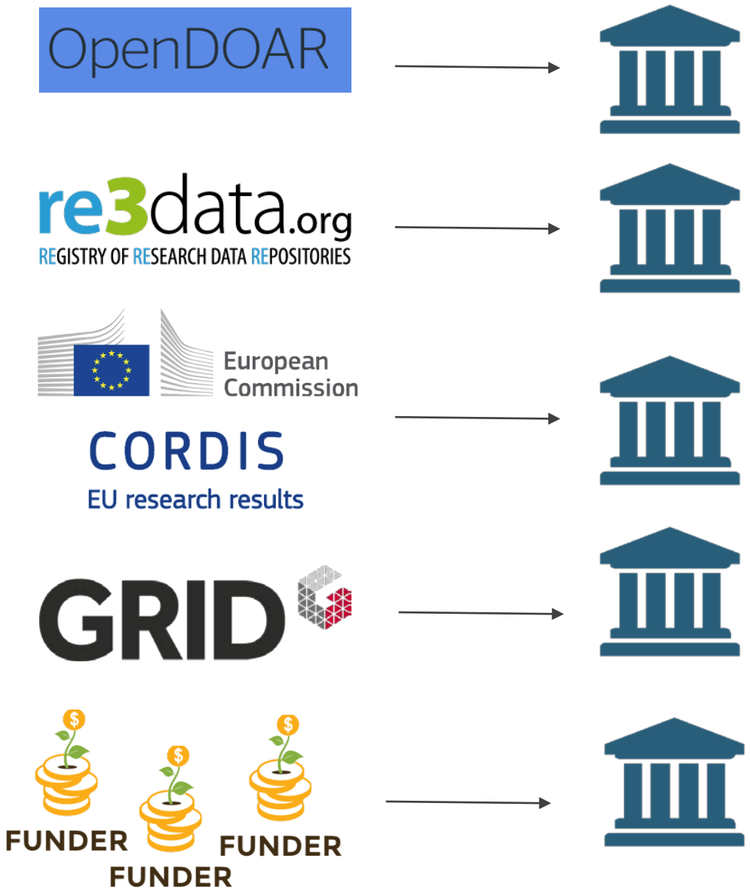
ROR → Coming soon

FUNDER FUNDER FUNDER → 86K, links to projects. No PIDs. OpenAIRE ID based on legal name, unless a local, stable identifier is aavilable

Imported 330K organisations
Lack of common identifiers
PIDs not referenced
=
Duplication

OpenAIRE

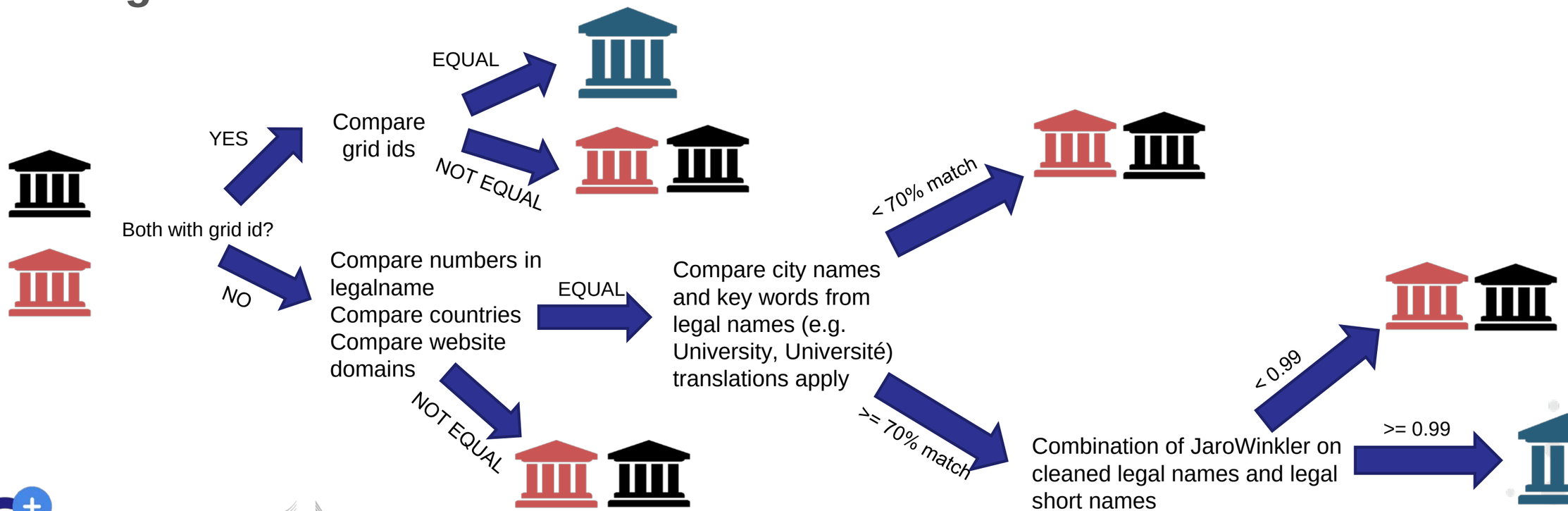
How to identify the duplicates?
A matter of data quality and completeness



