EPiC Series in Computing

Volume 109, 2026, Pages 172–183

Proceedings of EUNIS 2026 Annual Congress

Comparing Institutional Repository and OpenAIRE Research Products. Scuola Normale Superiore Case Study

Alessandro Cignoni[1], Daniele Marotta[1], and Donatella Tamagno[1]

Scuola Normale Superiore, Pisa, Italy
alessandro.cignoni@sns.it, daniele.marotta@sns.it, and donatella.tamagno@sns.it

Abstract

This paper presents an exploratory study conducted at the Scuola Normale Superiore (SNS) to evaluate the potential of OpenAIRE as a source for enriching institutional repository data. SNS uses IRIS as its Institutional Research Information System to collect and manage research outputs. In line with the Barcelona Declaration on Open Research Information, SNS initiated a collaboration with OpenAIRE to assess whether its Knowledge Graph and APIs can support the population of institutional repositories while reducing reliance on commercial platforms.

An open-source Python script was developed to retrieve and compare research products from IRIS and OpenAIRE. Data on the research outputs of SNS-affiliated researchers for the period 2021-2025 were extracted from IRIS. Other datasets from OpenAIRE were defined to collect the scientific outputs of SNS researchers, since OpenAIRE records organizational affiliations for research products but does not record researchers' affiliations. A record-matching model based on exact, sufficient, and weighted criteria was implemented to identify common and unique records across IRIS and OpenAIRE datasets.

Results show that OpenAIRE provides substantial coverage of SNS research outputs and could increase the number of records stored in the institutional repository. However, this study revealed some metadata inconsistencies, incomplete author identification, and affiliation ambiguities that currently limit the feasibility of fully automated bulk import.

Overall, OpenAIRE demonstrates potential as an open infrastructure for research information, but this study highlights the need for further improvements in metadata completeness, affiliation tracking, and deduplication processes that are necessary to enable reliable large-scale integration with institutional repositories.

1 Introduction

This exploratory study stems from the Scuola Normale Superiore (SNS)'s need to monitor the scientific outputs of its affiliated researchers.

SNS, like many Italian universities and research institutions, uses the Institutional Research Information System (IRIS) [7] as its institutional repository to collect, manage, and disseminate research outputs. IRIS is a proprietary software developed by Consorzio Interuniversitario del

Nord-Est per il Calcolo Automatico (CINECA) built on DSpace [9], an open-source system for scholarly repositories, and integrated with commercial platforms for harvesting of research outputs.

Following the SNS's adherence to the Barcelona Declaration [2], which promotes the use of open research information systems and infrastructures, SNS initiated a collaboration with OpenAIRE [15] to explore the potential offered by its APIs for retrieving and enriching data on SNS research outputs in its institutional repository IRIS.

OpenAIRE is a worldwide open scholarly communication infrastructure built on a knowledge graph that maps the research landscape through entities, including research products, organizations, projects, data sources, communities and persons, and relationships among them [12]. The data collected and processed by the OpenAIRE Graph are accessible via APIs through multiple entity-specific endpoints [13].

IRIS and OpenAIRE are already interconnected: IRIS retrieves metadata on European research projects from OpenAIRE, and OpenAIRE collects research product metadata from institutional repositories. However, IRIS's technical architecture does not support the harvesting of publication metadata from OpenAIRE, as it does from commercial platforms.

This work aims to assess the extent of coverage of SNS scientific production in OpenAIRE and the quality of its metadata, with the goal of progressively integrating OpenAIRE services into the institutional repository, reducing dependency on commercial platforms, enhancing data quality and transparency, and strengthening the role of open infrastructures in the research information ecosystem.

As an exploratory solution to address this limitation and ensure reusability by other universities, an open source Python script has been developed and uploaded to GitHub and Software Heritage [4] to retrieve OpenAIRE metadata on research products authored by researchers affiliated with a given organization.

This work addresses the following research questions:

1. To what extent can the OpenAIRE Graph support the population of institutional repositories? The case study of IRIS SNS.

2. How reliable are OpenAIRE affiliations when assessed against IRIS SNS data?

To answer these questions, various datasets were defined and collected from OpenAIRE and IRIS (§ 2). The technological choices were outlined (§ 3), and a model, along with a corresponding script for dataset comparison, was developed (§ 4). Based on this framework, the analysis of the SNS case study was conducted (§ 5), leading to relevant conclusions (§ 6).

