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

Ivana Končić, Bojan Macan
Centre for Scientific Information
Ruđer Bošković Institute - HR

OpenOrgs: bridging registries of research organizations

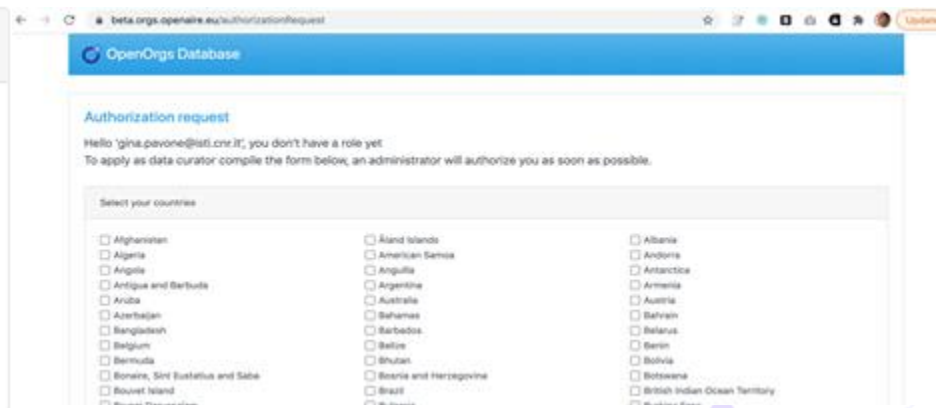## Improving the quality of data in OpenAIRE

OpenAIRE - 19 April 2023

OpenAIRE | NEXUS

@openaire_eu

How to join OpenOrgs

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**

The trainers

Gina Pavone

Research Fellow at the
Institute of Information
Science and Technologies
-

National Research Council
(CNR) - Italy

Ivana Končić

Graduate Librarian at the
Centre for Scientific
Information

Ruđer Bošković Institute
(RBI)

Croatia

Bojan Macan

Head of the Centre for
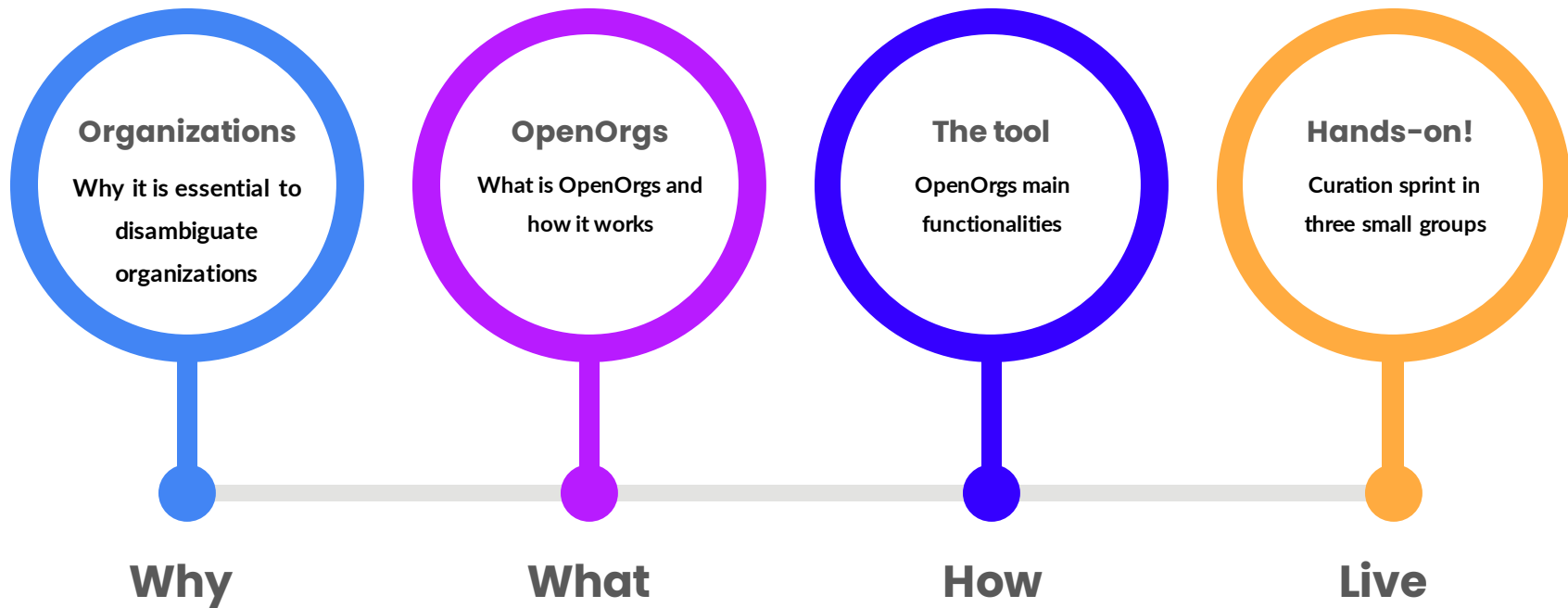Scientific Information

