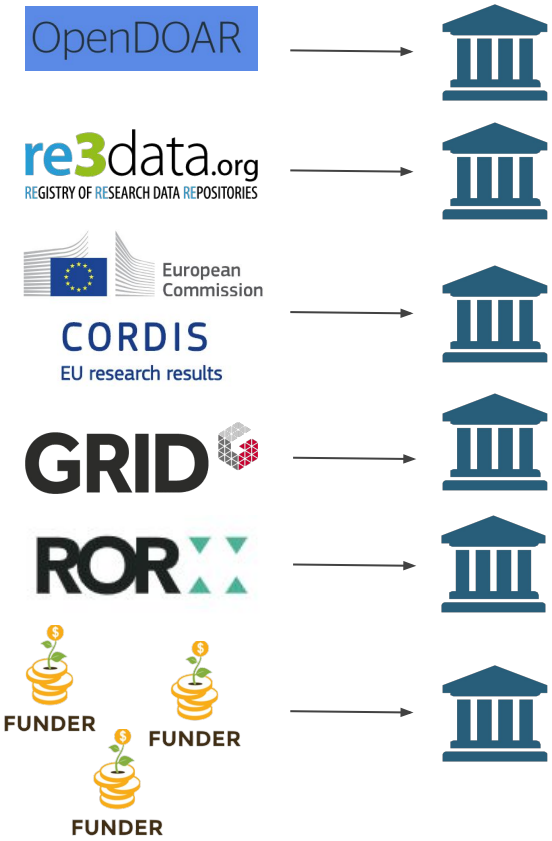
Claudio Atzori, Gina Pavone
Institute of Information Science and Technologies
Consiglio Nazionale delle Ricerche - IT

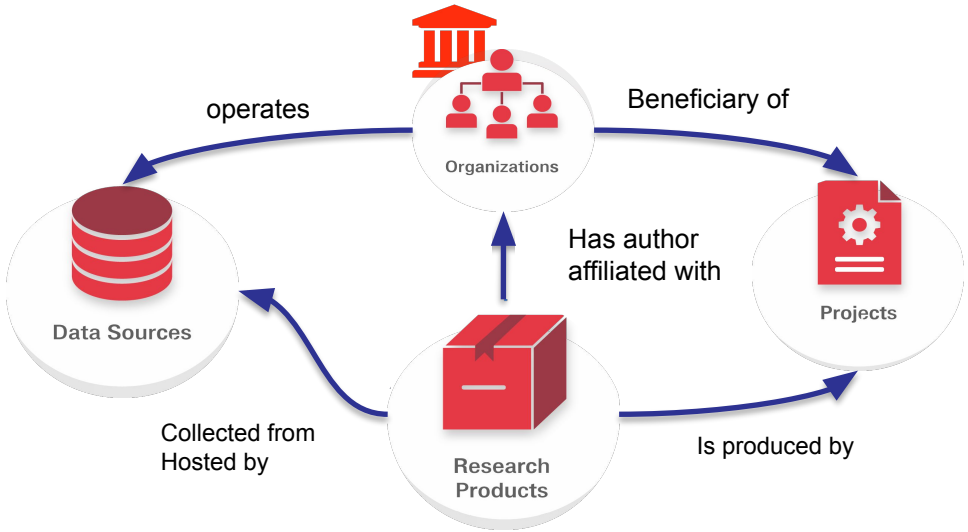
Bridging registries of research organizations

Supporting disambiguation and improving the quality of data

Open Science Fair, September 2021

Organizations in OpenAIRE





Organizations (mentions)

1. Responsible body for the management of a data provider
2. Project participant
3. Extracted from the article full-text as paper affiliation

Lack of common identifiers!