2 Data Collection

In IRIS, authorship for every author is stored as plain text, but affiliated researchers are also represented through a relationship with the corresponding instance of the internal author entity [6]. To ensure accuracy in matching records between IRIS and OpenAIRE, only internal authors with an ORCID were considered, as OpenAIRE identifies people through their ORCID identifiers. Non-internal authors, whose names are not standardized in IRIS, were excluded from the analysis. The dataset of research products by internal authors exported from the institutional repository has been named *Products from Institutional Repository* ($PfIR$). This dataset can be constructed on the basis of other institutional repositories, provided that the data comply with the defined model (§ 4).

At present, the OpenAIRE Graph does not record authors' affiliations within the Persons entity; instead, for each research product it stores the institution to which the authors were affiliated at the time of publication [10]. Therefore, two different approaches are available to collect SNS–affiliated authors' products in OpenAIRE:

1. Retrieving all products with at least one author affiliated with SNS at the time of publication (affiliations are specified in the research products). The resulting dataset is called *Affiliated Products* (AP).

2. Retrieving all products by affiliated researchers, listed in the institutional repository at a given time (*Authors List*, AL). The resulting dataset is called *Products from Authors List* ($PfAL$).

The two datasets extracted from OpenAIRE do not overlap perfectly: AP represents the research outputs resulting from the research activity carried out within the institution (including retired researchers and those now working at other institutions), whereas $PfAL$ includes research products authored by currently affiliated researchers throughout their entire careers.

$AP \cup PfAL$ is called *OpenAIRE* (OA), that is the set-theoretic union of the two datasets, without duplicates.

$PfIR$ is defined in accordance with the institution open access policy. Regarding SNS, it includes the research outputs of the last five years from the beginning of the affiliation and the research outputs resulting from activities carried out at SNS [16].

OA is comparable to $PfIR$, with the difference that $PfIR$ collects the research products of the last five years from the beginning of the affiliation, whereas OA collects research products covering researchers' entire career.

In OpenAIRE, author information for each research product is extracted directly as it appears in the source. `fullname` field containing the exact string as it appears in the product is always available. However, an author's complete name and surname, as well as their ORCID identifier, may not always be present. For this reason, only those authors who had at least specified either both given name and surname or an ORCID were considered as authors of a product in this analysis.

3 Technological Choices

Since the script end users are expected to be administrative staff, installing and using the Python code should be as simple as possible. The Python script requires as inputs:

1. Start and end year of the period under analysis.

2. The institution's ROR id, used to query the Research Products endpoint and retrieve AP. Thanks to affiliation propagation in OpenAIRE, it is possible to get metadata even of research products affiliated to child institutions [11].

3. A `CSV` file containing AL exported from IRIS or other institutional repository, provided that the data comply with the defined model (§ 4).

4. A `CSV` file containing $PfIR$.

With the exception of `JSON` files used for debugging, the script's only outputs are `XLSX` files containing the comparison of a pair of datasets, generated using the `openpyxl` library [8]. Excel (`XLSX`) format was preferred over `CSV` in order to produce a single output file for comparison and to enable graph construction. The only other library required is `requests` to access OpenAIRE APIs and retrieve AP, $PfAL$ and OA [3].

4 Comparison Model

To compare IRIS and OpenAIRE records, a set of fields useful for identifying research products was selected (§ 4.1). The resulting datasets were normalized prior to the comparison (§ 4.2), that is based on a record-to-record matching process to evaluate the likelihood of each pair referring to the same product (§ 4.3). For each comparison, a dataset of research products in common and those unique to each input dataset was built, and a summary analysis using graphs was implemented (§ 4.4). The Python script follows the comparison model, applying it at two dataset at a time (§ 2).

4.1 Field Selection

IRIS and OpenAIRE use different data models, so not all metadata useful to identify a publication is present in both systems. Through the OpenAIRE APIs, ISBN of books is not available. Both IRIS and OpenAIRE define type of products but the classification is different. IRIS distinguishes between a book chapter and a journal article, storing the title of the book and that of the journal in separate fields, whereas OpenAIRE records everything in a single field. Finally, OpenAIRE distinguishes between the title of a work and its subtitle, whereas IRIS uses a single field. Considering these differences, a new data model has been defined that collects information from both datasets for each record (§ Table 1). It should be noted that $PfIR$ has a record for each publication-author pair, so a conversion is needed.

