

Jun 26, 2025

Version 3

Open Science 24/25 - Protocol V.3

DOI

<https://dx.doi.org/10.17504/protocols.io.4r3l264ppv1y/v3>

Alberto Ciarrocca¹, Anna Nicoletti¹, Ahmadreza Nazari¹, Pietro Tisci¹, Lucrezia Pograri¹, Martina Pensalfini¹, Sergei Slinkin¹

¹University of Bologna

Open Science



Silvio Peroni

University of Bologna

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.4r3l264ppv1y/v3>

External link: <https://github.com/open-sci/2024-2025>

Protocol Citation: Alberto Ciarrocca, Anna Nicoletti, Ahmadreza Nazari, Pietro Tisci, Lucrezia Pograri, Martina Pensalfini, Sergei Slinkin 2025. Open Science 24/25 - Protocol. **protocols.io**

<https://dx.doi.org/10.17504/protocols.io.4r3l264ppv1y/v3> Version created by **Alberto Ciarrocca**

License: This is an open access protocol distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: June 25, 2025

Last Modified: June 26, 2025

Protocol Integer ID: 221022

Keywords: open science, university of bologna, unibo, alma mater studiorum, iris, researchers across various repository, citation impact of unibo, integration of these research object, research object, overlap among these repository, such research object, citation analysis tool, current research information system, various repository, much of such research object, analyzed metadata, repository, research output, open science practices at the university, existing repository, citation dynamic, citation pattern, dissemination of diverse research output, research, many citation, researcher, coverage of these research output, diverse research output, strengthening open science practice, citation impact, generalist repository, repository usage, open science, protocol purpose this study, research question, unibo personnel, metadata, current research, methodology the project team, university of bologna, dissemination, providing insight, dissemination pattern, unibo, database, project team, integration within unibo, opencitation

Abstract

Purpose

This study aims to evaluate the representation and dissemination of diverse research outputs - such as software, databases, exhibitions, audio-visual materials, and others - produced by the University of Bologna's (UNIBO) researchers across various repositories. It seeks to assess the extent of overlap among these repositories, analyze citation dynamics involving these research objects, and determine their integration within UNIBO's [Current Research Information System (IRIS)](<https://cris.unibo.it>).

The Research Questions (RQs) are formulated as follows:

1. What is the current coverage of these kinds of research objects created by UNIBO personnel in existing repositories?
2. Is there any overlap among these repositories (i.e., research objects deposited in more than one)?
3. How many citations (incoming and outgoing), as in OpenCitations, are these research objects involved in?
4. How much of such research objects are actually mapped in IRIS?

Methodology

The project team has systematically collected and analyzed metadata from selected institutional, disciplinary, and generalist repositories, including [AMS Acta](<https://amsacta.unibo.it/>), [Software Heritage](<https://www.softwareheritage.org/>), and [Zenodo](<https://zenodo.org/>). Relevant data and metadata were extracted using APIs and web scraping techniques. The analysis identified cross-repository depositions to assess overlaps and employ citation analysis tools, such as [OpenCitations](<https://opencitations.net/>), to examine citation patterns. Additionally, the collected data were cross-referenced with IRIS to evaluate its coverage of these research outputs.

Value

This study shed light on the dissemination patterns and citation impact of UNIBO's diverse research outputs, providing insights to enhance their visibility and accessibility. The findings can inform strategies to optimize repository usage and improve the integration of these research objects within IRIS, ultimately strengthening open science practices at the University of Bologna and promoting a more cohesive approach for maximizing the impact of its research outputs.



Before start

Note: Python 3.13.2 or previous versions are required to run the software.

Prepare the working environment by executing the following commands:

```
# Clone the repository
git clone https://github.com/open-sci/2024-2025

# Move to the repository folder
cd 2024-2025

# Install dependencies
pip install -r requirements.txt
```



Data Gathering

- 1 From **Zenodo** use the RESTful API protocol to harvest metadata in JSON format.

```
API_BASE = "https://developers.zenodo.org/#rest-api"
```

All the records were retrieved through the "**records**" operation of the API (<https://zenodo.org/api/records>), by querying for **affiliation** (of creators and contributors) and **ORCID IDs** extracted from IRIS (for creators and contributors).

The script used for the query is in [ZenodoDataExtraction.ipynb]

Expected result

ZenodoData.json file that follows the *JSON Metadata Schema* containing the properties to describe research objects, such as:

'title', 'doi', 'publication_date', 'description', 'access_right', 'license', 'creators', 'contributors', 'keywords', 'language', 'resource_type', 'relations', 'swb'

Dataset

ZenodoData.json

NAME

<https://zenodo.org/records/15741523/files/ZenodoData.json>^{LINK}

- 2 Metadata was harvested and filtered from the **AMS Acta** repository using the OAI-PMH API and the Python library Sickle.

```
API_BASE = "https://amsacta.unibo.it/cgi/oai2"
```

Metadata was collected to obtain **ePrint IDs**, which were then used to retrieve enriched JSON records. A two-step **filtering process** - based on author affiliations and ORCID identifiers - was applied to isolate records associated with the University of Bologna. The final dataset includes approximately 1,000 publications, highlighting various research outputs including datasets and software.