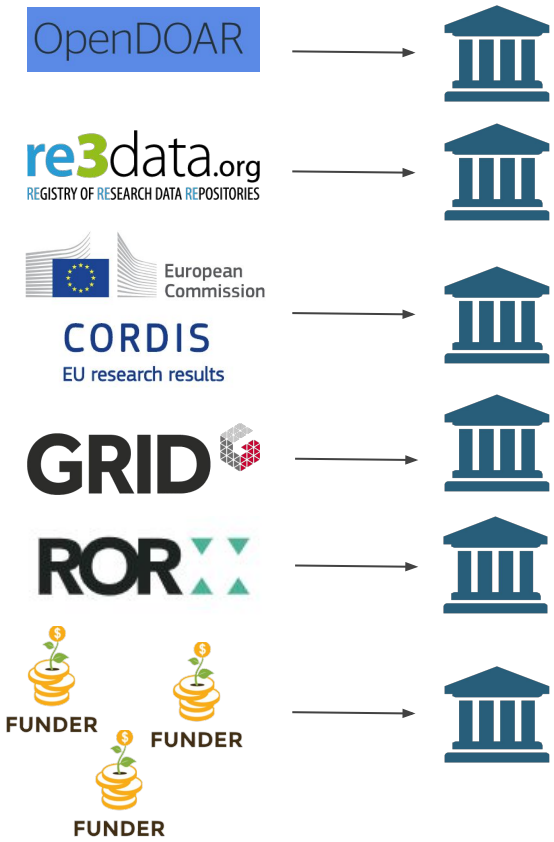
SAPIENZA
UNIVERSITÀ DI ROMA

1 Organization

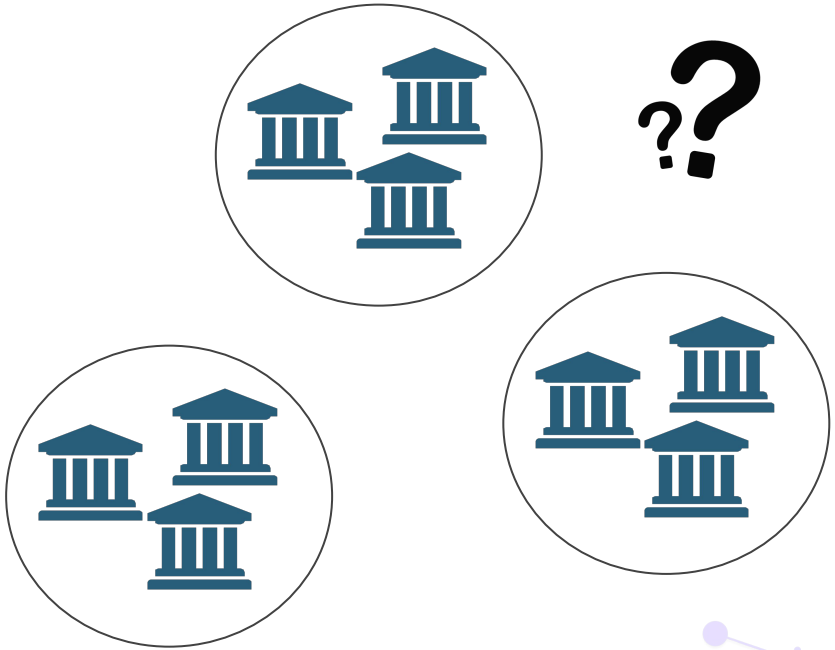
Many different names

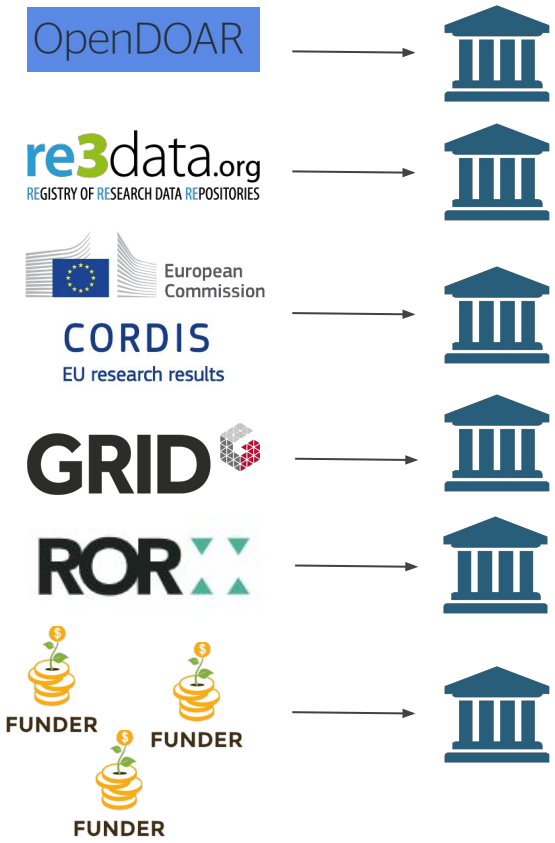
Many data sources

Related organization	Acronym	Country	Source
Sapienza University of Rome **URL:**		IT	**Original Id:** chistera____::8fe1974c2bdc8e69b4e2f0dde01c6f2c **Provenance:** CHIST-ERA
Sapienza Università di Roma **URL:**		IT	**Original Id:** chistera____::d216f9fa7ff12d036ad62203e1b6183d **Provenance:** CHIST-ERA
Sapienza Università di Roma **URL:**	Sapienza Università di Roma	IT	**Original Id:** doajarticles::Sapienza_Università_di_Roma **Provenance:** DOAJ-Articles
University La Sapienza of Rome **URL:**	University La Sapienza of Rome	IT	**Original Id:** doajarticles::University_La_Sapienza_of_Rome **Provenance:** DOAJ-Articles
Università degli Studi di ROMA 'La Sapienza' **URL:**		IT	**Original Id:** miur_____::49e607219c94f97237f12ffb82e29f69 **Provenance:** Ministero dell'Istruzione dell'Università e della Ricerca
Università degli studi La Sapienza di Roma **URL:** http://www.uniroma1.it/		IT	**Original Id:** opendoar____::Università_degli_studi_La_Sapienza_di_Roma_IT **Provenance:** OpenDOAR
Università degli Studi di ROMA "La Sapienza" **URL:** http://www.uniroma1.it	UNIROMA1	IT	**Original Id:** orgreg_____::IT0068 **Provenance:** OrgReg
Sapienza University of Rome **PID (GRID):** grid.7841.a **PID (OrgRef):** 1222318 **PID (ROR):** https://ror.org/02be6w209 **PID (FundRef):** 100010143 **PID (FundRef):** 501100004271 **PID (Wikidata):** Q209344 **URL:** http://www.uniroma1.it/	Sapienza University of Rome	IT	**Original Id:** ror_____::https://ror.org/02be6w209 **Provenance:** ROR
Sapienza University of Rome **URL:**		IT	**Original Id:** snsf_____::Sapienza_University_of_Rome **Provenance:** SNSF - Swiss National Science Foundation
University of Rome La Sapienza **URL:**		IT	**Original Id:** snsf_____::University_of_Rome_La_Sapienza **Provenance:** SNSF - Swiss National Science Foundation



