

biblio-glutton

Fast, local reference matching on Crossref data

We present biblio-glutton, a local lookup service with the same data as Crossref, plus extra identifiers and open access links. Give it an identifier or a raw citation; get back the Crossref record with its DOI.

Luca Foppiano · ScienziaLAB, Portugal · luca@scienzialab.com

01 Where the references come from: GROBID

GROBID, the open source library for scholarly PDFs, extracts each reference as structured fields. Its **consolidation** step then checks every reference against a metadata service: the public **Crossref API** by default, or **biblio-glutton**, running on your own machines.

Extract	Match	Enrich
GROBID: PDF → header, body and references as TEI XML.	Crossref API or biblio-glutton finds the real record.	DOI added, errors fixed: +12% F1 on the header.

02 Why biblio-glutton?

- **Scale:** in Q1 2026, Crossref received around **100 million** requests from GROBID clients.
- **Who needs it:** publishers, indexes, libraries, repositories, funders.
- **The real bottleneck:** large-scale citation lookup, one request per reference.
- **Requests do not all cost the same:** lookups are cheap, matches are expensive.

03 What biblio-glutton is

Written by Patrice Lopez and Luca Foppiano in 2018 · github.com/kermitt2/biblio-glutton · Apache 2.0

100% Crossref

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

Additional features

- PMID, PMCID, ISTEK or HAL ID → DOI → lookup
- Open access PDF link from OpenAlex
- Citation string → GROBID → lookup by “strong fields”

Daily sync with Crossref

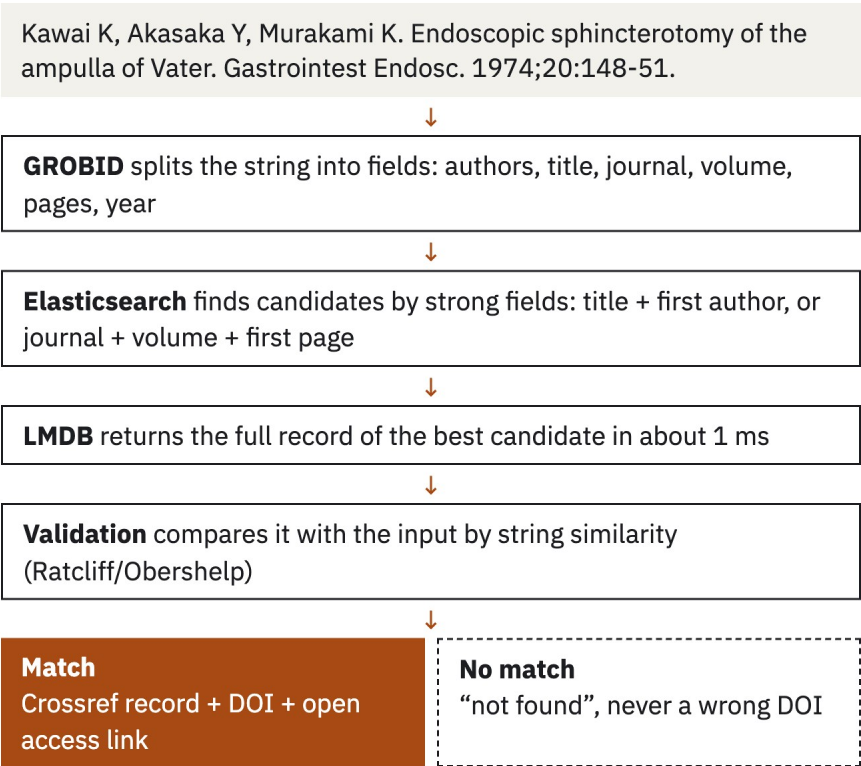
With the paid Metadata Plus plan

04 What we match against

Source	Brings	Records
Crossref	DOI metadata	~170 M
PubMed Central	PMID ↔ PMCID ↔ DOI	~42 M
OpenAlex	Open access PDF links	48 M
HAL	Curated French open archive	4.7 M
ISTEX	ISTEX IDs, ark, DOI, PII	33 M
ROR	Research organizations	144,717
Funder Registry	Funders	45,739

Counts from a fresh load, September 2026.

05 How a raw reference is matched



06 After 4 years, a new release: October 2026

More data	Lighter	Production-ready
• OpenAlex open access links replace Unpaywall • ROR and Funder Registry, refreshed monthly • Updated PMID ↔ PMCID mapping	• One Java service, no Node.js indexer • 60% less storage: Zstandard with a Crossref-trained dictionary • Dumps read straight from S3	• Java 21, Elasticsearch 8 with authentication • Health endpoint, clear 503 instead of timeouts • Resumable loads, tunable matching threshold

07 Future work

Improved matching. Pairwise matching trained on contrastive examples, with lightweight approaches; experiments with vector search and contextual embeddings.	Updated evaluation. The 2021 benchmark redone on today’s Crossref data, all approaches on the same references.	Continued development. Support for future Crossref dumps and new identifiers.
---	---	--