Field name	IRIS field	OpenAIRE field	Note
`id`	`HANDLE`	`id`	
`type`	`OWNING_COLLECTION_DES`	`type`	
`title`	`TITLE`	`mainTitle` + `subTitle`	
`lang`	`LAN_ISO_1`	`language[code]`	
`doi`	`IDE_DOI`	`pids`	With `"doi"` as `scheme`
`year`	`DATE_ISSUED_YEAR`	`publicationDate`	Only year
`publisher`	`PUB_NAME`	`publisher`	
`container_name`	`REL_ISPARTOFBOOK` or `REL_ISPARTOFJOURNAL`	`container[name]`	
`issn`	`REL_ISSN`	`container[issnPrinted]`, `container[issnOnline]`	List of ISSN
`vol_iss`	`REL_VOLUME` + `REL_ISSUE`	`container[vol]` + `container[iss]`	Field created with normalization

Table 1: Data model for comparison

Authors are not included in Table 1 because of the complexity of this metadata. In the comparison data model, authors are stored as a list of dictionaries, as shown in Table 2.

Field name	IRIS field	OpenAIRE field	Note
`name`	`FIRST_NAME`	`name`	
`surname`	`LAST_NAME`	`surname`	
`orcid`	`ORCID`	`authors[pid[id[value]]]`	With `"orcid"` as `scheme`

Table 2: Authors data model

4.2 Normalization

The comparison of each field was based on an exact match, in order to avoid the use of distance algorithms and likelihood management. Therefore, normalization is required. In Table 3, the normalization rule for each field in the comparison data model is listed.

Field name	Normalization rule
`title`	Small capitalization, removal of spaces and punctuation marks
`authors[name]`	Concatenation of `surname` and `name`, small capitalization, removal of spaces and punctuation marks
`publisher`	Small capitalization, removal of spaces and punctuation marks
`container_name`	Small capitalization, removal of spaces and punctuation marks
`vol_iss`	Concatenation of `volume` and `issue`, conversion of roman numbers to arabic digits and initial 0 digit removal

Table 3: Normalization rules

Normalized fields are used only in the comparison process; the output datasets contain the original IRIS and OpenAIRE metadata values. This improves the readability of the records and enables a manual verification of the script's results.

4.3 Matching Rules

Since AP and $PfAL$ originate from the same data source, the matching between them was performed using OpenAIRE identifiers.

For the other comparisons, the matching is obtained through a brute-force loop that perform a pairwise comparison between every record in the first dataset and every record in the second, in both directions. The result of this process consists of two datasets of matches, from which the common and unique products for each dataset are identified.

Field comparison is based on a strict equality check, except for `authors` and `issn`, since these metadata are not plain strings (Table 1). Authors from the IRIS database are conceptually different from those in OpenAIRE, as they include only internal authors rather than all authors to a product. Therefore, a match is obtained only if all authors from IRIS are contained in the OpenAIRE list, identified either by ORCID or by full name. Since IRIS assigns only a single ISSN to each research product, a match is obtained only if this ISSN is contained in the ISSN list provided by OpenAIRE.

When comparing a record with another there are three possible outcomes: exact match, sufficient match or weighted match.

Exact matches Records are considered to refer to the same research product if they share the same DOI, assuming they are correctly recorded in both databases.

Sufficient matches If there is no exact match, records are considered to refer to the same research product if they share one set of metadata (Table 4), that are deemed sufficient to identify a product.

Name	Set of fields	Type of products identified
suff1	`title`, `publisher`, `container_name`, `year`	Book chapters without DOI
suff2	`title`, `issn`, `vol_iss`	Papers in journals without DOI
suff3	`authors`, `title`, `year`	Products without DOI

Table 4: Sufficient sets of metadata to identify products without DOI

Weighted matches If there is no sufficient match, this rule can be useful to flag other pairs that might be a match when metadata do not allow automatic identification, thus requiring a manual check. A weight was assigned arbitrarily to each field (Table 5), bearing in mind that a weight of 1 should represent a high likelihood of a match and building on the sufficient matches.

