

Gina Pavone, Claudio Atzori

Institute of Information Science and Technologies

Consiglio Nazionale delle Ricerche - IT

OpenOrgs - the disambiguation tool for your organization

Improving the quality of data in the scholarly communication ecosystem

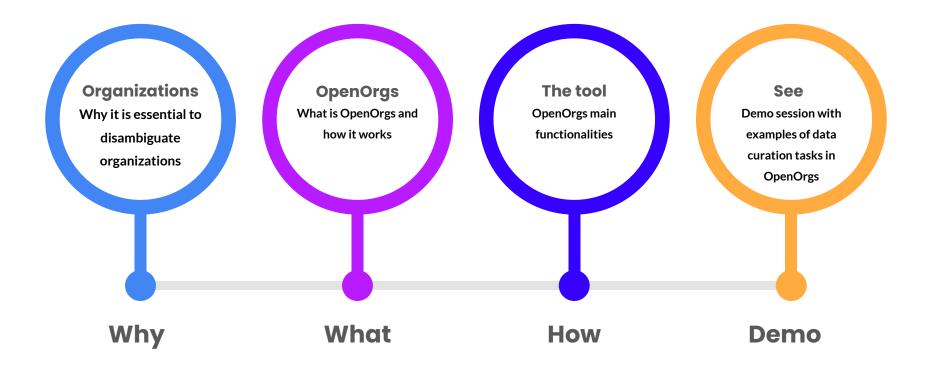
OpenAIRE - 28 June 2022





Today's agenda

OpenOrgs 2.0 training - June 28



Organizations - Information ambiguity

Information about the organizations is distributed in numerous data sources

Organizations may appear with different names (legal names, acronyms, short names, alternatives...)

The organization's structure may not be so clear: faculties, departments, branches, detachments etc. But in order to have an efficient scholarly communication system, it is necessary to disambiguate this information





