



ISTI Technical Reports

OpenOrgs: a tool for the disambiguation of organizations

Michele Artini, ISTI-CNR, Pisa, Italy

Sandro Fabrizio La Bruzzo, ISTI CNR, Pisa, Italy

Michele De Bonis, ISTI-CNR, Pisa, Italy

Gina Pavone, ISTI-CNR, Pisa, Italy



OpenOrgs: a tool for the disambiguation of organizations

Artini M., La Bruzzo S.F., De Bonis M., Pavone G.

ISTI-TR-2022/034

Organizations appear all over the Research & Innovation ecosystem in different shapes and formats: the same organization may appear with different metadata fields, different names - e.g., full legal name, short or alternative names, acronym. The ambiguity of organizations results in a huge deficiency in the exchange of information, the findability of research products, the monitoring of activities, and ultimately building a linked open scholarly communication system. OpenOrgs combines an automated process and human curation to compensate for the lack of information available and improve the organization's discoverability.

Keywords: Deduplication, Curation, Harmonization, Organization, OpenAIRE.

Citation

Artini M., La Bruzzo S.F., De Bonis M., Pavone G. *OpenOrgs: a tool for the disambiguation of organizations*. ISTI Technical Reports 2022/034. DOI: 10.32079/ISTI-TR-2022/034.

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

Area della Ricerca CNR di Pisa

Via G. Moruzzi 1

56124 Pisa Italy

<http://www.isti.cnr.it>

OpenOrgs

A tool for the disambiguation of organizations

*Michele Artini (ISTI CNR), Sandro Fabrizio La Bruzzo (ISTI CNR), Michele De Bonis (ISTI CNR),
Gina Pavone (ISTI CNR)*

The problem

Organisations appear all over the Research & Innovation ecosystem in different shapes and formats: the same organization may appear with different metadata fields, different names - e.g., full legal name, short or alternative names, acronyms. Even persistent identifiers may be of no help when different data sources identify organisations according to different PID schemas (ROR, ISNI, EC Pic numbers, and so on).

The ambiguity of organisations results in a huge deficiency in the exchange of information, the find-ability of research products, the monitoring of activities, and ultimately building a linked open scholarly communication system.

OpenOrgs and OpenAIRE

OpenAIRE is a technical infrastructure harvesting research output from connected data providers. OpenAIRE aims to establish an open and sustainable scholarly communication infrastructure responsible for the overall management, analysis, manipulation, provision, monitoring and cross-linking of all research outcomes.

OpenOrgs has been developed as a service of the OpenAIRE infrastructures and it is available at the address: <https://orgs.openaire.eu>.

It works on the organizations, aggregated by the OpenAIRE infrastructure, to disambiguate them and to improve their quality.

How it works

OpenOrgs works with a combination of an automated process and human curation. An algorithm does the first part of the work, grouping organizations with a certain degree of similarity in their metadata. After that, the curator has to determine if the identity is real or not. In most cases, it will be enough to look at the overall metadata collected from several sources to accept or reject the identity. But in some cases - and this is why it is impossible to rely only on the automated process - the user will need some additional research (or knowledge of the country being curated) to understand, for instance, whether two org entities are branches of the same organization or two independent institutions. Besides, the user can also suggest new duplicates that the algorithm has not found, curate and enrich metadata.

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;

Suggested org: is an organization to be approved by the curators; the approval produces an OpenOrgs ID.

Raw org: the organization with the original registry id (CORDA, re3data, ROR, ...). It will never change its ID, it identifies the org as it comes from the original source; it can be added as duplicate.

Duplicate: possible duplication of the same organization, to be resolved by a curator. It is created between an approved org and a raw org. Duplicates can be suggested by the dedup algorithm or added by the users.

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

Hidden org: is an organization not shown in the OpenAIRE public portals. It is the result of a resolution of a conflict. If the curator resolves this conflict by merging two organisations with different OpenOrgs IDs, a new org is created and the previous ones are set as hidden.

Curators

To become a curator the users have to:

- register themselves on the OpenAIRE portal
- request the authorisation to OpenOrgs choosing the countries that they want to curate
- wait for the approval

To each curator will be assigned one of the following roles:

- **Simple user:** he can perform basic operations on a subset of country,
- **National Admin:** he can perform all the operations but only on a subset of country, he can also approve other users (limited by country),
- **Super Admin:** he can operate on all the countries and all the users, he can also perform administrative tasks and configure the tool.

