Automated Construction of FAIR Digital Object Knowledge Graphs from Flat Cultural Heritage Records

Zeyd Boukhers[1,2] · Lingxiao Kong[1,3] · Xenophon Zabulis[4] · Georgios Toubekis[1]

[1]Fraunhofer Institute for Applied Information Technology FIT, Germany · [2]University Hospital of Cologne · [3]University of Cologne · [4]Foundation for Research and Technology (FORTH), Greece

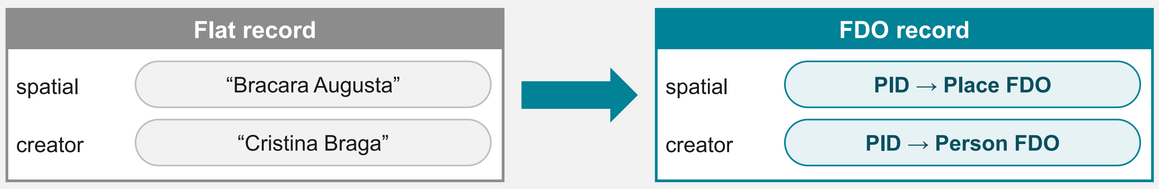
CIKM '26 · Rome, Italy · Paper DOI 10.1145/3799682.3840068

*Resolving metadata values to persistent identifiers buys **typed, resolvable, cross-lingually merged nodes, not more connectivity.** Graph connectivity, the usual proxy for machine actionability, does not separate this approach from byte-identical string matching.*

1 · The problem is a decision, not a linking task

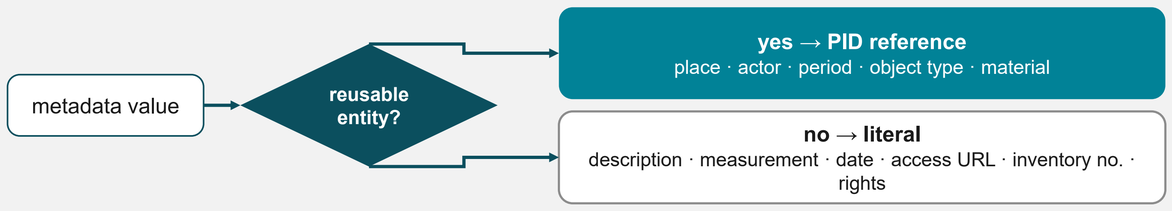
The gap

- Europeana's data model predates the FDO specification: most values are **plain text**.
- Fine for human browsing, but an automated agent has **nothing to follow** across records.
- No agent can tell that "Braga" and "Bracara Augusta" name the same place.



The PID/literal decision

- FDO specification: a value is a **PID reference if and only if** it denotes a reusable real-world entity; otherwise it stays a **literal**.



- Entity linkers get mentions already marked as entities; this pipeline must **first decide which values are entities at all.**

2 · Method: decide first, then resolve

1 Parse

europeana

637 records
→ **(field, value) slots**

2 Classify (LLM)

- semantic type (16 classes)
- PID or literal?
- target vocabulary

→ **one JSON call**

3 Resolve

- query the vocabulary
- fallback chains:

VIAF → Wikidata
AAT → Wikidata
PeriodO → AAT → Wikidata

- LLM reranks by context

4 Construct

- FDO kernel record
- Metadata FDO with CIDOC-CRM assertions

literal → bypasses resolution, stays a terminal leaf

Dataset and setup

- **637** archaeological records · **5** Europeana providers · **4** countries
- **12,720** metadata slots over content-bearing fields
- **9,197** already carry a Europeana vocabulary URI; **3,523** are plain strings
- gpt-oss-120b (120B, instruction-tuned) · temperature 0 · every response cached
- Code & outputs: **github.com/ZResearch/aFDO_CIKM**

Field restriction, reported

- Excluded: *organizations*, *rights*, *edmIsShownAt* (provenance, not content).
- *dcLanguage*, where the classifier fails: "en"/"English" → AAT *English Baroque* (334 records), "ro" → *Romanian (Cyrillic)* (148).
- **Relations are not generated:** the LLM returns type, PID/literal and vocabulary; CIDOC-CRM properties follow a fixed mapping (Place → P53, Person → P14, Group → P50).