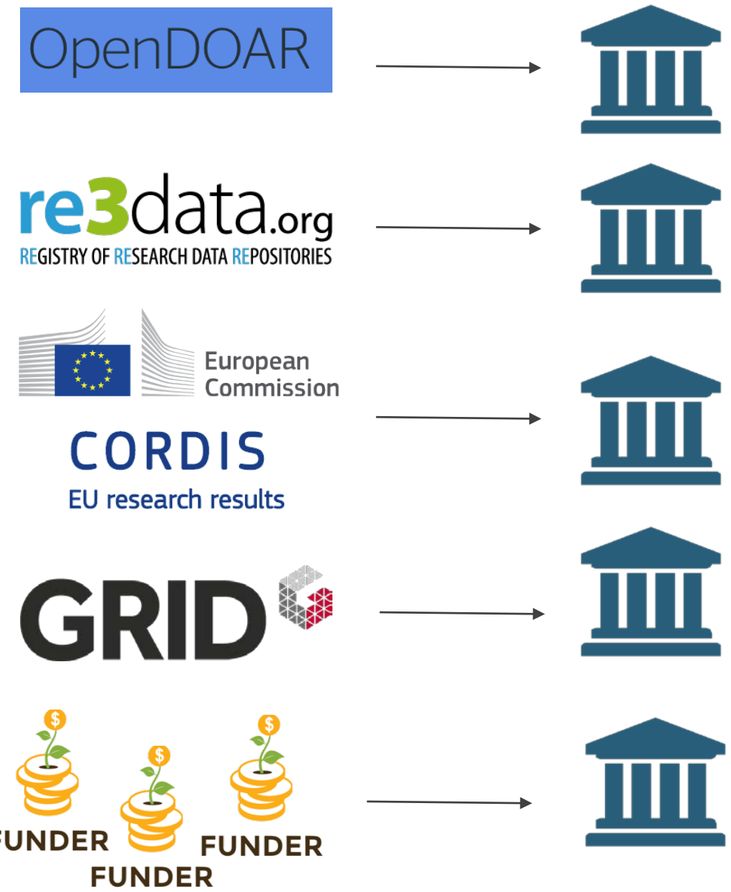
organization field completeness		
field	**missing**	**% missing**
legalname	254	0.1%
legalshortname	156,923	43.2%
alternativenames	319,825	88.0%
country	82,002	22.6%
websiteurl	142,047	39.1%
pid	184,536	50.8%

The deduplication algorithm in a nutshell

- Cluster organisations by legalname and by website URL (4 clustering functions)
- For each cluster we perform pairwise comparations with early exit strategies.