1 Organization

Many different names

Many data sources

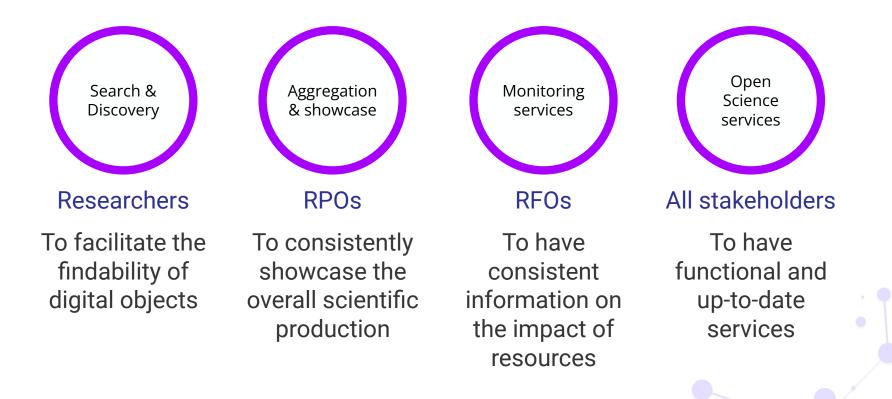


Related organization	Acronym	Country	Source
Sapienza University of Rome URL:		TI 🗖	Original Id: chistera::8fe1974c2bdc8e69b4e2f0dde01c6f 2c Provenance: CHIST-ERA
Sapienza Università di Roma URL:		т	Original Id: chistera::d216f9fa7ff12d036ad62203e1b618 3d Provenance: CHIST-ERA
Sapienza Università di Roma URL:	Sapienza Università di Roma	т 🔳	Original Id: doajarticles::Sapienza_Università_di_Roma Provenance: DOAJ-Articles
University La Sapienza of Rome URL:	University La Sapienza of Rome	т 🔳	OriginalId: doajarticles::University_La_Sapienza_of_Rome Provenance:DOAJ-Articles
Università degli Studi di ROMA 'La Sapienza' URL:		т 🗖	Original Id: miur::49e607219c94f97237f12ffb82e29f 69 Provenance: Ministero dell'Istruzione dell'Università e della Ricerca
Università degli studi La Sapienza di Roma URL: http://www.uniroma1.it/		т	Original Id: opendoar::Università_degli_studi_La_Sapi enza_di_Roma_IT Provenance: OpenDOAR
Università degli Studi di ROMA "La Sapienza" URL: http://www.uniroma1.it	UNIROMA1	تا 🌄	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	π	Original Id: ror::https://ror.org/02be6w209 Provenance:ROR
Sapienza University of Rome URL:		т	Original Id: snsf::Sapienza_University_of_Rome Provenance: SNSF - Swiss National Science Foundation
University of Rome La Sapienza URL:		т	Original Id: snsf::University_of_Rome_La_Sapienza Provenance: SNSF - Swiss National Science Foundation

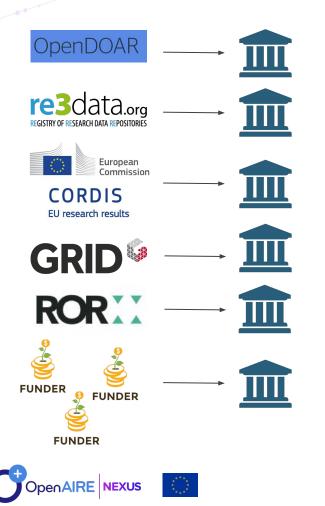
This ambiguity affects all the scholarly communication system



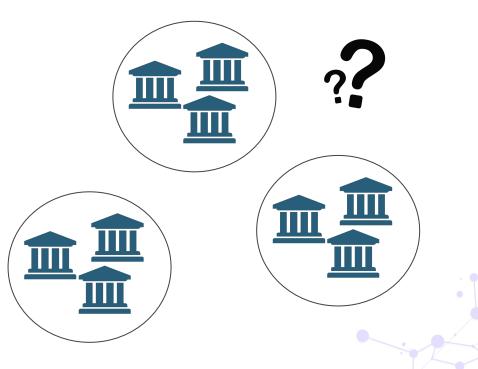
Services needing unambiguous information

