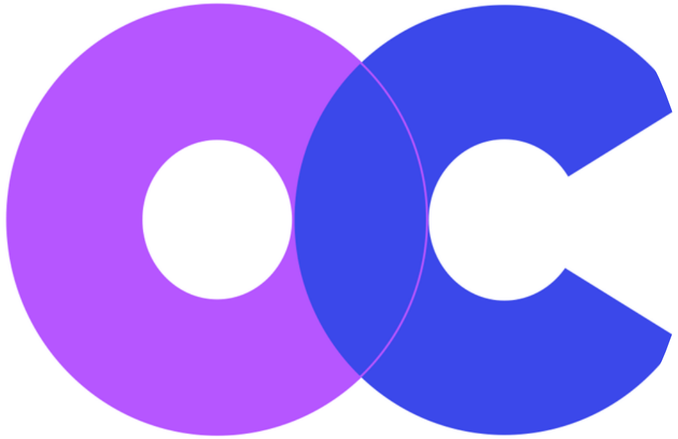
OpenCitations

- is an **independent not-for-profit infrastructure** organization
- managed by the Research Centre for Open Scholarly Metadata at the **University of Bologna**
- dedicated to the **publication of open bibliographic and citation data,** and it is a founding member of the Initiative for **Open Citations (I4OC)**
- exploits **Semantic Web** (Linked Data) technologies
- currently counts:
  - **1.8 B** unique citations
  - **84.7 M** bibliographic records
  - **68.4 M** cited entities
  - **67.9 M** citing entities



Datasets Evolution

Origins

In **2011** we released the **Open Citations Corpus**, a repository of scholarly citation data openly provided under a CC0 1.0 public domain license.
The **initial data** comprised the reference lists of **204,637 articles** from the Open Access Subset of PubMed Central, with a total of **6,325,178 citation data**

COCI and CROCI

A **citation index** is a bibliographic collection of citation records between documents.
In 2018, we launched **COCI**, our inaugural citation index, derived from open reference lists in the Crossref database that also identified documents using DOIs.
The following one was CROCI, for collecting crowdsourced open citations.

Since 2022, our systematic integration of new data sources has necessitated a workflow revision to optimize execution time, storage space, and coding effort

Data Sources

COCI (2018)

Crossref

DOI

DOCI (2022)

DataCite

DOI

POCI (2022)

NIH-OCC

PMID

OROCI (2023)

OpenAIRE

DOI, PMID, PMC, ARXIV

JOCI (EOY 2023)

JALC

DOI

Each of the sources provides data for a citation index (i.e.: COCI, DOCI, POCI, OROCI, JOCI, now also converging in a unified index), and bibliographic entity metadata, managed by **META**.

OpenCitations Meta (2022) is a software tool and database for metadata curation and management, housing over 99 million publications with rich bibliographic details. It covers more than 300 million contributors, including authors, editors, and publishers.