Problem statement disambiguation





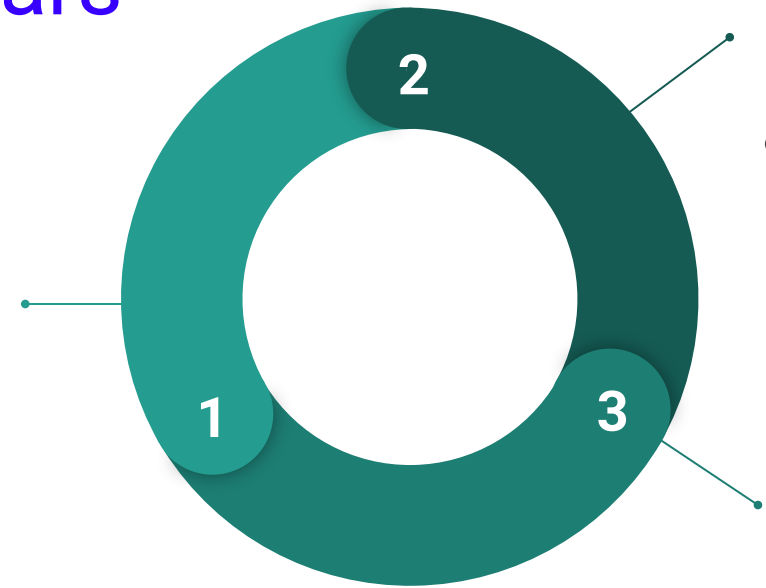
Problem statement quality / completeness

organization field completeness (pre - dedup)		
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%

OpenOrgs activity pillars

Automatic suggestion of duplicates

The process for the automatic identification of possible duplicates will periodically produce new suggestions for the data curators.

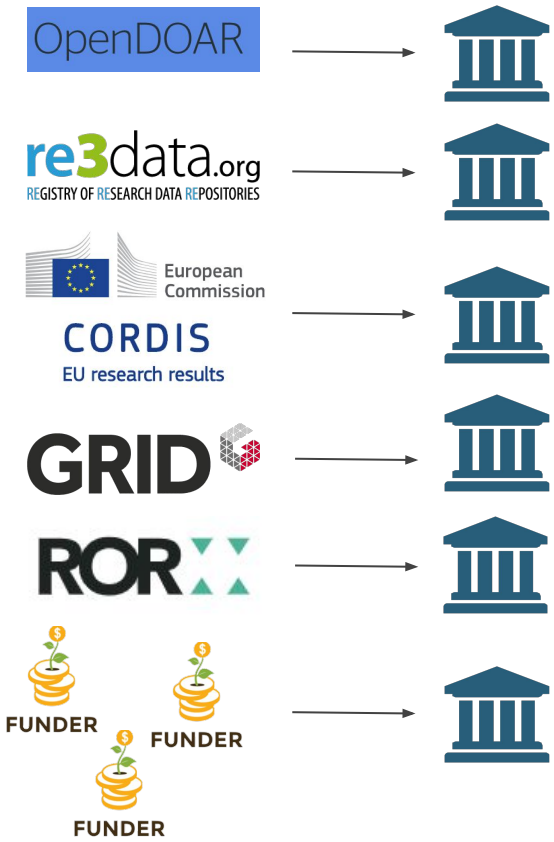


Management of duplicates

Deciding whether two or more digital representations are duplicates of the same object can be a task that only humans can carry out precisely. The OpenOrgs tool allows data curator to manage the ambiguities in the data lack of the organization mentions aggregated by OpenAIRE.

Metadata curation

In OpenOrgs data curators can enrich the metadata description of the organization entities, compensating the lack of information available from the sources and improving the discoverability / completeness of the organization records.