Field name	Weight
`title`	0.5
`lang`	0.1
`year`	0.25
`authors`	0.25
`publisher`	0.25
`container_name`	0.25
`issn`	0.25
`vol_iss`	0.25

Table 5: Fields weights

Deduplication It may occur that an IRIS record obtains both an exact and a sufficient match or two different sufficient matches with different OpenAIRE records. These duplicated instances have been removed from the matches list and presented separately.

4.4 Automatic Summary Analysis

In the `XLSX` output file, there is a dedicated sheet for a summary analysis composed of the following tables and graphs:

- **Dataset composition table**
  Provided for every comparison, the table specifies the number of identified products that are common, unique, and total for each source dataset.

- **Dataset composition table (percentage)**
  Provided for every comparison, the table specifies the percentage over the total of identified products that are common and unique for each source dataset.

- **Dataset composition clustered column barchart**
  A clustered column barchart built upon the previous table.

- **Exact and sufficient matching table**
  Not provided for the comparison between AP and $PfAL$, the table reports the number of exact matches and the counts for each type of sufficient match.
- **Exact and sufficient matching piechart**
  A piechart built upon the previous table.
- **Weighted matches equal or greater than 0.75 confidence distribution table**
  Not provided for the comparison between AP and $PfAL$, the table reports the number of weighted matches with a confidence equal to or greater than 0.75, classifying them according to their confidence category.
- **Weighted matches equal or greater than 0.75 confidence distribution piechart**
  A piechart built upon the previous table.
- **Weighted matches equal or greater than 1 matched fields distribution table**
  Not provided for the comparison between AP and $PfAL$, the table reports the number of weighted matches with a confidence equal to or greater than 1, classifying them according to the matched fields (Table 5).
- **Weighted matches equal or greater than 1 matched fields distribution piechart**
  A piechart built upon the previous table.

5 SNS Case Study: Results Analysis

All authors affiliated with the SNS who have an ORCID, including administrative staff, were considered in constructing $PfAL$, and all products in IRIS were included in the construction of $PfIR$, regardless of the internal authors' roles.

Data from IRIS and OpenAIRE were extracted for a five-year period, from 2021 to 2025. A year-by-year extraction was preferred to reduce computing time and obtain more detailed information. The `year` metadata always matches, since records are compared only with others from the same year.

All the analyzed data were extracted from IRIS and OpenAIRE on 19$^{\text{th}}$ February 2026 and they are available on the Zenodo community of SNS [5].

5.1 Institutional Repository Enrichment Using OpenAIRE

As shown in Figure 1, OA and $PfIR$ are of comparable magnitude, even though IRIS contains fewer products, and the number of items in common resulting from the comparison corresponds to roughly half of the publications.

In principle, this suggests that all records in $OA_{uniques} = OA - (OA \cap PfIR)$ could be integrated into IRIS, implying a 65% increase in the number of products stored in IRIS.

However, the number of importable research products is actually smaller, due to mismatches that can only be identified by a human operator.

It is worth noting that the number of products for 2025 is considerably lower in both $PfIR$ and OA. Regarding IRIS, this can be attributed to reduced researcher incentive to upload their most recent outputs, as the latest Evaluation of Research Quality (*Valutazione della Qualità della Ricerca* - *VQR*) covered only the years 2020–2024 [1]. OpenAIRE also shows fewer items, likely due to harvesting latency: records must first be indexed by a source and only later harvested by OpenAIRE.