Ruđer Bošković Institute
(RBI)

Croatia

Today's agenda

OpenOrgs curation sprint - April 19

Organizations

Why it is essential to disambiguate organizations

OpenOrgs

What is OpenOrgs and how it works

The tool

OpenOrgs main functionalities

Hands-on!

Curation sprint in three small groups

Why

What

How

Live

Organizations - Information ambiguity
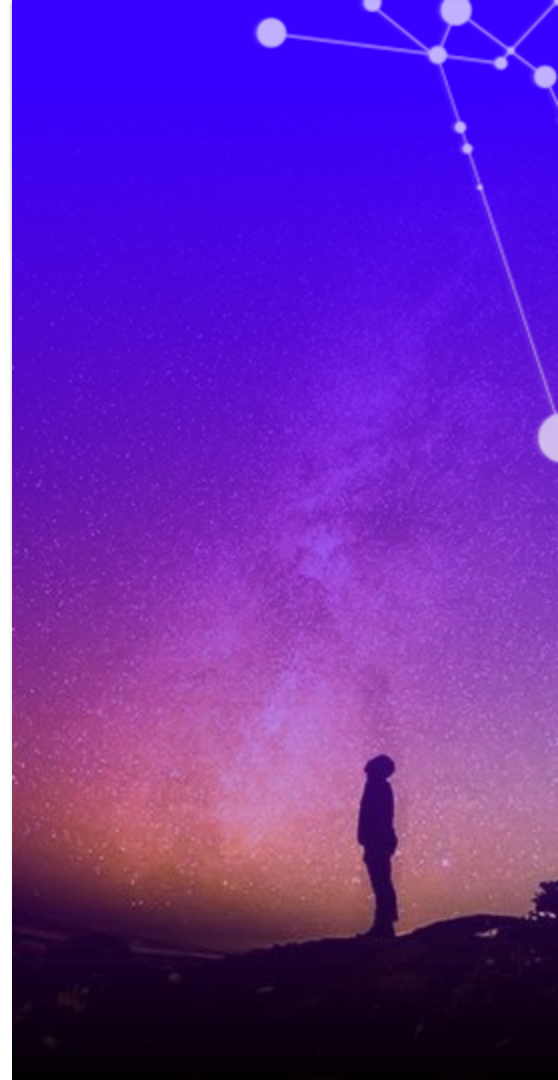
Information about the organizations is distributed in numerous data sources