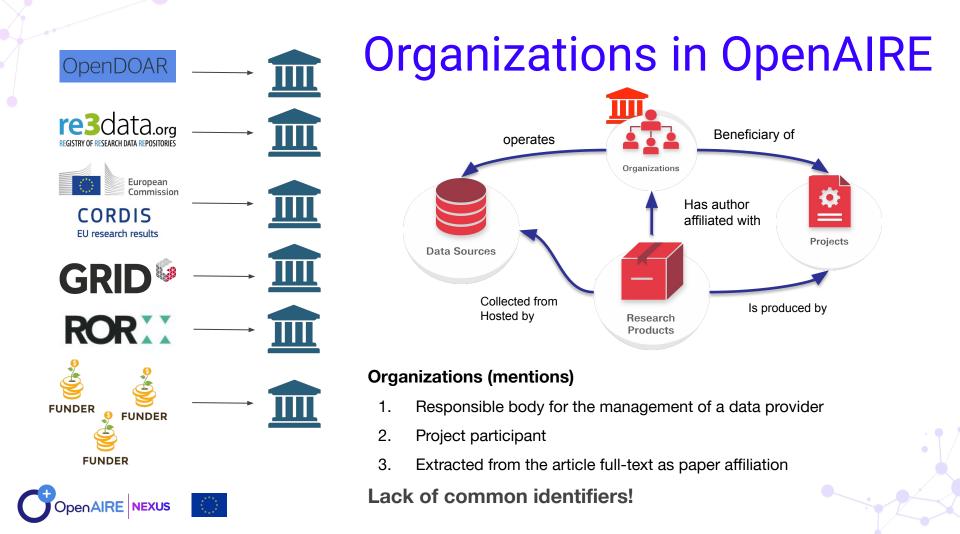


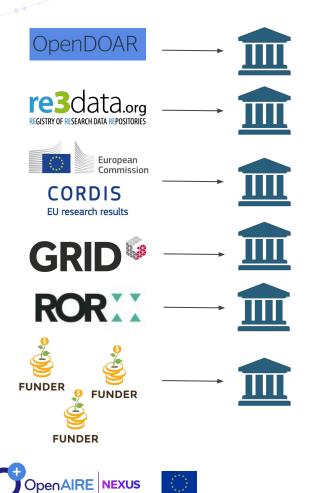




Problem statement disambiguation







Problem statement quality / completeness

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

re3data.org European Commission CORDIS EU research results GR RO FUNDER FUNDER FUNDER



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!

OpenAIRE services needing organization disambiguation

Explore

The portal where to explore the research graph, from scientific products to institutions and funding streams

Institutional d.

A dashboard where single institutions can track their research activities (grants, publications, datasets, software and other), and their performance in areas such as OS, EOSC, collaborations, reach and visibility and so on



Monitor Dashboard

Funders can have an overall view of the scientific production of the resources they allocate

OS observatory

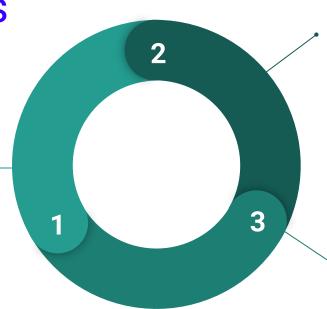
To understand the European research landscape and track trends for open access to publications, data, software



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.



Manual 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.



OpenOrgs curation tasks

Actions in OpenOrgs

All data curators can

- Curate and enrich orgs metadata
- Resolve duplicates suggested by the algorithm in the <u>duplicates section</u>
- Manually add duplicates in the single org metadata page
- Approve <u>suggested orgs</u> and give them stable OpenOrgs IDs
- Insert <u>new orgs</u> inside OpenOrgs
- Leave comments in the <u>note section</u> of each organization
- See <u>history</u> of the actions made on the single org page
- Resolve <u>conflicts</u>



Metadata curation & enrichment

Recommendations and Tips

- Use the English version of the organization name as the main name, and fill in the name in the national language in aliases
- <u>Always try to include the URL</u> of the organization or check that it is correct.
- Use UTF-8 encoding
- If you are not sure what type of organization it is, you can check the EC flags.
- You can add notes inside an org page to keep track of the choices made.
- See in history the actions performed in the org page



Metadata curation & enrichment

← → C
a orgs.openaire.eu/main#!/edit/0/openorgs____::0000097669

🗅 ☆ 🖻 🕘 🖸 🖉 🥕 🖪 🌘

C OpenOrgs Database Search - Curation - Create -

\$* ****** 8*

University of Pisa

NEXUS

nen AIRE

ID: openorgs____::0000097669 Created at July 16, 2020 11:48:19 by import:grid.ac Modified at October 8, 2020 15:04:15 by michele.artini@isti.cnr.it

Metadata Manage	ment Duplicates new Conflicts	Note History 5		
Official name	and type			
name Universit	y of Pisa		type Education	¢ (EC flags
Geographical	location			
city Pisa	country	Italy	♦ lat 43.716429	Ing 10.398687
Other names	and identifiers			
		Acronyms		
UniPi			1	
new acronym			+	
		Aliases		
name		Allases	language	
Università di Pisa			it 🔳	

Connecting different ids of the same org

🗅 🖈 🖹 🛞 🖬 🙆 🕻 🥕 👫 🖪 🍘

← → C
a orgs.openaire.eu/main#!/edit/0/openorgs___::000009766

language name Î Università di Pisa it Î University of Pisa en Ť Université de Pise fr **T** Universität Pisa de + new alias... language ... \$

id type Ť 0000 0004 1757 3729 ISNI Ť grid.5395.a GRID T 453158 OraRef T 501100007514 FundRef Î 0645663 Wikidata new id... \$ + type. Ť. http://www.unipi.it/

Metadata curation & enrichment

Establishing relationships between organizations

In OpenOrgs only parent-child relationships are allowed ('is parent of' or 'is child of')

Please note: this means that <u>everything produced by the child will be included in the</u> <u>parent's production</u>.