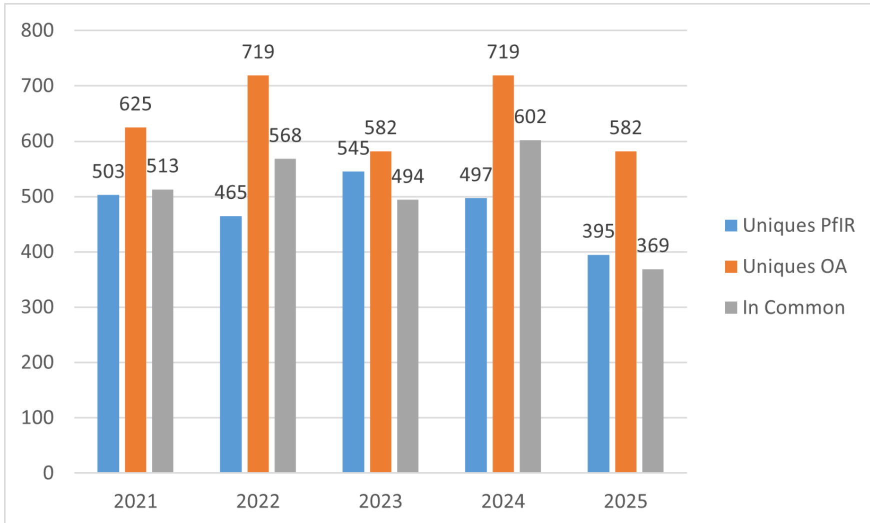

Figure 1: $PfIR$ and OA Composition per Year

As shown in Figure 2, it is possible to assess the effectiveness of the sufficient rules presented in Table 4. Compared with exact matching via `doi` field, these rules were less effective in identifying products: in particular, the rules designed to identify book chapters without DOI (suff1) and journal papers without DOI (suff2) produced only a handful of matches. This can be explained by two main factors:

1. the script's logic prioritizes DOI matching: if two records match via `doi`, the comparisons required for the sufficient rules are not executed, meaning that the rules could have been used to identify a larger number of products, but their application was not necessary;

2. with regard to book chapters without DOI (suff1), it is plausible that OpenAIRE has greater difficulty harvesting their metadata, as they form part of a larger, unified work.

The high number of matches obtained using the rule based on `authors`, `title`, and `year` (products without DOI - suff3) may partly be explained by the occurrence of false positives. The same authors may have published a contribution with the same title in different venues, or may have produced both a publication and, for example, its associated dataset or software. To achieve proper disambiguation, it would therefore be necessary to consider `issn` and `type` as well. However, considering `issn`, many correct matches would likely be lost due to the incompleteness of metadata in OpenAIRE's sources. Considering `type`, it would require a mapping between the product types used in OpenAIRE and those used in IRIS (since the script is not customized for IRIS SNS, but is designed to fit any institutional repository).

5.2 Reliability of OpenAIRE affiliations

Of the 295 authors in AL, only 88 have at least one publication linked to their ORCID in OpenAIRE. This suggests that $PfAL$ retrieves fewer research products than AP, as shown in Figure 3. For this reason, OA is largely composed of records originating from AP, which includes authors whose current affiliation with the SNS cannot be guaranteed.

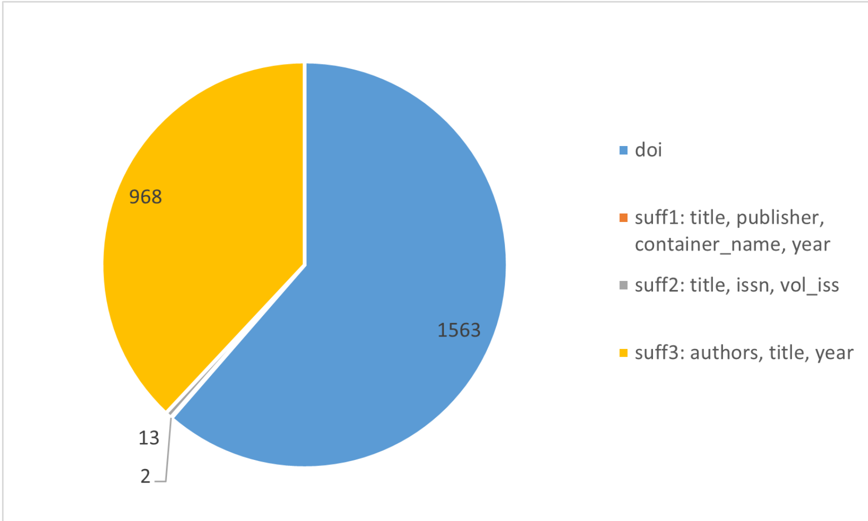

Figure 2: Matching Rules Frequency (2021-2025 combined)

These considerations do not imply that OpenAIRE fails to index the publications of the authors under examination. Firstly, some authors have no publications at all, as they belong to the technical and administrative staff. For most researchers, however, this phenomenon can be explained by the fact that the source from which the metadata are harvested does not expose the ORCID and therefore does not allow for the unambiguous attribution of research products. OpenAIRE, unlike some commercial platforms, does not assign publications based solely on name strings, in order to avoid mismatches caused by homonymy and potential inconsistencies arising from abbreviated or publisher-dependent name formats.