Limitations of a fully automated approach

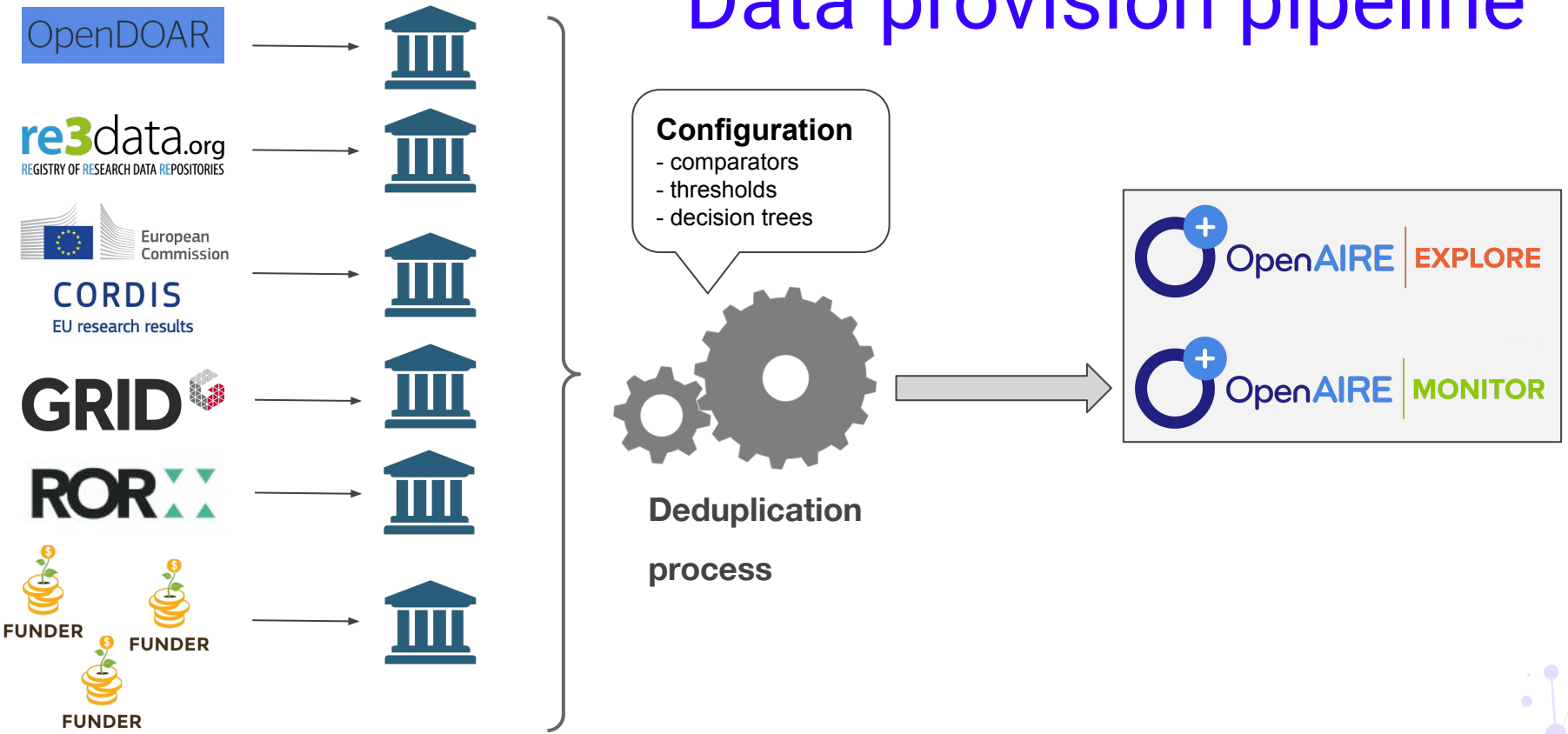
False positives

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

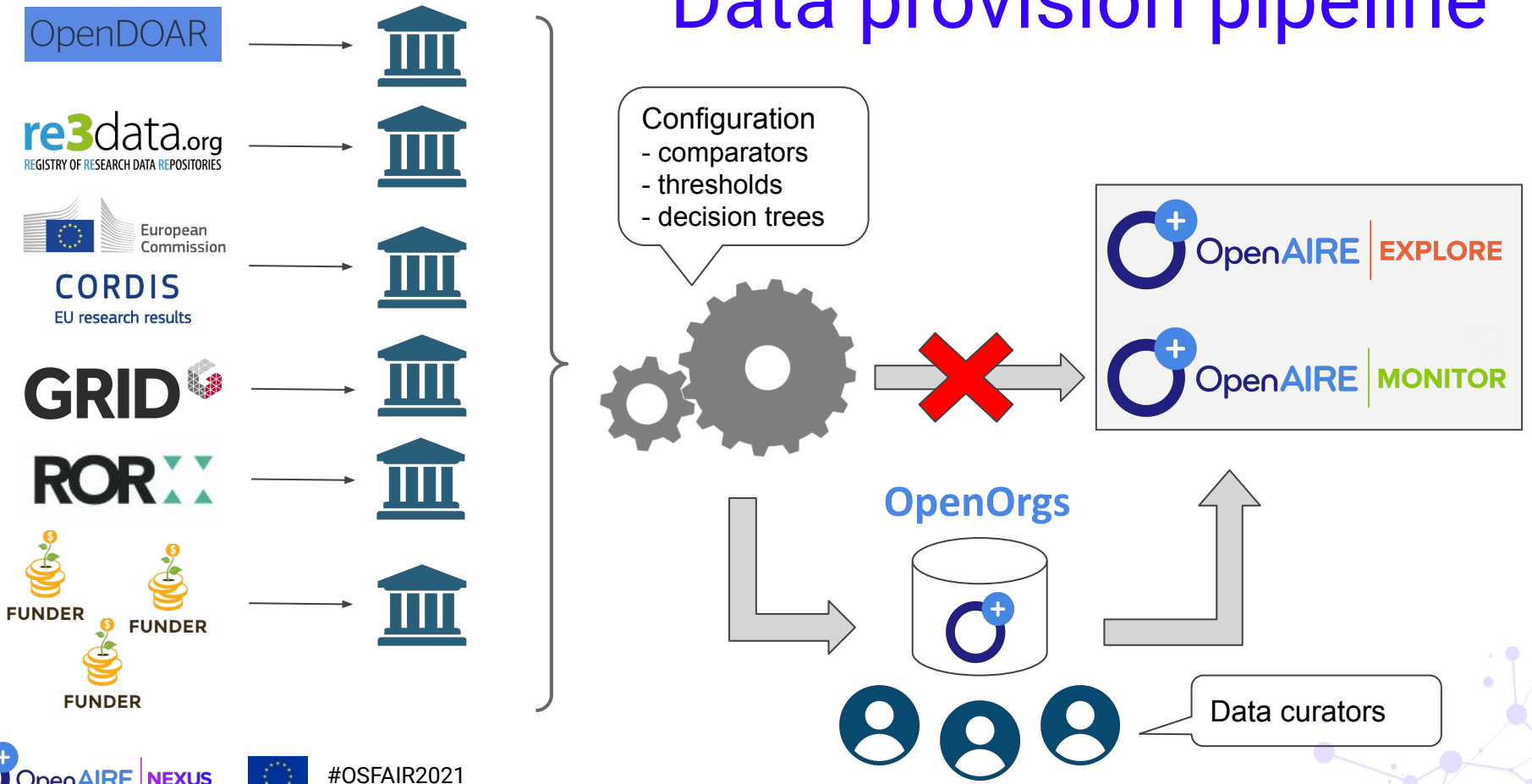
False negatives

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

Data provision pipeline



Data provision pipeline



User roles

Simple User

Scope of data curation is limited to a set of countries
Metadata editing & enrichment
Approval of suggestions from the automated duplicate identification
Creation of a new Organization → to be approved by national admin

National admin

Scope of data curation is limited to a set of countries