Drawbacks



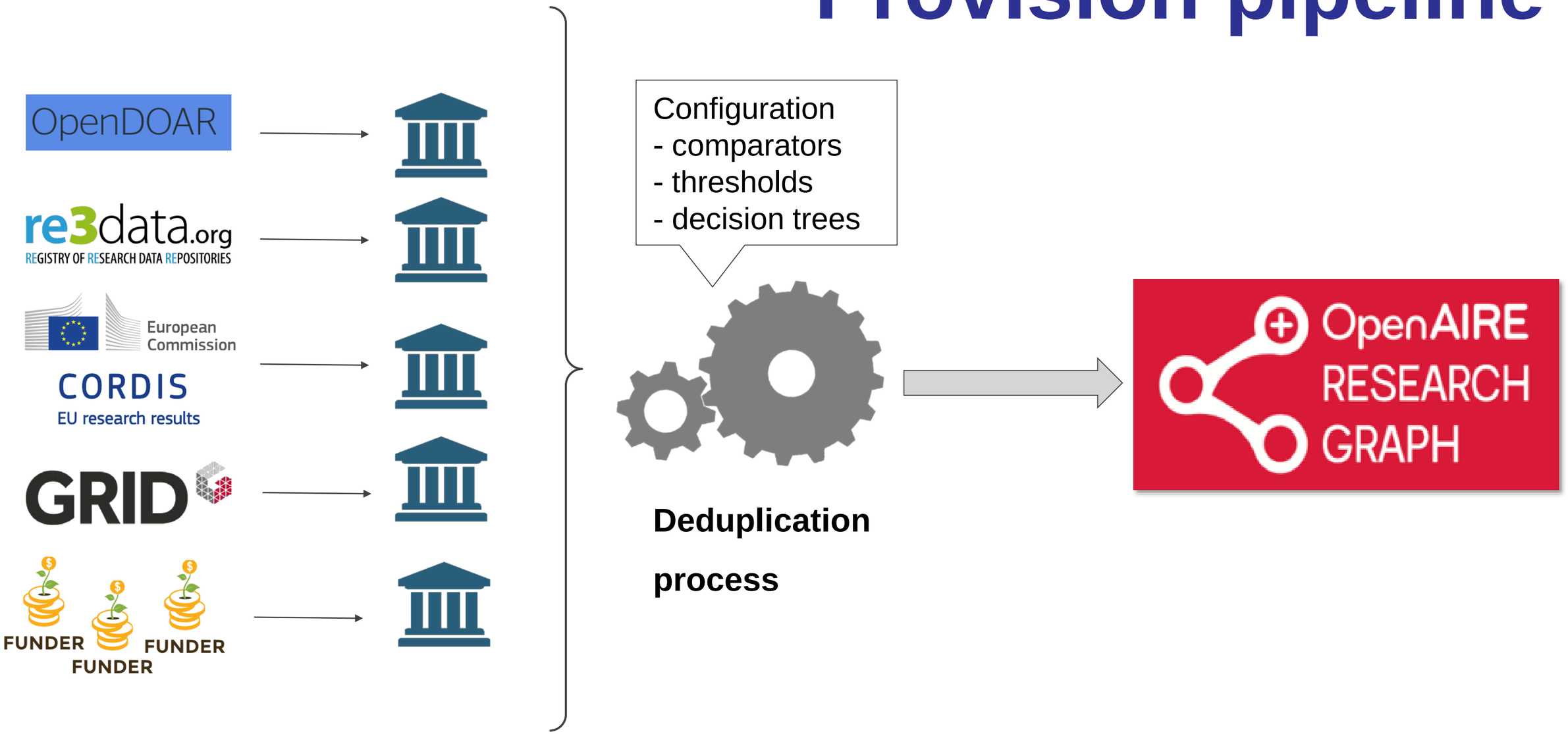
False positives

- Mess up statistics (e.g. for institutional and country monitors)
- Mess up search functionality (e.g. finding only the US Concordia University and not the Canadian one)