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

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

1 Organization

Many different names

Many data sources

This ambiguity affects all the scholarly communication system

Services needing unambiguous information

Search & Discovery

Aggregation & showcase

Monitoring services

Open Science services

Researchers

RPOs

RFOs

All stakeholders

To facilitate the findability of digital objects

To consistently showcase the overall scientific production

To have consistent information on the impact of resources
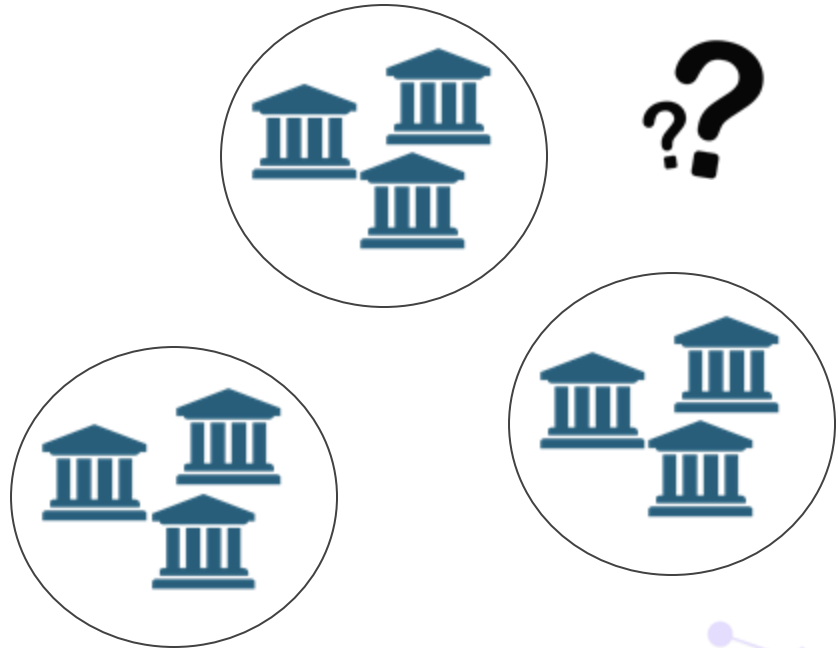
To have functional and up-to-date services

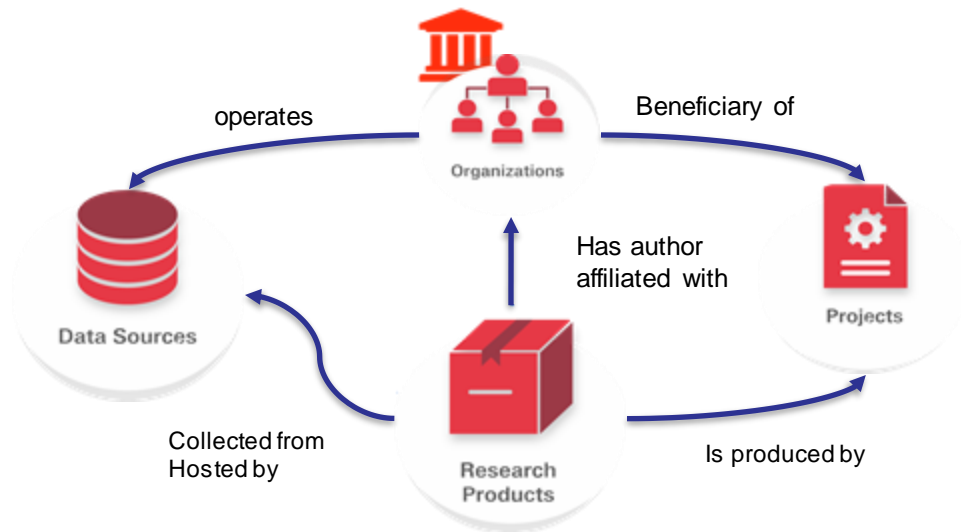OpenAIRE | NEXUS

Problem statement
disambiguation

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!

Problem statement
quality / completeness

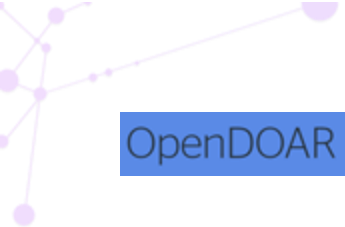| organization field completeness		
field	**missing**	**% missing**
legalname	254	0.1%
legalshortname	156,923	43.2%
alternativenames	319,825	88.0%
country	82,002	22.6%
website url	142,047	39.1%
pid	184,536	50.8%