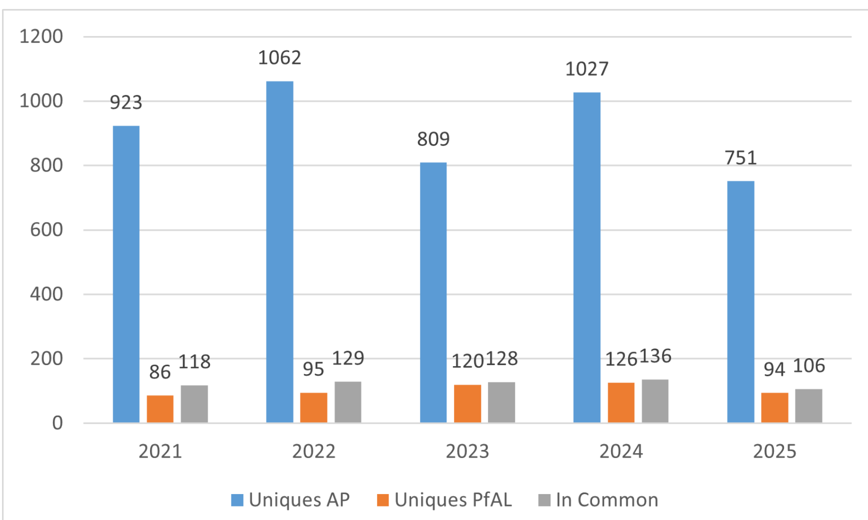

Figure 3: AP and $PfAL$ Composition per Year

6 Conclusion and Future Developments

To answer the first research question, the script could help populate institutional repositories by successfully identifying a substantial number of research products in OpenAIRE for integration.

However, the actual expansion of the archive must be subject to human oversight and a bulk import from OpenAIRE is not feasible yet because the correspondence is not entirely certain.

Regarding exact matches, some records may have missing or incorrect DOI, for example the DOI of the book instead of the specific DOI of the book chapter. In other instances, disambiguation failed because DOI was stored in the `alternateIdentifiers` field instead of `pids`[14]; it would be useful if OpenAIRE had merged `pids` with `alternateIdentifiers`. A potential development of the script to identify false positives could be to compare the other fields as well, even for exact matches via DOI.

Regarding sufficient matches, the reliability of the results depends on the information quality of the metadata drawn from OpenAIRE data sources.

The weighted matches can be useful for identifying additional products to be imported, but the integration must be carried out with great caution. This functionality of the script should be further developed: on the one hand, a more accessible format than a list of identifier pairs would be necessary to facilitate the evaluation of the association; on the other hand, a reassessment of the weights assigned to each field would be required.

The current weighting system (§ Table 5) was assigned arbitrarily and, as shown by the various confidence distributions (§ 4.4), the threshold value of 1 is often exceeded. A recalibration of the weights is therefore necessary. A potential development would be to analyze the various non-secure matches proposed and determine which sets of matched fields most frequently correspond to an actual match, in order to build a Gold Standard from which the appropriate weights could be calculated through statistical methods.

The analysis reveals also that several products in IRIS have multiple corresponding records in OpenAIRE in some cases where two or more records originating from different sources have either different DOIs or provide a DOI for only one of them. A potential development could involve identifying which of the duplicate record contains the most complete metadata and reporting it to OpenAIRE to support its deduplication efforts. This could also be useful for assessing the quality of the data sources on which OpenAIRE relies upon, thereby improving the ranking algorithm used for deduplication [14].

To answer the second research question about the reliability of OpenAIRE affiliation, both $PfAL$ and AP do not represent authors' affiliations, since at the time of writing OpenAIRE does not record authors' affiliations directly within the Persons entity. Work on this feature has already begun on the OpenAIRE side, and this study highlights the need for such functionality, further supporting the decision to pursue this line of development.

To conclude, OpenAIRE may become a viable alternative to commercial platforms once the quality of its data will enable accurate identification of products, affiliations, and authors. Collaborations such as the one between SNS and OpenAIRE help move closer to achieving this goal. On the side of institutional repositories, accelerating progress toward this goal would require implementing product validation, thereby ensuring a shared standard not only for which metadata are exposed but also for the quality of their content.