The script used for the query is in [AMSActaReduced.ipynb] and the script for the filtering is in [filteringAMSActa.ipynb]

Expected result

amsacta_filtered_affiliation_or_orcid_doubles.json file in *Dublin Core* format containing the properties to describe research objects, such as:

'uri', 'creators', 'monograph_type', 'publisher', 'title', 'date', 'doi', 'eprintid', 'keywords', 'type', 'abstract'

Dataset

amsacta_filtered_affiliation_or_orcid_doubles.json

https://zenodo.org/records/15741523/files/amsacta_filtered_affiliation_or_orcid_doubles

- 3 Query **Software Heritage** API endpoint to identify software repositories affiliated with the University of Bologna.

```
API_BASE = "https://archive.softwareheritage.org/api/1/"
```

An initial query on the API endpoint was performed on relevant keywords, and later filtered using heuristic-based keyword searches and email domain analysis: the pipeline filters candidate origins by inspecting **README content** and **commit metadata** for institutional markers.

The script used for the query and filtering is in [swh_data_extraction.ipynb]



Expected result

unibo_repositories_swh.json file of validated repositories containing the following properties:

'url', 'rev', 'dir_id'

The '**authors**' object is added during the processing and contains the '**name**' and '**email**' properties.

Dataset

unibo_repositories_swh.json

NAME

https://zenodo.org/records/15741523/files/unibo_repositories_swh.json

LINK

- 4 The newest version (May 30th 2025) of the **IRIS** data dump was retrieved to be later used in our research.

Dataset

UNIBO IRIS bibliographic data dump

NAME

<https://amsacta.unibo.it/id/eprint/8374/>

LINK

It is comprised of six CSV files describing 412,272 bibliographic entities and with a total uncompressed size of 413 MB (125 MB zipped).

5 Convert and Normalize Data across datasets

- 5.1 Read all datasets obtained in previous steps with **pandas** and extract necessary information from nested key-value pairs.

- 5.2 **Normalize the text** (deleting white spaces, removing html tags from text...), **drop duplicates** and **normalize all the column labels**.



Additionally, each entry has been flagged with the corresponding origin dataset, expressed in the '**src_repo**' column.

- 6 **Merge vertically** all four datasets and output **one csv file** of 419,829 records with a size of 120 MB.

The script for the conversion, normalization and merging is in [mashup.ipynb]

Expected result

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
	title	id	doi	creators	orcid	date	description	resource_type	url	type	rights	publisher	relation	communities	sw_h_id	keywords	src_repo	issn	pmid

mashup.csv

Dataset

mashup.csv

NAME

<https://zenodo.org/records/15741523/files/mashup.csv>^{LINK}

Data Analysis

- 7 **Filter the mash-up data** to obtain research objects meaningful to our research purposes.

To do so we created the dictionary '**iris_map**' lists the 15 IRIS categories relevant to our project. For every incoming row the helper '**classify_row()**' concatenates the '**type**' and

'resource_type' fields, lowers the case and tests the strings against the regular-expression list attached to each IRIS code.

Rows that match any pattern receive a new label 'iris_cat', the others are discarded on the fly.

All labelled chunks are concatenated.

Expected result

mashup_IRIS_subset_v3.csv (8 699 rows, 20 columns – the original metadata + iris_cat)

- 8 **Answer RQ1** (*"What is the current coverage of these kinds of research objects created by UNIBO personnel in existing repositories?"*) by counting in **mashup_IRIS_subset_v3.csv** how many research objects we have extracted from each dataset.

The script for the filtering of the data and for answering RQ1 is [RQ_1.py]

- 9 **Answer RQ2** (*"Is there any overlap among these repositories (i.e., research objects deposited in more than one)?"*) by detecting duplicates among all repositories in **mashup_IRIS_subset_v3.csv**.

We accomplished this by comparing both the 'doi' and 'title' columns and producing a new table outlining the amount of overlapping objects for each pair of datasets.

Expected result

duplicate_objects_v3.csv

The script for answering RQ2 is [RQ_2.py]

- 10 **The Answer to RQ4** (*"How much of such research objects are actually mapped in IRIS?"*) is already present in **duplicate_objects_v3.csv**, since IRIS is part of **mashup_IRIS_subset_v3.csv**.
- 11 **Answer RQ3** (*"How many citations (incoming and outgoing), as in OpenCitations, are these research objects involved in?"*)

The script for answering RQ3 is [RQ_3.py]



- 11.1 We query **OpenCitation**'s API to retrieve the number of **ingoing and outgoing citation** for each DOI available.

```
API_BASE = "https://opencitations.net/index/api/v2"
```

We used '**citations**' operation to retrieve ingoing citation and '**references**' to retrieve outgoing citations; the resulting JSON file was then converted in CSV format.

Expected result

	A	B	C	D	E	F
	doi	cit_num_outgoing	cit_num_ingoing	oci_outgoing	oci_ingoing	status

CitationsCount.csv

- 11.2 After collecting all the citations from the API we needed to **reassign each DOI to its original repository** connecting the list of DOI used to the corresponsive '**src_repo**' from **mashup_IRIS_subset_v3.csv**
- 11.3 With the method '**.iterrows()**' we calculated the total of ingoing and outgoing citations for each repository and stored it in two dictionaries '**ingoing**' and '**outgoing**' where the keys are the 4 repositories and the values are the sum of the citations

Data Visualisation

- 12 In order to **visualise key findings** from our research, we chose to represent the data obtained in the Data Analysis section in **bar charts**, using **pandas** and **matplotlib.pyplot**.
- 13 A Jupyter-Notebook was created representing the results with visualisations

Expected result

DataVisualisation.ipynb



Data Publication

- 14 All software has been published [here](#)
All datasets have been published [here](#)