Metadata editing & enrichment
Approval of pending organizations → generates approved Orgs
Grant access to simple users for the selected countries
Resolution of conflicts

A look at the main functionalities. Let's see:

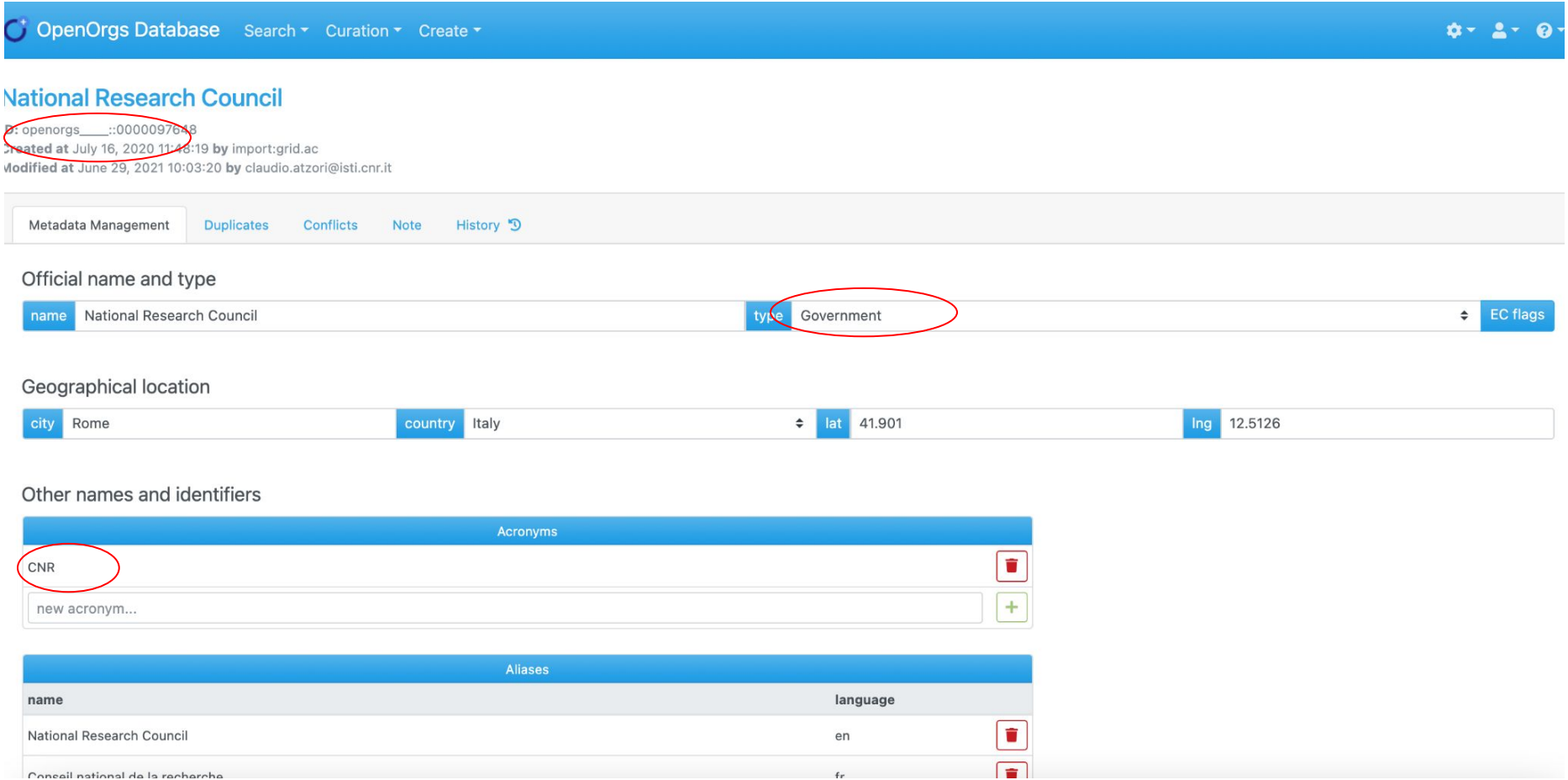
- How a curated org appears: **National Research Council** - Italy
- An organization with wrong duplicates suggested by the algorithm: **Istituto Ortopedico Galeazzi**
- How to curate metadata and approve a new org inside OpenOrg: **Center for Research and Telecommunication Experimentation for Networked Communities**
- Resolution of potential conflicts**: San Camillo Forlanini**