7 Acknowledgments

The authors would like to thank Giulia Malaguarnera, Stefania Amodeo, and Miriam Baglioni, from OpenAIRE, for their valuable support and for providing clarifications on how the Ope-

nAIRE Graph collects and structures data.

References

[1] ANVUR. Bando vqr 2020-2024. https://www.anvur.it/sites/default/files/2024-11/Bando-VQR-2020-2024_31ottobre.pdf, 2024.

[2] Barcelona Declaration Network. Barcelona Declaration on Open Research Information. https://barcelona-declaration.org/, 2026.

[3] Cory Benfield, Ian Stapleton Cordasco, and Nate Prewitt. Requests: Python http for humans. https://pypi.org/project/requests/, 2025.

[4] Alessandro Cignoni, Daniele Marotta, and Donatella Tamagno. Irvoapy. https://github.com/alessandrocignoni-sns/IRVOApy, 2026. Software Heritage persistent identifier: swh:1:dir:6f8f0d6cd9bbf75aefa921074978a7977cc4b665;.

[5] Alessandro Cignoni, Daniele Marotta, and Donatella Tamagno. Irvoapy: script input and output. https://doi.org/10.5281/zenodo.18716057, 2026.

[6] CINECA. Iris data structure ods ir-l1. https://wiki.u-gov.it/confluence/display/public/UGOVHELP/ODS+-+IR-L1, 2026.

[7] CINECA. Iris institutional research information system. https://wiki.u-gov.it/confluence/display/public/UGOVHELP/IRIS+-+Institutional+Research+Information+System, 2026.

[8] Charlie Clark and Eric Gazoni. Openpyxl: A python library to read/write excel 2010 xlsx/xlsm files. https://pypi.org/project/openpyxl/, 2024.

[9] DSpace. DSpace. https://dspace.org/, 2024.

[10] Myrto Kallipoliti, Serafeim Chatzopoulos, Miriam Baglioni, Eleni Adamidi, Paris Koloveas, and Thanasis Vergoulis. From raw affiliations to organization identifiers. https://doi.org/10.1007/978-3-032-05409-8_8, 2026.

[11] OpenAIRE. Openaire graph documentation - propagation. https://graph.openaire.eu/docs/graph-production-workflow/deduction-and-propagation/propagation, 2022.

[12] OpenAIRE. Openaire graph. https://graph.openaire.eu/, 2026.

[13] OpenAIRE. Openaire graph apis documentation. https://graph.openaire.eu/docs/apis/home/, 2026.

[14] OpenAIRE. Openaire graph documentation - deduplication. https://graph.openaire.eu/docs/graph-production-workflow/deduplication/, 2026.

[15] OpenAIRE. OpenAIRE: Open Access Infrastructure for Research in Europe. https://www.openaire.eu/, 2026.

[16] Scuola Normale Superiore. Regolamento in materia di accesso aperto alla letteratura scientifica. https://www.sns.it/sites/default/files/2026-01/regolamento_accesso_aperto_modificato_23122025_web.pdf, 2025.

8 Author Biographies

Alessandro Cignoni works in the Research Assessment and Open Science Service of Scuola Normale Superiore. As a Digital Humanist, he has worked on Historical GIS with University of Pisa's Laboratorio di Cultura Digitale and history of computing with independent research project Hackerando la Macchina Ridotta. As a programmer, he specializes in web development, relational databases, and Python scripting.

Daniele Marotta works in the Research Assessment and Open Science Service of Scuola Normale Superiore. He has a background in the humanities and has twenty years of experience in digital humanities projects. In particular, he contributed to the creation of a digital archive of pre-Islamic South Arabian inscriptions and to the digital edition of the Ilāhiyyāt of the Kitāb al-Šifā' by Avicenna, within the framework of ERC projects DASI and PhiBor.

Donatella Tamagno is Head of the Research Assessment and Open Science Service at Scuola Normale Superiore. After earning a PhD in Ancient Studies, she worked for a decade in the academic publishing sector—particularly in open access—and developed extensive expertise in scholarly communication. She coordinates support services for researchers on copyright and open access, research data management, open science, and research assessment. She is a member of the university's Open Science Committee.