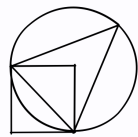


Crossref 2026 annual meeting and board election · 22 October 2026

biblio-glutton

Fast, local reference matching on Crossref data

Luca Foppiano · ScienzaLAB (www.scencialab.com), Portugal



ScienzaLAB



Crossref



GROBID

WHERE THE REFERENCES COME FROM

GROBID turns PDFs into structured data

1 · Extract

PDFs are made for humans.
GROBID recovers header, body
and references as TEI XML.

2 · Match

Extraction is strong but not
perfect. Each reference is
looked up in **Crossref** or
biblio-glutton.

3 · Enrich

Matched records fill gaps and
fix errors: **+12% F1** on header
[1].

[1] <https://grobid.readthedocs.io/en/latest/Consolidation/>

WHY BIBLIO-GLUTTON?

Citation consolidation with Grobid

- **Who needs it** Publishers, indexes, libraries, repositories, funders.
- **The real bottleneck** Large scale citation lookup: 20 request/article → 20M requests/1M articles to Crossref
- **Requests do not all cost the same:** lookups are cheap, matches are expensive

In Q1 2026, Crossref received around 100 million requests from Grobid clients

WHAT BIBLIO-GLUTTON IS

A local metadata lookup service built on Crossref

100% Crossref

- Based on the Crossref public dump
- Identifier or metadata lookup → full Crossref record

Additional features

- {PMID, PMCID, ISTEEX, HAL}ID → DOI → lookup;
- Includes OpenAlex
OpenAccess PDF link
- Citation string → Grobid → lookup by "strong fields"

Daily sync with crossref

with the paid Metadata Plus plan

WHAT WE MATCH AGAINST

Open records, from many sources

References are incomplete, abbreviated, misspelled: records must be **searchable by content**, not only reachable by identifier.

Source	What it brings	Records
Crossref	DOI metadata: title, authors, journal, volume, pages, funders	~170 M
PubMed Central	PMID ↔ PMCID ↔ DOI mapping	~42 M
HAL	French open archive, curated records	4.7 M
ISTEX	ISTEX identifiers and ark, with DOI and PII	33 M
OpenAlex	Open access link to paper	48 M
ROR · Funder Registry	Research organizations and funders	144,717 · 45,739

Counts from a fresh load in September 2026

WHAT'S NEXT

New version after 4 years

More data

- Open access links from OpenAlex, replacing Unpaywall
- ROR and the Crossref Funder Registry, refreshed monthly
- Updated PMID / PMC mapping from the deprecated FTP

Lighter

- One Java service: loading and indexing included, no Node.js indexer
- Storage 60% smaller: Zstandard with a dictionary trained on Crossref
- Dumps read straight from S3

Production-ready

- Java 21, Elasticsearch 8 with authentication
- Health endpoint, clear 503 instead of timeouts
- Resumable loads, tunable matching threshold

FUTURE WORK

Smarter matching, evaluated again

Improved matching

- Pairwise matching trained on contrastive examples, using lightweight approaches.
- Experiments with vector search and contextual embeddings

Updated evaluation

The 2021 benchmark redone on today's Crossref data, comparing all approaches on the same references.

Continued development

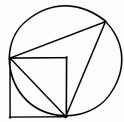
- Support future Crossref dumps
- Support new IDs

Thank you · Questions?

Open source, Apache 2.0 · no rate limits, no data leaving your network

Star us on github.com/kermitt2/biblio-glutton

Luca Foppiano · luca@scienzialab.com



ScienziaLAB



Crossref



GROBID