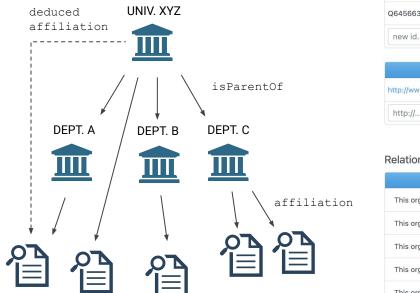
Taking care of relationships, such as between universities and departments or between the main org and its institutes, is very important for example for production monitoring.

• When curating relations between orgs (ie. departments or institutes inside an org), <u>the department or institute must first be approved as an organization</u>, and then linked as child to the main org.



An example of relations:

orgs.openaire.eu/main#!/edit/0/openorgs____::0000097669 C



Q645663	Wikidata
new id	type +

	Urls
http://www.unipi.it/	Î
http://	+

Relations

Relations							
This organizazion is parent of Department of Chemistry and Industrial Chemistry - Università di Pisa							
This organizazion is parent of Dipartimento di Civiltà e Forme del Sapere - Università di Pisa							
This organizazion is parent of Dipartimento di Civiltà e Forme del Sapere - Università di Pisa							
This organizazion is parent of Dipartimento di Civiltà e Forme del Sapere - Università di Pisa							
This organizazion is parent of Dipartimento di scienze economiche - Università di Pisa							
This organizazion is parent of Dipartimento di Studi Italianistici - Università di Pisa							
This organization rel type + related organization	+						

Duplicates curation

Main concept: duplicates

They are possible duplication of the same organization, to be resolved by the user.

A duplicate is created between an approved org and a raw org (an external org with no OpenOrg ID, ie coming from Corda, Re3data, ROR...).

Duplicates can be suggested by the deduplication algorithm or added by the user inside an org page when he/she finds that there are duplicates not detected by the automatic process.



Duplicates curation in two steps

1. Complete the automated part

Duplicates Section \rightarrow There is a list of organizations that an algorithm suggests are the same.

So it is an automated process to be confirmed or rejected by the data curator.

2. Manual entry

Duplicates addition in the metadata org $page \rightarrow$ The data curator manually adds duplicates in the main org page.

Hand-added duplicates do not go into the list of possible duplicates suggested by the algorithm



Potential new duplicates section

Curation menu -> Organizations with new duplicates

\leftarrow \rightarrow C $($ orgs.openaire.eu/main#!/sea	Û	☆	2	B (Ö	e ,	* *	L 🧃) E	
C OpenOrgs Database Search	Curation Create						¢٠	1 - (9 -
	Suggested organizations 9137 Organizations that have been suggested by an User or by the Dedup OpenAIRE Workflow								
	Organizations with new duplicates 160 Organizations that presents ingulicates suggested by the Dedup OpenAUTE Workflow								
	Potential conflicts 1 Approved Organizations that need to be merged								
	Searc	ch							





Registered organization	
Name	Consortium GARR
Туре	Nonprofit
Place	Rome, IT
Acronyms	
Also known as	Gruppo per l'Armonizzazione delle Reti della Ricerca Consortium GARR
Urls	http://www.garr.it/
Other identifiers	grid.423642.5 (GRID) https://ror.org/03x9xd924 (ROR)