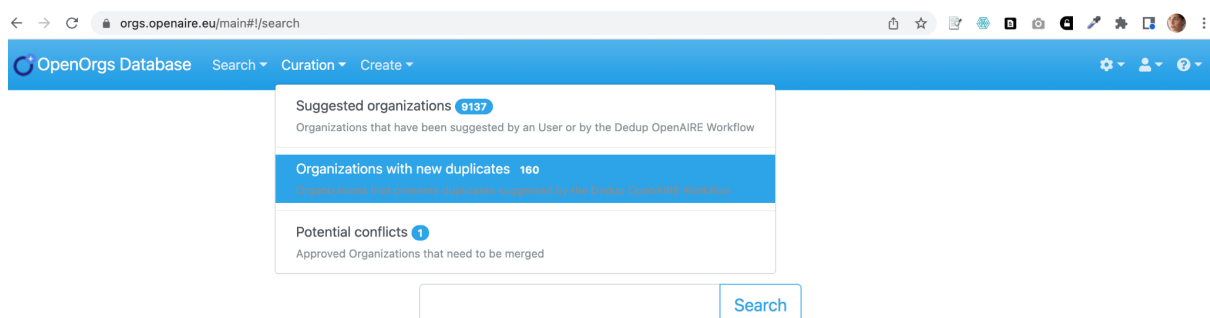
	Simple User	National Admin	Super Admin
Available countries	a limited set of countries	a limited set of countries	ALL
Metadata curation	x	x	x
New organizations	only suggestion	x	x
Duplicate resolution	x	x	x
Conflict resolution		x	x
Show Org History		x	x
Users configuration		limited	x
System configuration			x

Architecture

The tool has been developed as a SpringBoot Application with data stored in a relational database, the actual version has been deployed using the following component/framework:

- Java 11
- PostgreSQL 13
- SpringBoot 2.5.2
- Bootstrap 4.3.1 (for the UI)
- AngularJS 1.8.2 (for the UI)

Page: Main Page



The main page contains a form that permits to search the organizations.

For simple users and national administrators the results are limited by their countries.

Under the “Curation” menu the curators can find the a summary of the operations suggested by the automatic algorithm

Page: Metadata Editor

The screenshot shows the OpenOrgs Database interface for editing the metadata of the University of Pisa. The top navigation bar includes 'OpenOrgs Database', 'Search', 'Curation', and 'Create'. The main content area is titled 'University of Pisa' and displays the following information:

- ID:** openorgs____:0000097669
- Created at:** July 16, 2020 11:48:19 by import:grid.ac
- Modified at:** July 5, 2022 12:21:15 by noad-it@openaire.eu
- OA Graph Node ID:** openorgs____:5c351d85f02db01ca291acd119f0bd78 [try on OA Explore]

The interface is divided into several sections for metadata management:

- Official name and type:** A form with 'name' set to 'University of Pisa' and 'type' set to 'Education'. There is an 'EC flags' button.
- Geographical location:** A form with 'city' set to 'Pisa', 'country' set to 'Italy', 'lat' set to '43.716429', and 'lng' set to '10.398687'.
- Other names and identifiers:** This section contains three sub-tables:
 - Acronyms:** A table with one entry 'UniPi' and a trash icon. Below it is a text input field 'new acronym...' with a plus icon.
 - Aliases:** A table with columns 'name' and 'language'. It lists four entries: 'Università di Pisa' (it), 'University of Pisa' (en), 'Université de Pise' (fr), and 'Universität Pisa' (de). Each entry has a trash icon. Below the table is a text input field 'new alias...' and a language dropdown menu with a plus icon.
 - Identifiers:** A table with columns 'id' and 'type'. It lists one entry: '0000 0004 1757 3729' (ISNI) with a trash icon.



The curators are invited to respect some recommendations during the editing of the metadata, for example:

- prefer the English version of the organization name as the main name, the names in other languages can be added as “Other names”



OpenOrgs: A tool for the disambiguation of organizations

- include the URL of the organization and/or check that it is correct.
- use UTF-8 encoding
- the tool permits to model the parent/child relationships between the organisations, such as between universities and departments or between the main organisation and its institutes. These relationships are very important for the OpenAIRE production monitoring.