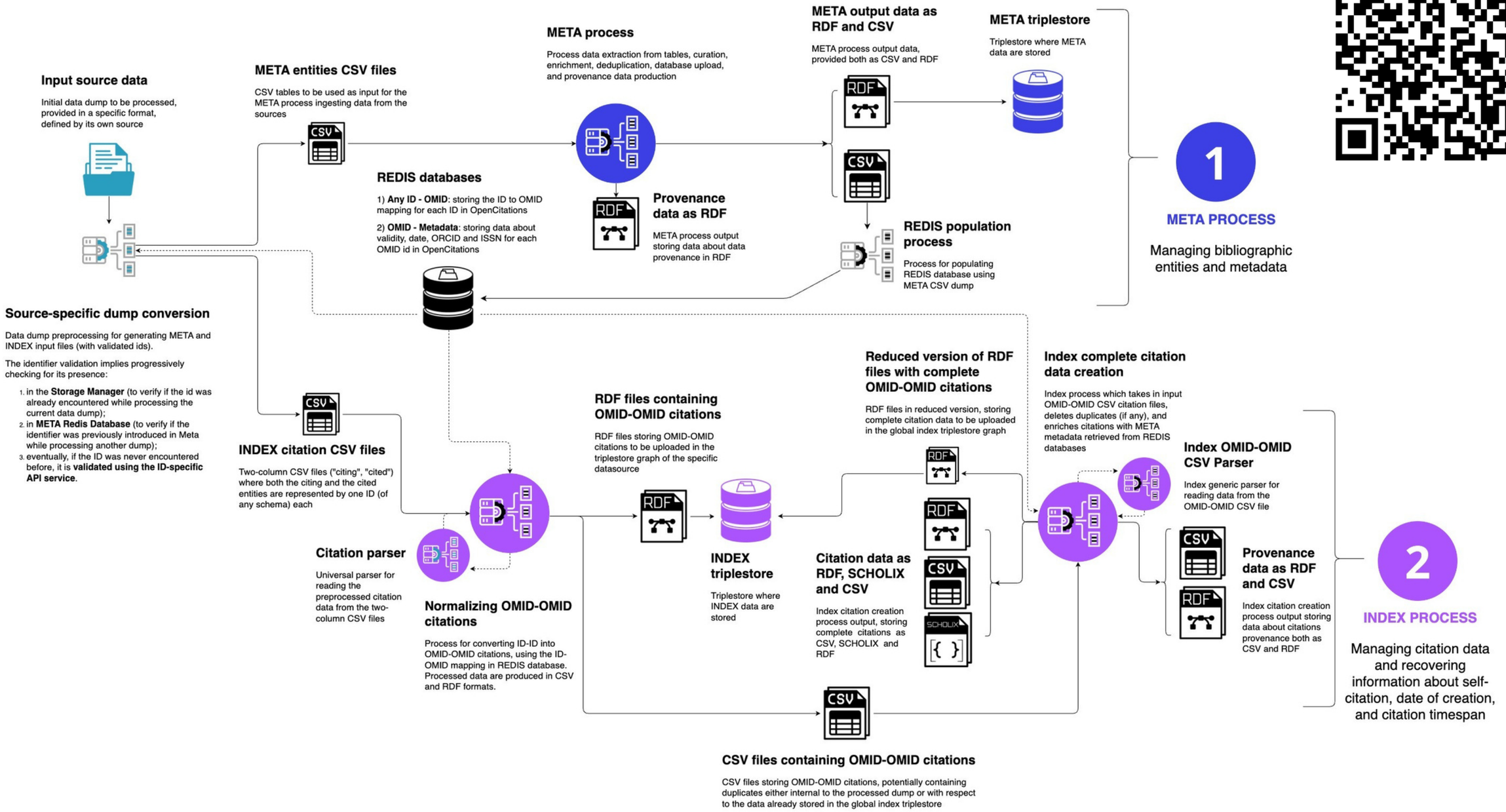
Workflow Objective

The main aim of the renewed ingestion workflow was to **perform an efficient metadata crosswalk** from the specific data models and file formats adopted by the data sources toward the OpenCitations data model (OCDM) and the formats we use as input of our subprocesses (CSV) to publish data (CSV, SCHOLIX, RDF)