Besides you can:

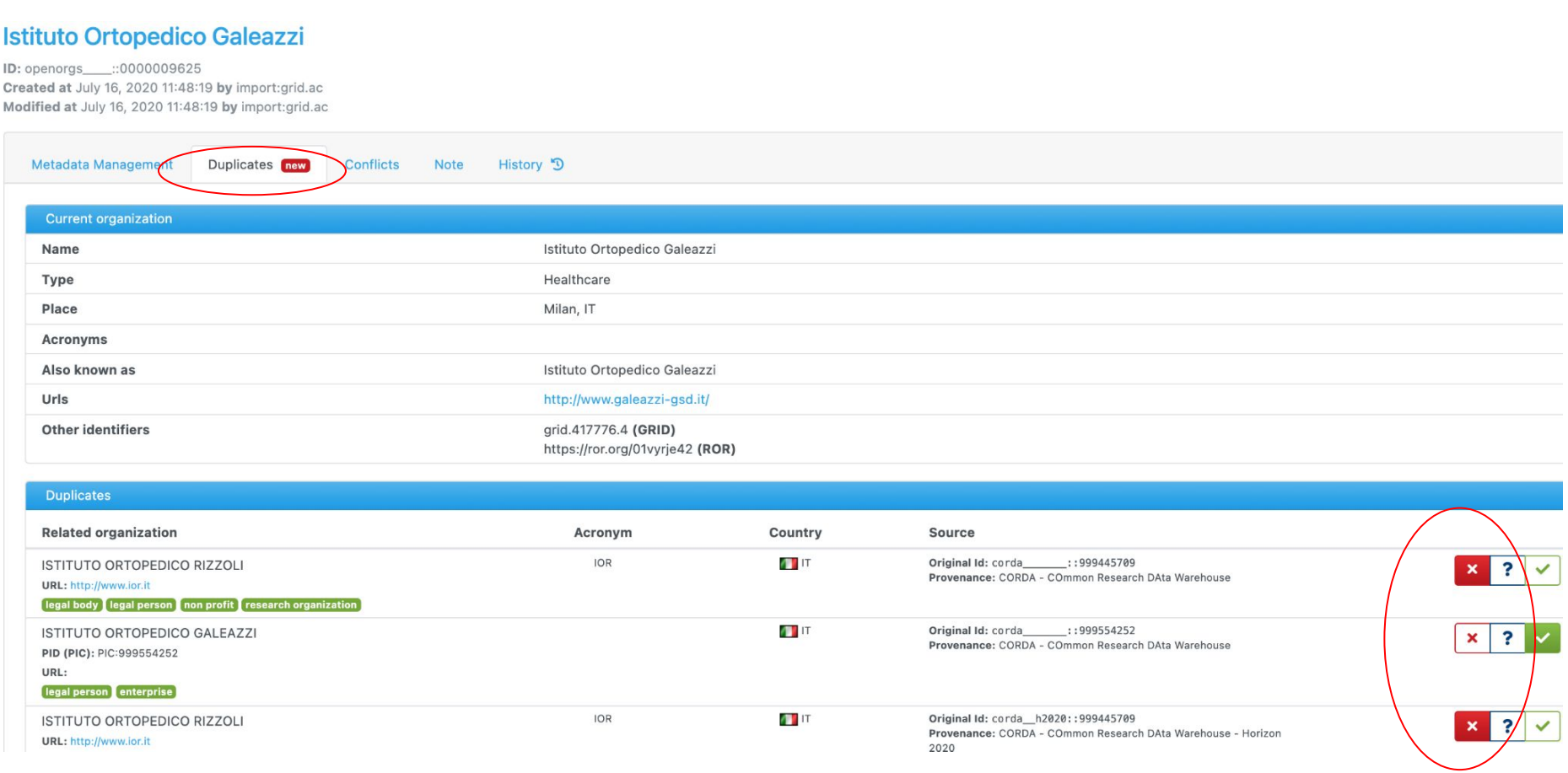
- Perform a simple search (also with part of the name)
- Browse by type of org (and by country if you are curating more than one county)
- Browse only pending orgs (orgs to be approved)
- Browse only duplicates
- Insert a new org