Duplicates				
Related organization	Acronym	Country	Source	
CONSORTIUM GARR PID (PIC): 999579084 URL: http://www.garr.it legal body (legal person non profit)	GARR	п	Original Id: corda::999579084 Provenance: CORDA - COmmon Research DAta Warehouse	× ? ×
CONSORTIUM GARR PID (PIC): 999579084 URL: http://www.garr.it legal person non profit	GARR, GESTIONE AMPLIAMENTO RETE RICERCA	п	Original Id: corda_h2020::999579084 Provenance: CORDA - COmmon Research DAta Warehouse - Horizon 2020	× ? 、
CONSORTIUM GARR PID (PIC): 999579084 URL: http://www.garr.it legal person non profit	GESTIONE AMPLIAMENTO RETE RICERCA	т	Original Id: cordahe::999579084 Provenance: CORDA - COmmon Research DAta Warehouse - Horizon Europe	× ? ×
Consortium GARR PID (GRID): grid.423642.5	Consortium GARR	TI 🚺	Original ld: ror::https://ror.org/03x9xd924	× ? .

×

An org with wrong suggested duplicates

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

Duplicates new

Metadata Management

Conflicts Note History ⁵

Current organization	
Name	Istituto Ortopedico Galeazzi
Туре	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	TI 🔽	Original Id: corda::999445709 Provenance: CORDA - COmmon Research DAta Warehouse	× ? ~
ISTITUTO ORTOPEDICO GALEAZZI PID (PIC): PIC:999554252 URL: (legal person) enterprise		TI	Original Id: corda::999554252 Provenance: CORDA - COmmon Research DAta Warehouse	× ? ✓
ISTITUTO ORTOPEDICO RIZZOLI URL: http://www.ior.it	IOR	т 👔	Original Id: corda_h2020::999445709 Provenance: CORDA - COmmon Research DAta Warehouse - Horizon 2020	× 2 ×

Manual entry of duplicates

Single org page -> Duplicates tab -> Add button (bottom of the page) -> type the search

c) OpenOrgs Database	e Search -	Curation -	Create 🕶				¢+ &+ 0+
Cr	niversity of Pisa openorgs::0000097669 eat8th at July 16, 2020 1146-19 odified at October 8, 2020 15:0							
	Metadata Management	Duplicates new	Conflicts	Note	History 🔊			
	Current organization							
	Name				University of Pisa			
	Туре				Education			
	Place				Pisa, IT			
	Acronyms				UniPi			
	Also known as				Università di Pisa University of Pisa Université de Pise Universität Pisa			
	Urls				http://www.unipi.it/			
	Other identifiers				0000 0004 1757 3729 (ISNI) grid.5395.a (GRID) 453158 (OrgRef) 50110007514 (FundRef) Q645663 (Wikidata)			
	Duplicates							
	Related organization				Acronym	Country	Source	
- (C) -	Université de Pise URL:					UNKNOWN	Original Id: anr::19c7778f9dbf4e28f0e9a8d520e7307b Provenance: Agence Nationale de la Recherche	× ? ✓

\rightarrow C

→ C 🌲	orgs.openaire.eu/main#!/edit/0/openorgs::0000097669		仓 😒 📑 🚳 🗈	o C 🌶	' 🛪 🛛	3	:
PID (NIF): NIF:NU URL:		Provendnice, Pr	or - nunuação para a ciencia e a rechologia	i.			
Università di Pl URL:	Select Organization			×	× ?	~	
Universit�� c URL:	Pisa		Page: 1/1 - Total	Search elements: 7	× ?	~	
UNIVERSITY O		« Previous Next »			× ?	<u>~</u>	
Università degl URL: https://www	Click the organization name to add the organization (multiple selection is allo	owed)			× ?	~	
11.1	Organization name	Place	Acronyms	Туре			
Università di Pi URL: http://www.u	Conservatorio di VENEZIA "Benedetto Marcello" - palazzo pisani URL: http://www.conservatoriovenezia.net/portale/index.php/it/	- , IT		UNKNOWN	× ?		
Università di Pl PID (OrgReg): ITC URL: http://www.u	IRCCS San Raffaele Pisana URL: http://www.sanraffaele.it/assistenza/strutture-sanitarie/1/irccs-san-raffaele-pisana	🚺 -, IT		UNKNOWN	× ?	~	
Università di Pi URL: https://www non profit	OPERA DELLA PRIMAZIALE PISANA URL: http://www.opapisa.it	💶 -, IT	OPERA DELLA PRIMAZIALE PISANA	UNKNOWN	× ?	~	
University of P PID (ISNI): 0000	PROVINCIA DI PISA	🚺 -, IT		UNKNOWN	× ?	<u>~</u>	
PID (GRID): grid. PID (ROR): https: PID (OrgRef): 45	Scuola Normale di Pisa VRL: https://api.eurocris.org/dris/orgunits/ou00332	— -, IT		UNKNOWN			
PID (FundRef): 5 PID (Wikidata): 0	UNIVERSITA DI PISA URL: http://www.unipi.it	-, IT	UNIPI	UNKNOWN			
URL: http://www.u University of P	UniverSity of Pisa, Department of Civilization and forms of knowledge URL: https://www.cfs.unipi.it/	-, IT		UNKNOWN	× 2		