Limitations of a fully automated approach

OpenDOAR

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

European Commission
CORDIS
EU research results

GRID

ROR

FUNDER   FUNDER   FUNDER

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 | NEXUS

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
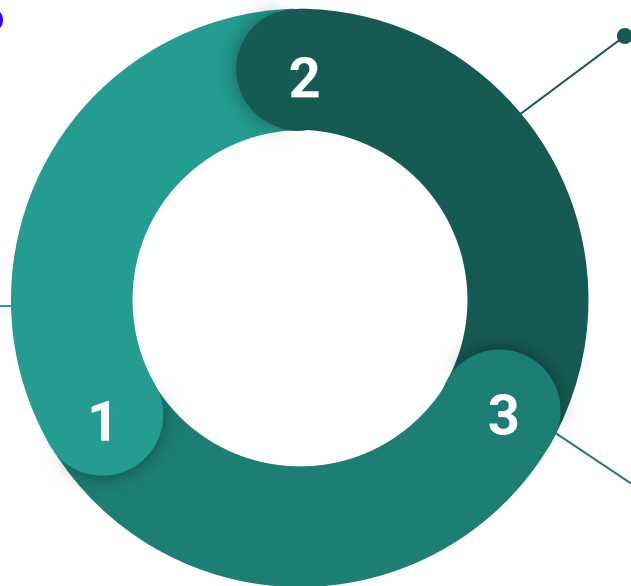
OS observatory

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

OpenAIRE | NEXUS

OpenOrgs activity pillars

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.

Automatic suggestion of duplicates

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

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

Data curators can

- Curate and enrich orgs <u>metadata</u>
- Resolve duplicates suggested by the algorithm in the <u>duplicates section</u>
- Manuall<u>y add duplicates</u> in the single org metadata page
- Add parent-child relations (if any) in the 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>

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

Suggested organizations 168624
Organizations that have been suggested by an User or by the Dedup OpenAIRE Workflow

Organizations with new duplicates 12812
Organizations that presents duplicates suggested by the Dedup OpenAIRE Workflow

Potential conflicts 1024
Approved Organizations that need to be merged

Search

Metadata curation & enrichment

Other names and identifiers

Acronyms	
IRB	🗑
RBI	🗑
new acronym...	+

Aliases		
name	language	
Institut Ruđer Bošković	hr	🗑
Rudjer Boskovic Institute	en	🗑
new alias...	language... ⇅	+

Identifiers		
id	type	
4763329	OrgRef	🗑
grid.4905.8	GRID	🗑
143735567	VIAF	🗑
Q7383690	Wikidata	🗑
501100007259	FundRef	🗑
999875031	PIC	🗑
0000 0004 0635 7705	ISNI	🗑
new id...	type... ⇅	+

Urls	
http://www.irb.hr/	🗑
http://...	+

How it looks on OpenAIRE EXPLORE

Metadata curation & enrichment

Recommendations and Tips

- Use the **English** organization name as the main name, and include the local name as an alias.
- Include the organization **URL** and use **UTF-8** encoding.
- Establish a consistent "local" naming **rules** (e.g. University of Zagreb, Faculty of Law).

- Check **EC flags** for organization type.
- Add **notes** to org page for information tracking.
- See in **history** the actions performed in the org page.

Connecting different IDs of the same org
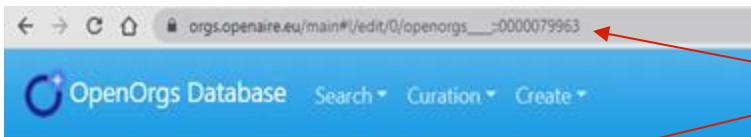
OpenOrgs ID

Where to find org's IDs?
- orgs official websites, documents, etc.
- other databases
  - **ISNI** https://isni.org/page/search-database/
  - **ROR** https://ror.org/
  - **PIC** https://bit.ly/41kE71Y
  - **VIAF** https://viaf.org/

CNRS
DGEEC
FundRef
GRID
HESA
ISNI
LinkedIn
NIF
OrgRef
OrgReg
Participant Identification Code
Ringgold
RingGold
RNSR
ROR
RRID
UCAS
UKPRN
VIAF
Wikidata
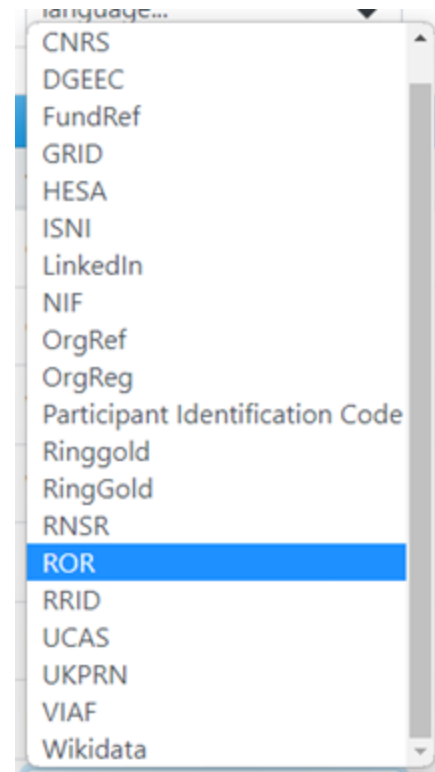
Metadata curation & enrichment