Key aspects and challenges

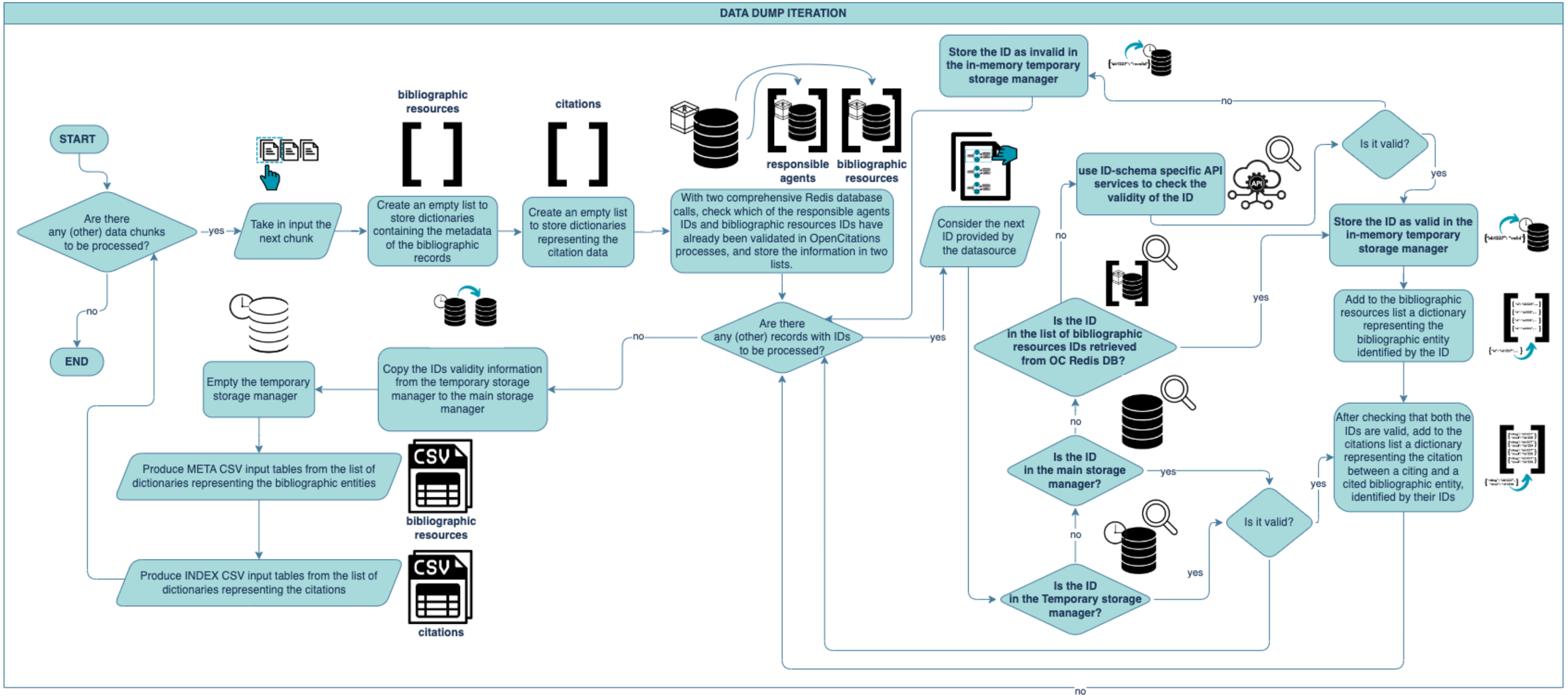
1. **Code revision for** enhancing **reusability** across multiple processes
2. **Parallel execution** enabled through duplicate minimization.
3. Data **duplication prevention** rather than post-removal.
4. **Streamlined process** for all sources, **easily extendable** with source-specific **plugins**.
5. **Identifier validation** system with cascading checks and data reliability heuristics, aiming to **minimize API usage**.

Workflow Overview



IDs Validation

In the source-specific dump conversion phase, we perform **metadata crosswalks** toward the OpenCitations data model, with the aims of (1) **producing input material** for the further steps of the ingestion workflow, and (2) **validating the IDs**



Output Data

Currently maintained datasets

Citations

- Unified OMID-OMID index
- RDFs specifying the data source provenance

Bibliographic Entities

- META database for bibliographic entity metadata management and curation

OMID: new OpenCitations unique identifier for entity disambiguation, since when OC manages other PIDs beyond DOIs

How to query OC data

- Download data dump from figshare
- Query the SPARQL endpoint
- Use the API services
- Search and browse interfaces

Output data formats

- CSV
- RDF
- SCHOLIX

	INDEX	COCI	DOCI	POCI	OROCI
INDEX	1,836,126,096	1,414,884,127	169,814,412	695,988,810	14,645,838
COCI	1,414,884,127	962,047,039	26,703	449,058,531	3,751,854
DOCI	169,814,412	26,703	169,663,603	9,623	114,483
POCI	695,988,810	449,058,531	9,623	237,208,867	9,711,789
OROCI	14,645,838	3,751,854	114,483	9,711,789	1,067,712