URL:

Close

Adding new orgs

If an org is not present in OpenOrg

data curators can add it by clicking on create -> a new org.

There are three mandatory metadata: name, country and type.

Recommendations:

- use the English version of the organization name as the main name, and fill in the name in the national language in aliases
- Always try to include the URL of the organization or check that it is correct.
- use UTF-8 encoding

Tips:

- When you are not sure what type of organization it is, you can check the EC flags.
- You can add notes inside an org page to keep track of the choices made.



Approval of new orgs

Main concepts:

Pending or suggested org: is an organization to be approved by a data curator; the approval produces an OpenOrgs ID.

Approved org: is an organization that was confirmed by a curator. It is persisted with a stable identifier (OpenOrgs ID) and its metadata fields can be enriched/curated;



← → C 🏻 orgs.openaire.eu/main#!/pendings/IT

CopenOrgs Database Search Curation - Create -

Suggested organizations 180244

Suggested Organizati

Current country:

Organizations with new duplicates (12953)

Organizations that presents duplicates suggested by the Dedup OpenAIRE Workflow

Potential conflicts 1040

Approved Organizations that need to be merged



← → C		🖞 🖈 📴 🛞 🖬 🙆 🧭 🥕
O OpenOrgs Database Search ▼ Curation ▼ Create ▼		¢~ 4
Suggested Organizations		
Current country: IT -		downloa
Pisa		
Organization name 🛩	Place	Acronyms
Conservatorio di VENEZIA "Benedetto Marcello" - palazzo pisani	📶 -, IT	
IRCCS San Raffaele Pisana	-, IT	
IRCCS SAN RAFFAELE ROMA SRL	-, IT	IRCCS SAN REAFFAELE PISANA
KUVERA SPA	[11 -, IT	
OPERA DELLA PRIMAZIALE PISANA	-, IT	OPERA DELLA PRIMAZIALE PISANA
PROVINCIA DI PISA	-, IT	
Scuola Normale di Pisa	💶 -, IT	
University of Pisa, Department of Civilization and forms of knowledge	🚺 -, IT	

In this case also a relation must be established!



Beware!

The same org must not have more than one OpenOrgs id

So please approve the org only after checking that the org does not already exist with the OpenOrg ID).

If the same org has more than one OpenOrgs IDs, this will result in a conflict.



Resolution of conflicts

A conflict is created when the same organization is approved more than once, thus giving to the same org different OpenOrgs IDs.

The section dedicated to conflicts is called 'potential conflicts' because they are suggested by the deduplication algorithm and the curator must confirm that they actually are the same organization.

If a curator finds that the same organization accidentally has two different OpenOrgs IDs (that is it has been improperly approved more than once), the curator can add a conflict as follows: select the main organization, then select the conflicts tab on the org page, click add, search for the name of the conflicting org, select it, and save.

Adding or resolving potential conflicts set as "hidden" the conflicting orgs and creates a new OpenOrgs ID.