Establishing relationships between organizations

- OpenOrgs only permits parent-child relationships ('**is parent of**' or '**is child of**').
- Child production is **included** in parent production.
- Establishing relationships between organizations is important for **monitoring production**, particularly between universities, departments, and institutes.
- To link departments or institutes within an organization, they must first be approved as an organization and then linked as a child to the main org.
- Consider naming children orgs with a specific rule (e.g. *University of Zagreb, Faculty of Law*).
- One org can have **more than one parent** (for "hybrid" cases).

An example of relations:

Relations

Relations
This organizazion **is parent of** University of Split, Arts Academy
This organizazion **is parent of** University of Split, Catholic Faculty of Theology
This organizazion **is parent of** University of Split, Faculty of Chemistry and Technology
This organizazion **is parent of** University of Split, Faculty of Civil Engineering, Architecture and Geodesy
This organizazion **is parent of** University of Split, Faculty of Economics, Business and Tourism
This organizazion **is parent of** University of Split, Faculty of Electrical Engineering, Mechanical Engineering an
This organizazion **is parent of** University of Split, Faculty of Humanities and Social Sciences
This organizazion **is parent of** University of Split, Faculty of Kinesiology
This organizazion **is parent of** University of Split, Faculty of Maritime studies
This organizazion **is parent of** University of Split, Faculty of Science
This organizazion **is parent of** University of Split, School of Medicine

This organization | rel type... ⇕ | related organization...
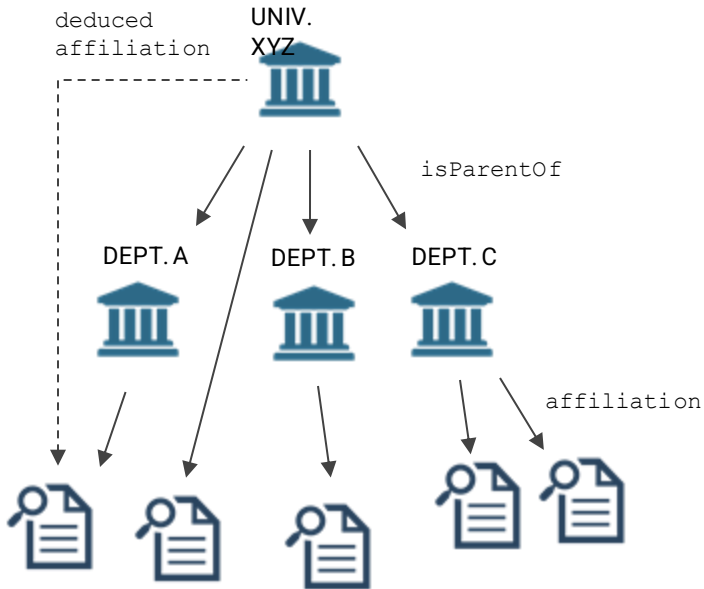
Adding new orgs

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:
- For the org's name, it is recommended to use the English version as the main name and include the name in the national language as an alias.
- Always try to include the organization's URL or check that it is correct and use UTF-8 encoding.

Tips:
- To determine the organization's type, check the EC flags.
- Notes can be added within an orgs page to keep track of choices made.

OpenAIRE | NEXUS

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;

OpenOrgs Database   Search ▾   Curation ▾   Create ▾

Suggested organizations 168624
Organizations that have been suggested by an User or by the Dedup OpenAIRE Workflow

Organizations with new duplicates 12812
Organizations that presents duplicates suggested by the Dedup OpenAIRE Workflow

Potential conflicts 1024
Approved Organizations that need to be merged

OpenAIRE | NEXUS

In this case also a relation must be established!