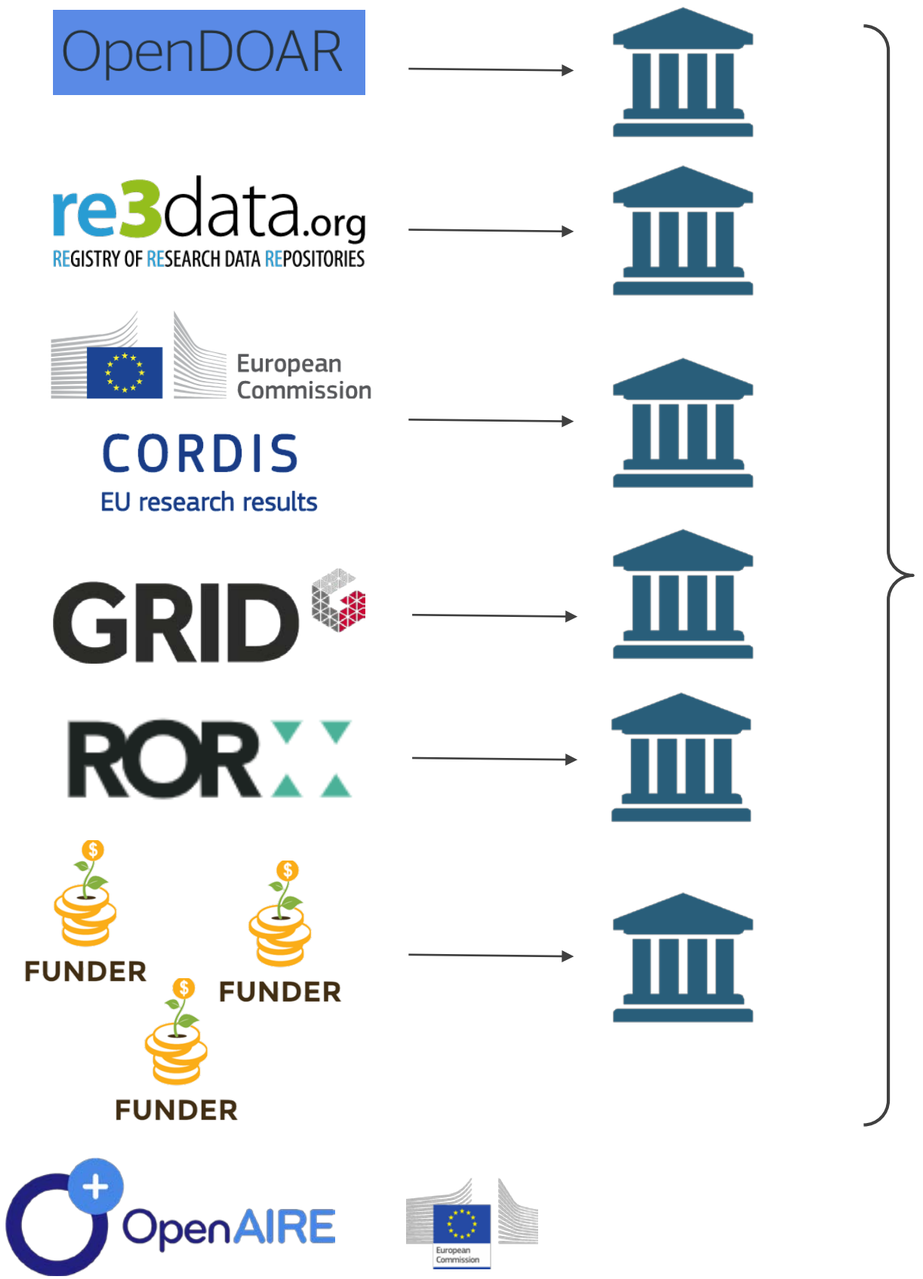
False negatives

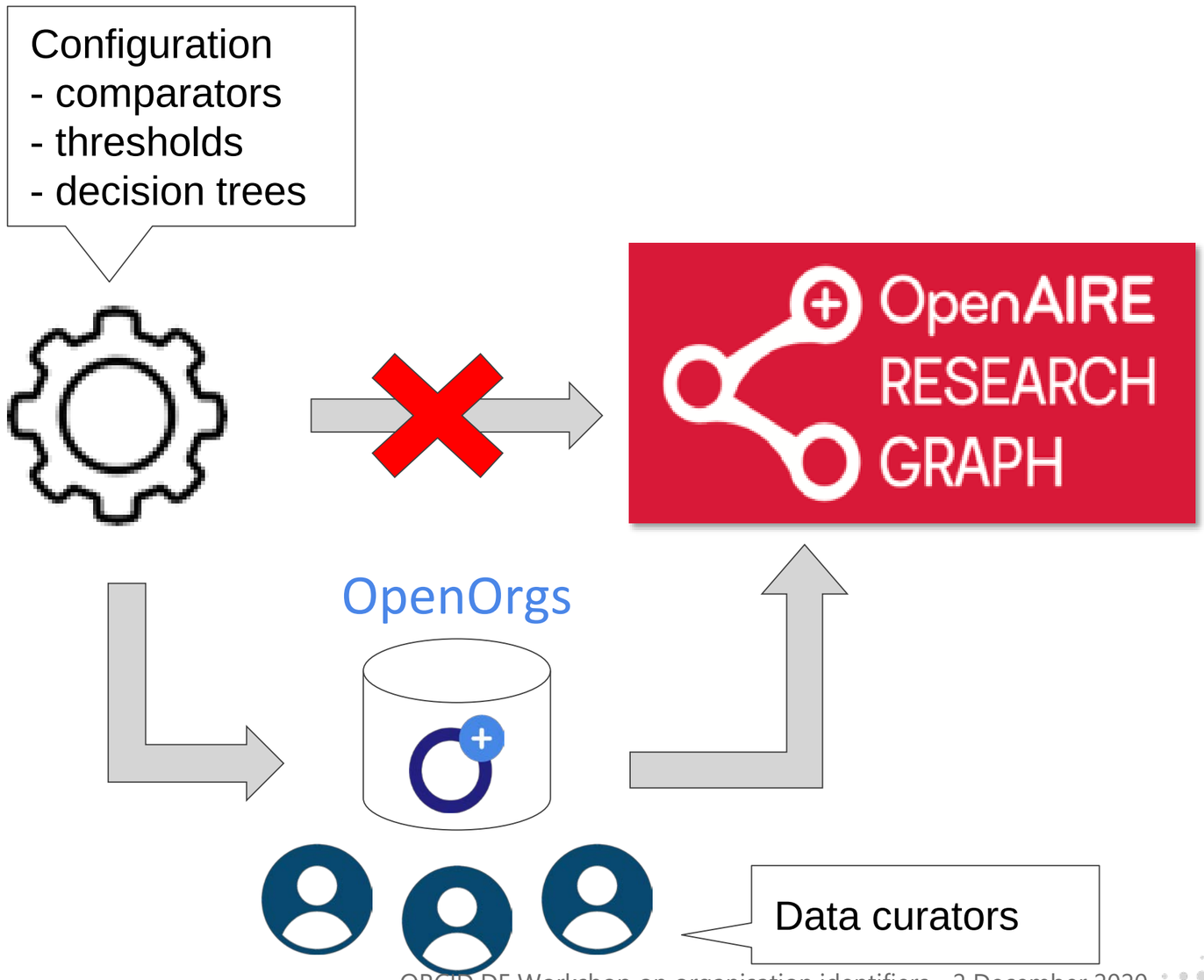
- One instance is linked to the papers, one to the project…but it is the same organization!

Provision pipeline



Provision pipeline





Next steps

- Operational OpenOrgs for OpenAIRE National Open Access Desks
- Import organisation mentions from repository records (CRIS systems, repository compliant to OpenAIRE guidelines 4)
- Notify missing PIDs to OpenAIRE sources (via the OpenAIRE Broker)
- Provide the sub-graph of entities with PIDs and bridge among different PID systems

Thank you!

Alessia Bardi (CNR-ISTI)

Alessia.bardi@isti.cnr.it

graph.openaire.eu