Potential conflict resolution

Conflicts		
Current country:		
Filter		
Group 1		
#1 Azienda Ospedaliero-Universitaria Careggi	🛐 Florence, IT	He
#2 Azienda Ospedaliera Universitaria Senese	🚺 Siena, IT	He
#3 Meyer Children's Hospital	Florence, IT	He
#4 Azienda Ospedaliero Universitario Mater Domini	🔲 Catanzaro, IT	He
#5 Azienda Ospedaliero Universitaria Pizana add resolve manually mergt all all different	Pisa, IT	He

Recap - Main concepts

Approved org: is an organization that was confirmed by a curator. It is persisted with a stable identifier (OpenOrgs ID) and its metadata fields can be enriched/curated;

Pending or suggested org: is an organization to be approved by the national admin; the approval produces an OpenOrgs ID.

Duplicate: possible duplication of the same organization, to be resolved by the user. It is created between an approved org and a raw org (an external org with no OpenOrg ID, ie coming from Corda, Re3data, ROR...). Duplicates can be suggested by the dedup algorithm or added by the user inside an org page.

Raw org: the organization with the original registry id. It will never change its ID (those changing ID with the approval are the pending or suggested orgs), it identifies the org as it comes from the original source; it can be added as duplicate.

Hidden org: is an organization not shown in the OpenAIRE public portals. It is the result of a resolution of a conflict. A conflict can be created for example when the same organization is approved by mistake more than once inside OpenOrgs (thus giving different stable OpenOrgs IDs to the same organization). If the curator resolves this conflict by merging two organizations with different OpenOrgs IDs, a new org is created and the previous ones are set as hidden.



Suggested workflow

1. Resolution of duplicates suggested by the algorithm;

2. Metadata curation and, if necessary, manual adding of duplicates - please note that this is a different task than the previous one;

3. Resolution of conflicts (if any);

4. Approval of new orgs (only after checking that the org do not already exist with the OpenOrg ID)



Quick guide for institutional dashboard managers

List of actions

1. Metadata curation for your organization

- a. Main name in english
- b. Name in the National language in aliases
- c. Add or update the URL
- d. Add extra PIDs (e.g. ROR identifier)
- e. Set the parent/child relationships
- 2. Management of the duplicates
 - a. On the "duplicates" tab confirm or reject the suggestions
 - b. Add new duplicates (if needed)

How to join

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

	→ C	🔮 🚳 🖸 🙆 🗲 🌲 🌘 Update
Sign In Info	OpenOrgs Database	
Access Guide To use this service you have to perform the following steps:	Authorization request Hello 'gina,pavone@isti.cnr.it', you don't have a role yet To apply as data curator compile the form below, an administrator will authorize you as soon as possible.	
 Register on the OpenAIRE portal. Login using the OpenAIRE credentials Compile the Authorization Request Form An administrator will authorize you as soon as possible 	Select your countries Afghanistan Áland Islands Albania Afghanistan American Samoa Andora Angola Angulla Anterican Samoa Andora Angola Angulla Anterican Samoa Andora Angola Argentina Armenia Australia Australia Azerbaijan Bahmans Bahrain Bahrain Banjadesh Balza Belra Belnin Belgium Bilze Belnin Belvize Belnin Bonaire, Sint Eustatius and Saba Bohaire and Herzegovina Bolivian Group British Indian O Brandi Deumetine Tracili Buistine from O Buistine from O Buistine from O	Territory

Demo session

- How a curated org appears: University of Pisa Italy
- An organization with **wrong duplicates suggested by the algorithm**: Istituto Ortopedico Galeazzi
- How to **curate metadata and approve a new org** inside OpenOrg: a department of the University of Pisa
- Avoid conflicts: university of L'Aquila
- Resolution of potential conflicts: Italian section
- Manually add a conflict: Accademia di belle arti dell'Aquila

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 suggested orgs (orgs to be approved)
- Browse only duplicates
- Insert a new org

Thank You! gina.pavone@isti.cnr.it claudio.atzori@isti.cnr.it