Beware!

The same org must not have more than one OpenOrgs ID!

- Only approve the organization after checking that it does **not already exist** with the OpenOrg ID.
- Multiple OpenOrg IDs for the same organization will result in a **conflict**.

OpenAIRE | NEXUS

Duplicates curation

Main concept: duplicates

- Duplicates can occur between **approved** and **raw organizations**.
- Raw organizations are external organizations without an OpenOrg ID, sourced from platforms such as Corda, Re3data, and ROR.
- The deduplication **algorithm may suggest** duplicates, or **users may add** them manually within an organization's page.

Duplicates curation in two steps

1. Complete the automated part

Duplicates section → lists organizations suggested by an algorithm to be duplicates.

It is an automated process that requires confirmation or rejection by the data curator.

2. Manual entry

Duplicates addition in the metadata org page → The data curator manually adds duplicates in the main org page.

Hand-added duplicates do not appear in the list of possible duplicates suggested by the algorithm.
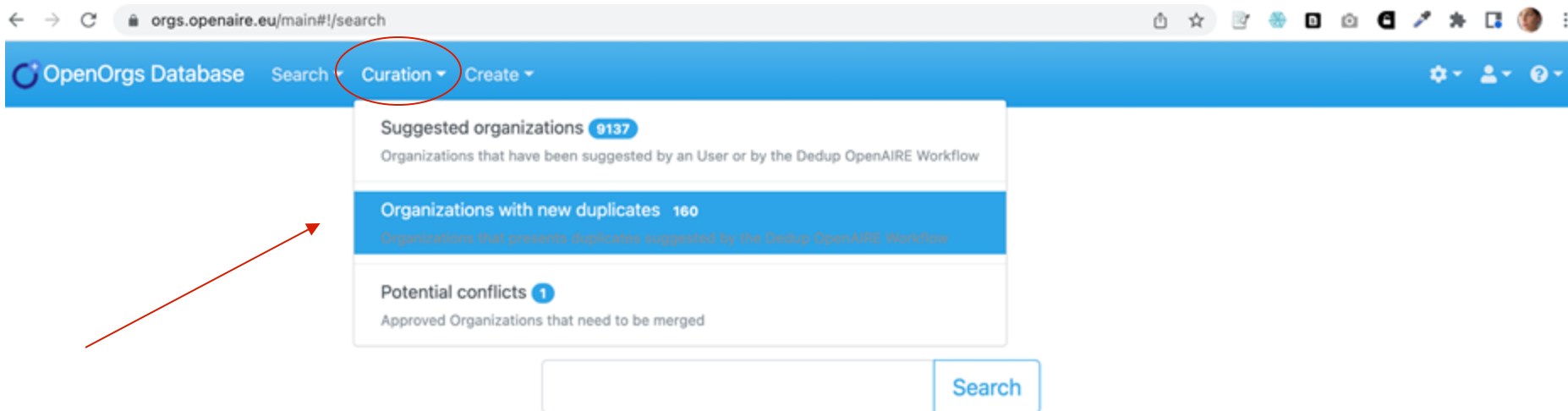
OpenAIRE | NEXUS

Potential new duplicates section

Curation menu → Organizations with new duplicates

University of Zadar

ID: openorgs____-0000016041
Created at July 16, 2020 11:48:19 by import:grid.ac
Modified at July 16, 2020 11:48:19 by import:grid.ac
OA Graph Node ID: openorgs____:5e91b35f34a6d4ead6432ef2f77d24bf [try on OA Explore]

| Metadata Management | Duplicates | Conflicts | Note | History 🕒 |

Current organization

Name	University of Zadar
Type	Education
Place	Zadar, HR
Acronyms	
Also known as	Sveučilište u Zadru
	Sveučilište u Zadru
	University of Zadar
	Universitas Studiorum Iadertina
	Свеучилиште у Задру
Urls	http://www.unizd.hr/
Other identifiers	3070362 (OrgRef)
	grid.424739.f (GRID)
	Q2496233 (Wikidata)
	0000 0001 2159 1688 (ISNI)
	https://ror.org/00t89vb53 (ROR)