3 · Results: what resolution buys, and what connectivity misses

58.5% resolved by the pipeline, of values Europeana left as plain text (1,798 / 3,075)

89.6% combined, including Europeana's pre-linked URIs (10,995 / 12,272)

800 entity FDOs · 9,054 CIDOC-CRM edges

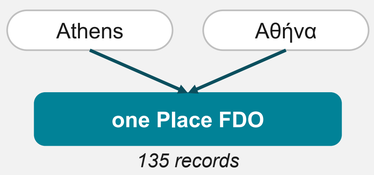
176 surface forms merged into 33 shared entities

17 / 33 of those merges correct on manual review

What resolution buys

- 33 entity FDOs absorb **176** surface forms that byte-identical matching keeps apart.
- All 33 reviewed by hand: **17 correct**: cross-lingual and variant alignments of places and organisations.

Merged forms	Records
Athens · Αθήνα	135
National Theatre of Greece · Εθνικό Θέατρο	29
Epidaurus · Επίδαυρος	12



- Every FDO node: resolvable identifier + semantic type (baseline: 63.8% of nodes).
- Label similarity not reported: it scores correct cross-lingual links at zero.

Where it fails

- The 16 wrong merges are not a language problem: they absorb values that are **not names**, e.g. 59 identifiers *share3d:*/TMP.1* → *Human parainfluenza virus 1*; 47 → *ISO 2-numbers*.
- Numerous but shallow: 142 forms, 163 records. Correct merges: 34 forms, 327 records.
- Type-vocabulary consistency is measurable only for PERIOD: **62.5%** on 331 slots; all 124 observed type errors fall in it.
- Strict vocabulary routing: **73.1%** (1,315 / 1,798).

Takeaway and next steps

- The weakness is deciding whether a value is a name, not reading names across languages.
- What separates the graphs is what the nodes are, not how many records connect.
- No human gold set yet: this bounds every quality figure.
- Next: gold set · dedicated entity linkers · downstream traversal or discovery task.
- One domain (archaeology), one aggregator: no claim of generalisation.

Connectivity is a weak discriminator

Table 1. Connectivity over 637 records, content-bearing fields, values in >10% of records removed. $G_{flat}^{=}$: same slots as G_{FDO}; G_{flat}^{+} adds content values the pipeline left unlinked.

Metric	G_{flat}^{+}	$G_{flat}^{=}$	G_{FDO}
Slots used	12,720	10,995	10,995
Distinct shared nodes	479	359	369
Connected components	8	32	**20**
Largest component (% records)	71.6%	69.4%	**73.6%**
Cross-provider bridges	13	12	13
Nodes with resolvable URI	63.8%	63.8%	**100.0%** †
Surface forms merged	0	0	**176**

† induced by construction, not measured.

- Same 10,995 slots: resolution cuts components **32 → 20**.
- All 12,720 content values: string matching reaches **8**, since unlinked values still connect records as byte-identical strings.
- The direction depends on the slot population, which the metric itself does not fix.
- Europeana ships URIs for 65% of slots, byte-identical across records: string matching recovers them without any resolution step.
- 1 of 13 cross-provider bridges is spurious (*Human parainfluenza virus 1*, 59 records, 2 countries).

This poster

Scan for the poster
doi.org/10.5281/zenodo.24033368

zeyd.boukhers@fit.fraunhofer.de

Acknowledgement

Supported by the European Union's Horizon Europe programme through the ARXIVE project (Grant Agreement No. 101233418) and the ARGUS project (Grant Agreement No. 101132308), and by the German Research Foundation (DFG) through NFDI4DataScience (project no. 460234259).

Contributing organisations · Fraunhofer FIT · Universität zu Köln · Uniklinik Köln · FORTH Foundation for Research and Technology - Hellas

Funded by · Funded by the European Union · ARXIVE · ARGUS