For more details please see: https://zenodo.org/record/5101096#.YUNC_p4zY6E

- How a curated org appears: **National Research Council** - Italy



- An organization with wrong duplicates suggested by the algorithm: I**stituto Ortopedico Galeazzi**



Istituto Ortopedico Galeazzi

ID: openorgs____::0000009625
Created at July 16, 2020 11:48:19 **by** import:grid.ac
Modified at July 16, 2020 11:48:19 **by** import:grid.ac

Metadata Management | Duplicates new | Conflicts | Note | History

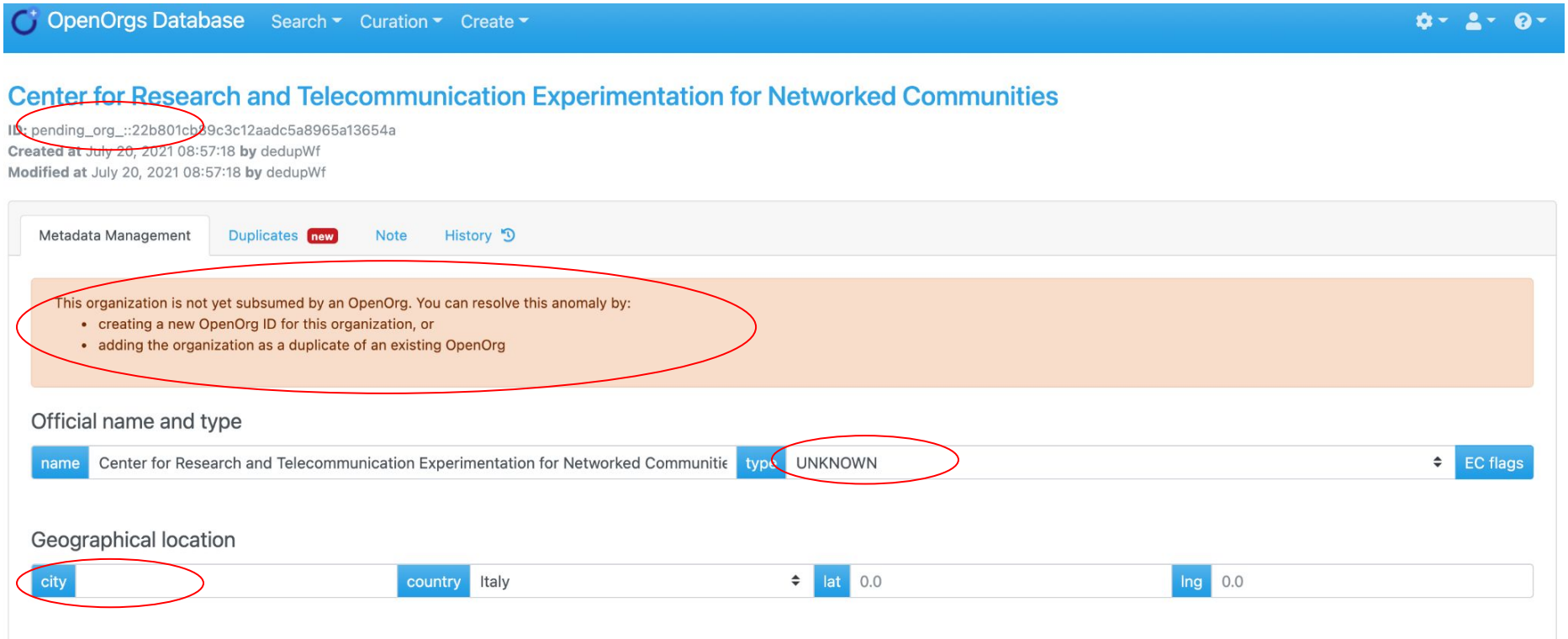
Current organization

Name	Istituto Ortopedico Galeazzi
Type	Healthcare
Place	Milan, IT
Acronyms	
Also known as	Istituto Ortopedico Galeazzi
Urls	http://www.galeazzi-gsd.it/
Other identifiers	grid.417776.4 **(GRID)** https://ror.org/01vyrje42 **(ROR)**