Duplicates

Related organization	Acronym	Country	Source	
University of Zadar PID (PIC): 991684060 URL: http://www.unizd.hr legal body legal person non profit research organization higher education	UNIZD	HR	Original Id: corda_____::991684060 OA Graph Node ID: corda_____::4cec14029b16076cc6534f34d5e52d0c [try] Provenance: CORDA - COmmon Research DAta Warehouse	✗ ? ✓
SVEUCILISTE U ZADRU PID (PIC): 991684060 URL: http://www.unizd.hr legal body legal person non profit research organization higher education	UNIZD	HR	Original Id: corda__h2020::991684060 OA Graph Node ID: corda__h2020::4cec14029b16076cc6534f34d5e52d0c [try] Provenance: CORDA - COmmon Research DAta Warehouse - Horizon 2020	✗ ? ✓
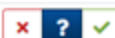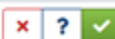
An org with wrong suggested duplicates

University of Zagreb, Faculty of Agriculture

ID: openorgs____:0000101403
Created at October 6, 2022 11:30:46 **by** Martina.Zugaj@irb.hr
Modified at October 6, 2022 11:30:46 **by** Martina.Zugaj@irb.hr
OA Graph Node ID: openorgs____:c55be4e1c4433a491bcce8e71df78ed4 [try on OA Explore]

Metadata Management	Duplicates	Conflicts	Note	History ⟲

Current organization

Name	University of Zagreb, Faculty of Agriculture
Type	Education `legal body` `legal person` `non profit` `higher education`
Place	Zagreb, HR
Acronyms	
Also known as	Sveučilište u Zagrebu, Agronomski fakultet
Urls	http://www.agr.unizg.hr/
Other identifiers	999534561 **(PIC)**

Duplicates

Related organization	Acronym	Country	Source	
SVEUCILISTE U ZAGREBU - GEODETSKI FAKULTET **PID (PIC):** 973026207 **URL:** http://www.geof.hr `legal body` `legal person` `non profit` `research organization` `higher education`	UNIVERSITY OF ZAGREB-FACULTY OF GEODESY	HR	**Original Id:** corda_____::973026207 **OA Graph Node ID:** corda_____::84dde38fa21b40fad483aa5267d51239 [try] **Provenance:** CORDA - COmmon Research DAta Warehouse **Added by:** Martina.Zugaj@irb.hr	✖ ? ✔

Manual entry of duplicates

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

Select Organization

university of zagreb | Search

Page: 1 / 1 - **Total elements:** 20

« Previous | Next »

Click the **organization name** to add the organization (multiple selection is allowed)

Organization name	Place	Acronyms	Type
Centre for Croatian Studies of University of Zagreb	-, HR		UNKNOWN
Criatian Social Science Data Archive, University of Zagreb Faculty of Humanities and Social Sciences URL: https://crossda.hr	-, HR	Criatian Social Science Data Archive, University of Zagreb Faculty of Humanities and Social Sciences	UNKNOWN
Croatian Sociological Society, Department of Sociology at Faculty of Humanities and Social Sciences, University of Zagreb	-, HR	Croatian Sociological Society, Department of Sociology at Faculty of Humanities and Social Sciences, University of Zagreb	UNKNOWN
Department of Chemistry Faculty of Science University of Zagreb	-, HR		UNKNOWN
Department of Geography Faculty of Science University of Zagreb	-, HR		UNKNOWN
Department of Mineralogy, Petrology and Geochemistry University of Zagreb	-, HR		UNKNOWN
Dept. of Comparative Politics Faculty of Political Science University of Zagreb	-, UNKNOWN		UNKNOWN
Faculty of electrical engineering and comput Department of Control and Computer Engineeri University of Zagreb	-, UNKNOWN		UNKNOWN
Faculty of electrical engineering + computer Department of Control + Computer Engineering University of Zagreb	-, HR		UNKNOWN
Faculty of Science Department of Geology&Paleontology University of Zagreb	-, HR		UNKNOWN
Faculty of Science Department of Geology&Paleontology University of Zagreb	-, UNKNOWN		UNKNOWN
National and University Library in Zagreb URL: https://www.nsk.hr/	-, HR	NSK	UNKNOWN
National and University Library in Zagreb URL: http://www.nsk.hr/	-, HR		UNKNOWN
SVEUCILISTE U ZAGREBU, UCITELJSKI FAKULTET URL: http://www.ufzg.unizg.hr	-, HR	UNIVERSITY OF ZAGREB, FACULTY OF TEACHER EDUCATION	UNKNOWN
University Hospital Center - Pathobiochemical and Clinical Research Unit, Zagreb	-, HR		UNKNOWN
University of Zagreb	-, UNKNOWN		UNKNOWN

