Investigating Bibliographic Entities Without Persistent Identifiers

Erica Andreose[1], Leonardo Zilli[1]

[1]*Digital Humanities and Digital Knowledge, Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy*

Purpose

The identification and reconciliation of research outputs without persistent identifiers (PIDs), such as DOI (Digital Object Identifier), are critical challenges in scholarly communication. This study proposes an initial investigation of the IRIS No ID dataset, which contains research publications from the University of Bologna's IRIS repository lacking a DOI or other standardized PIDs.

The Institutional Research Information System (IRIS) is the University of Bologna's implementation of a Current Research Information System (CRIS), designed to collect and manage data related to research activities and outputs. The importance of IRIS is closely linked to the broader Open Science movement, which emphasizes transparency, research integrity, and collaboration. According to the university's Open Access Policy, all research outputs must be archived in publicly accessible repositories, particularly when supported by public funding. This not only ensures transparency but also plays a key role to the university's research evaluation system, where archived publications and datasets contribute to internal and external assessment exercises, influencing funding decisions and academic rankings (Bollini et al. 2016).

However, a significant portion of IRIS data lacks persistent identifiers such as DOI, ISBN, or PMID. This research aims to explore the factors contributing to the absence of PIDs, identify potential correlations with metadata attributes, and assess the feasibility of retrieving missing identifiers by querying external bibliographic databases.

Research Questions

RQ1. Is there a correlation between the different metadata attributes of the Iris No ID (INOID) dataset and the absence of persistent identifiers?

RQ2. To what extent do the metadata available in the INOID dataset allow for successful identification of publication in external bibliographic databases?

RQ3. What is the coverage of the publications contained in INOID in other sources of open research information such as Crossref and OpenCitations?

Materials and methods

Data – The data used in this study derives the work described in Andreose et al. (2025) which analysed the coverage of the IRIS data dump within other sources of open research information, namely OpenCitations Meta (Massari et al., 2024) and Index (Heibi et al., 2024). The analysis led to the creation of four distinct datasets derived from the initial IRIS dump

(Amurri et al., 2024). The dataset under examination in the current study is Iris No ID (INOID) (Zilli et al., 2024), which includes metadata for the 103,481 entries from the IRIS dump that do not have a DOI, ISBN or PMID identifier.

Methodology – A first inspection of the dataset reveals two entries with "null" values in both the title and authors fields, which we discard. To standardize formatting, all author's names are normalized to lowercase.

We check for duplicate entries by identifying rows with identical values in both the title and author fields, finding 2,915 duplicates. We deduplicate these by retaining only the first occurrence of each, resulting in a refined dataset consisting of 101,742 unique bibliographic records (BRs).

To explore potential answers to Research Question 1 (RQ1), we analyze the deduplicated dataset's attributes to identify potential correlations between the metadata of the BRs and the absence of persistent identifiers. We aggregate values from selected attributes in the dataset to determine the number of unique entries for each. Specifically, we count the number of unique values in the fields *type* (column "OWNING_COLLECTION_DES"), *publishing country* (column "PUB_COUNTRY") and *publication year* (column "DATE_ISSUED_YEAR"). The resulting value distributions were then visualized using plots to assess whether any patterns or correlations emerge.

To address Research Question 2 (RQ2) and 3 (RQ3), we make use of Crossref's public REST API to search for the identifiers of the INOID entries within their databases. For each BR, a search query is constructed by concatenating the *title* with the list of *authors*, separated by a comma. Following the guidelines provided by Crossref, we use the "query.bibliographic" parameter and limit the number of requested matches to a maximum of two results per query. From the returned result, we extract the *author, title, DOI, ISSN* and *relevance score* fields. Records with a relevance score greater than a threshold of 85 were considered positive matches. This threshold was determined based on a preliminary manual evaluation of a sample of queries.

The retrieved DOIs are normalized following OpenCitation's naming convention (*doi:identifier*) and each identifier is searched within the OpenCitations Meta data dump (OpenCitations, 2024) following the same methodology described in Andreose et al. (2025).