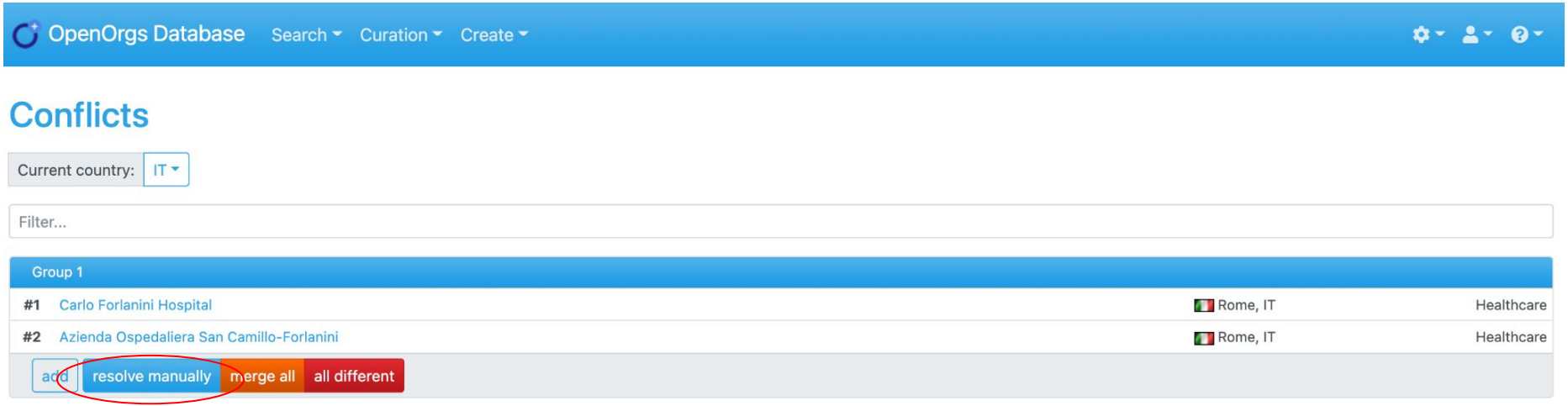
Duplicates

Related organization	Acronym	Country	Source
ISTITUTO ORTOPEDICO RIZZOLI **URL:** http://www.ior.it legal body · legal person · non profit · research organization	IOR	IT	**Original Id:** corda_____::999445709 **Provenance:** CORDA - COmmon Research DAta Warehouse
ISTITUTO ORTOPEDICO GALEAZZI **PID (PIC):** PIC:999554252 **URL:** legal person · enterprise		IT	**Original Id:** corda_____::999554252 **Provenance:** CORDA - COmmon Research DAta Warehouse
ISTITUTO ORTOPEDICO RIZZOLI **URL:** http://www.ior.it	IOR	IT	**Original Id:** corda__h2020::999445709 **Provenance:** CORDA - COmmon Research DAta Warehouse - Horizon 2020

- How to curate metadata and approve a new org inside OpenOrg: **Center for Research and Telecommunication Experimentation for Networked Communities**



- Resolution of potential conflicts**: San Camillo Forlanini**

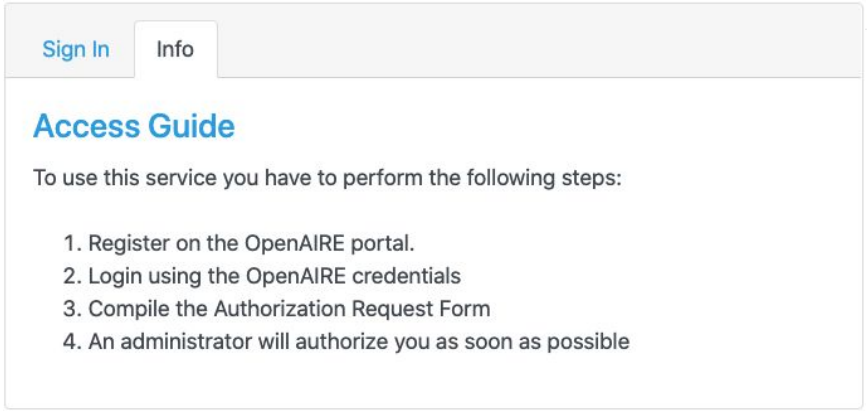


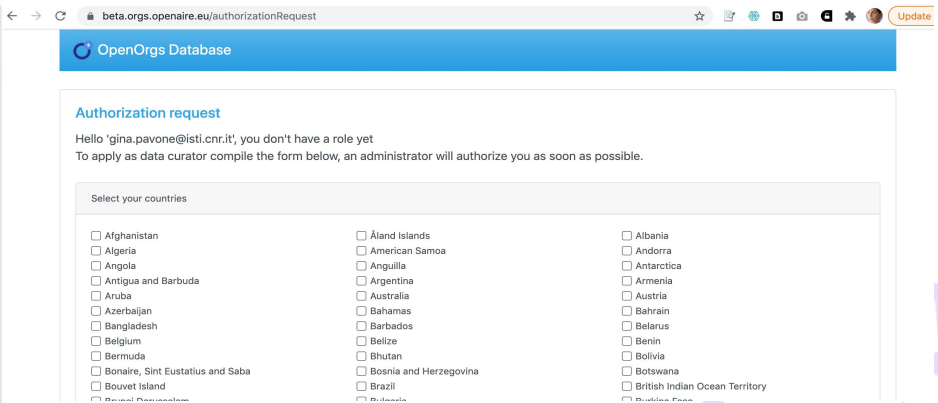
Next steps & Future developments

- Promotion to production
- Engage new curators, extend the data curation teams
- Integrate with national institution registries
- Front face portal to explore the data
- Develop an API for 3rd parties to consume the PID correspondences

How to join

1. Create an OpenAIRE account
2. Sign in @ https://beta.orgs.openaire.eu
3. You may choose different countries that you want to curate
4. Get in contact with admins: **openorgs-admin@openaire.eu**





Thanks!

claudio.atzori@isti.cnr.it

gina.pavone@isti.cnr.it

Support: openorgs-admin@openaire.eu