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 deduplication algorithm suggests **potential conflicts**, which require confirmation from the curator.
- To add a conflict, curators can select the main organization and add the conflicting org through the **conflicts tab** on the org page.
- Resolving potential conflicts hides the conflicting orgs ('**hidden**') and creates a new OpenOrg ID.

OpenAIRE | NEXUS

Potential conflict resolution

Resolve conflicts

Organizations

openorgs___:0000043562

Name	Iscte - Instituto Universitário de Lisboa
Type	Education
Place	Lisbon, PT
Acronyms	Iscte
Also known as	Instituto Universitário de Lisboa Lisbon University Institute Higher Institute of Business and Labour Sciences Instituto Superior de Ciências do Trabalho e da Empresa
Urls	https://www.iscte-iul.pt/
Other identifiers	0000 0001 2220 8863 **(ISNI)** grid.45349.3f **(GRID)** https://ror.org/014837179 **(ROR)** 21625121 **(OrgRef)** Q2886844 **(Wikidata)**

select >>

openorgs___:0000098645

Name	ISCTE - Instituto Universitário de Lisboa
Type	Education
Place	-, PT
Acronyms	ISCTE
Also known as	
Urls	http://www.iscte-iul.pt
Other identifiers	PT0014 **(OrgReg)**

select >>

New group

Main organization

Name	Iscte - Instituto Universitário de Lisboa
Type	Education
Place	Lisbon, PT
Acronyms	Iscte
Also known as	Instituto Universitário de Lisboa Lisbon University Institute Higher Institute of Business and Labour Sciences Instituto Superior de Ciências do Trabalho e da Empresa
Urls	https://www.iscte-iul.pt/
Other identifiers	0000 0001 2220 8863 **(ISNI)** grid.45349.3f **(GRID)** https://ror.org/014837179 **(ROR)** 21625121 **(OrgRef)** Q2886844 **(Wikidata)**

Secondary organization 1

Name	ISCTE - Instituto Universitário de Lisboa
Type	Education
Place	-, PT
Acronyms	ISCTE
Also known as	
Urls	http://www.iscte-iul.pt
Other identifiers	PT0014 **(OrgReg)**

reset   all different   merge

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.

OpenAIRE | NEXUS

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)

Let's start!

1. Choose an organization to curate (e.g. your organization)

1. Check and edit (eventually enrich) the metadata - and in this point maybe it is worth recalling the metadata curation guidelines
    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)

1. Set parent/child relations (if any)

1. Go to the duplicate section and approve/reject duplicates

1. Manually add duplicates

1. Are there any conflicts? If yes, resolve them.

Are you up for a challenge?

We are challenging you to curate 10 organizations in 15 days!

Take on the challenge by choosing organizations to curate, checking and editing metadata including resolving duplicates, adding extra PIDs, setting parent/child relations...

Once you complete the challenge, tell us what you think!

Share your thoughts and provide feedback here:
https://bit.ly/3osnWAU

OpenAIRE | NEXUS

Demo session

- How **a curated org** appears: University of Pisa - Italy

- An organization with **wrong duplicates suggested by the algorithm**: Alma mater studiorum

- How to **curate metadata and approve a new org** inside OpenOrg:
  Department of Biology, University of Rome Tor Vergata, Rome, Italy

- **Avoid conflicts**: Aristotle University of Thessaloniki or university of L'Aquila

- Resolution of potential conflicts**:** Italian section
- Manually add a conflict: A2A

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
ikoncic@irb.hr
Bojan.Macan@irb.hr