An implementation of the adopted workflow is provided in the form of a Jupyter Notebook, accessible via nbviewer.org. (https://nbviewer.org/gist/leonardozilli/dd4ce14e9d6240a4c4136e36acf58838 (accessed 30/01/2025))

Limitations

One of the main challenges in utilizing the IRIS dataset is the presence of duplicate records, which complicates the retrieval of unique BRs. Problems such as the lack of a standardized naming convention for authors exacerbates this issue, making disambiguation particularly difficult. A more thorough methodology could be developed by integrating the additional metadata describing the individuals associated to the IRIS BRs, as contained in the original IRIS dataset.

Regarding the matching of BRs within Crossref, while generally the internal scoring mechanism (based on the BM25 algorithm) provides valuable results, we found many cases in which results with a high relevance score corresponded to incorrect matches. We found this issue to be especially common in queries containing a long list of authors, which can inflate the lexical similarity score leading to erroneous matches. To improve the accuracy of retrieval, a more refined search methodology should be devised.

Results

The typology analysis revealed that the most common entry types are 1) journal articles, 2) abstracts, and 3) conference proceedings. While journal articles may lack DOIs due to indexing delays, abstracts and conference papers often remain without PID by default. The most frequent publishing country was “unspecified” (73.4%) followed by Italy, USA and UK. The temporal analysis showed a significant increase in missing DOIs between 2004 and 2008. A drastic spike in overall IRIS uploads can be observed in 2004, likely due to the migration of records from the old internal research registry into IRIS. Since 2006, the trend of entries without PIDs has been steadily declining, reflecting the growing adoption of Open Access practices. It is worth noting that 2,498 IRIS BRs presented a placeholder date “9999”, and were not included in the analysis. The answer to RQ1 is therefore negative, however these analyses revealed curious insights into metadata completeness and the evolution of Open Access practices through time.

Concerning RQ2, preliminary results from the retrieval attempt of the INOID BRs in Crossref revealed a total of 18,664 successfully reconciled DOIs, indicating an initial level of success in linking the metadata to external bibliographic databases. However, 5,754 of these DOIs were associated with multiple IRIS entries, indicating potential metadata inconsistencies or duplicate entities.

The mapping of these DOIs within OpenCitations Meta was able to identify 10,387 records already included in OpenCitations Meta, providing an answer to RQ3.

Value

Our findings reveal that a significant portion of the IRIS No ID dataset can be successfully linked to external bibliographic databases, with 18% of entries reconciled with CrossRef and 10% with OpenCitations.

While no strong direct correlation was found between specific metadata attributes and the absence of persistent identifiers, the temporal analysis revealed a positive trend of decline over time of records with missing PIDs.

This study serves as a starting point for future efforts in reconciling and integrating IRIS entities into other open research information sources, increasing the coverage of Italian publications within Open Science infrastructures and thus fostering transparency and accessibility.

References

Amurri, A., Giachino, E., & Peroni, S. (2024). UNIBO IRIS bibliographic data dump, dated 4 June 2024 (Version 4 June 2024) [Tabular data; CSV]. AMSActa. https://doi.org/10.6092/unibo/amsacta/7736

Andreose, E., Di Marzo, S., Heibi, I., Peroni, S., & Zilli, L. (2025). Analysing the coverage of the University of Bologna's publication metadata in an existing source of open research information. arXiv preprint. https://arxiv.org/abs/2501.05821

Bollini, A., Mennielli, M., Mornati, S., & Palmer, D. (2016). IRIS: Supporting & managing the research life-cycle. Universal Journal of Educational Research, 4(10), 738–743. https://doi.org/10.13189/ujer.2016.040410

Heibi, I., Moretti, A., Peroni, S., & Soricetti, M. (2024). The OpenCitations Index: Description of a database providing open citation data. Scientometrics, 129(12), 7923–7942. https://doi.org/10.1007/s11192-024-05160-7

Massari, A., Mariani, F., Heibi, I., Peroni, S., & Shotton, D. (2024). OpenCitations Meta. Quantitative Science Studies, 5(1), 50–75. https://doi.org/10.1162/qss_a_00292

OpenCitations. (2024). OpenCitations Meta CSV dataset of all bibliographic metadata (Version 8) [Tabular data; CSV]. Figshare. https://doi.org/10.6084/m9.figshare.21747461.v8

Zilli, L., Andreose, E., & Di Marzo, S. (2024). Iris No ID [Dataset]. Figshare. https://doi.org/10.6084/m9.figshare.25897759.v2