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

Q645663	Wikidata	
<input type="text" value="new id..."/>	<input type="text" value="type..."/>	

Uris

http://www.unipi.it/	
<input type="text" value="http://..."/>	

Relations

Relations	
This organization is parent of Department of Chemistry and Industrial Chemistry - Università di Pisa	
This organization is parent of Dipartimento di Civiltà e Forme del Sapere - Università di Pisa	
This organization is parent of Dipartimento di Civiltà e Forme del Sapere - Università di Pisa	
This organization is parent of Dipartimento di Civiltà e Forme del Sapere - Università di Pisa	
This organization is parent of Dipartimento di scienze economiche - Università di Pisa	
This organization is parent of Dipartimento di Studi Italianistici - Università di Pisa	
This organization <input type="text" value="rel type..."/>	<input type="text" value="related organization..."/> 

Page: Suggested Organizations

Current country: IT download as CSV

Pisa

Organization name	Place	Acronyms	Type
Conservatorio di VENEZIA "Benedetto Marcello" - palazzo pisani	-, IT		UNKNOWN
IRCCS San Raffaele Pisana	-, IT		UNKNOWN
IRCCS SAN RAFFAELE ROMA SRL	-, IT	IRCCS SAN RAFFAELE PISANA	UNKNOWN
KUVERA SPA	-, IT		UNKNOWN
OPERA DELLA PRIMAZIALE PISANA	-, IT	OPERA DELLA PRIMAZIALE PISANA	UNKNOWN
PROVINCIA DI PISA	-, IT		UNKNOWN
Scuola Normale di Pisa	-, IT		UNKNOWN
University of Pisa, Department of Civilization and forms of knowledge	-, IT		UNKNOWN

Only the administrator can see the Suggested Organizations and decide their approval.

These organisations can be suggested by automatic algorithms or by simple users.

When suggested organization is approved, the system assign a new OpenOrgs identifier to it.

Page: Duplicates curation

Consortium GARR



Registered organization	
Name	Consortium GARR
Type	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	IT	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	IT	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	IT	Original Id: corda_____he:999579084 Provenance: CORDA - COmmon Research DATA Warehouse - Horizon Europe
Consortium GARR PID (GRID): grid.423642.5	Consortium GARR	IT	Original Id: ror_____:https://ror.org/03x9xd924 -----

The organization contains 3 approved and 1 pending 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

Metadata Management			
Duplicates	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.iior.it legal body legal person non profit research organization	IOR	IT	Original Id: corda_____:999445789 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.iior.it	IOR	IT	Original Id: corda__h2020:999445789 Provenance: CORDA - COmmon Research DATA Warehouse - Horizon 2020

The organization contain 1 approved and 2 rejected duplicates

Page: Conflict resolution

The screenshot shows the OpenOrgs Database interface for conflict resolution in Italy (IT). The page title is "Conflicts". A dropdown menu shows "Current country: IT". Below this is a "Filter..." input field. A table titled "Group 1" lists five organizations in a conflict:

Group 1	
#1	Azienda Ospedaliero-Universitaria Careggi
#2	Azienda Ospedaliera Universitaria Senese
#3	Meyer Children's Hospital
#4	Azienda Ospedaliero Universitario Mater Domini
#5	Azienda Ospedaliero Universitaria Pisana

At the bottom of the table are four buttons: "add", "resolve manually", "merge all", and "all different".

The curator can resolve the conflict perform one of the following action:

- **resolve manually:** he can choose the master and decide to merge only a subject of the suggested orgs.
- **merge all:** the conflict is automatically resolved merging all the organizations
- **all different:** the conflict is totally rejected

The curators can also add other organizations to the conflict before its resolution.

Table Of Contents

The problem	1
OpenOrgs and OpenAIRE	1
How it works	2
Main Concepts	2
Curators	3
Architecture	4
Page: Main Page	4
Page: Metadata Editor	5
Page: Suggested Organizations	7
Page: Duplicates curation	8
Page: Conflict resolution	9
Table Of Contents	10