AI-Assisted Metadata Enrichment in Collaborative Knowledge Graphs

Making implicit bibliographic knowledge explicit while preserving transparency

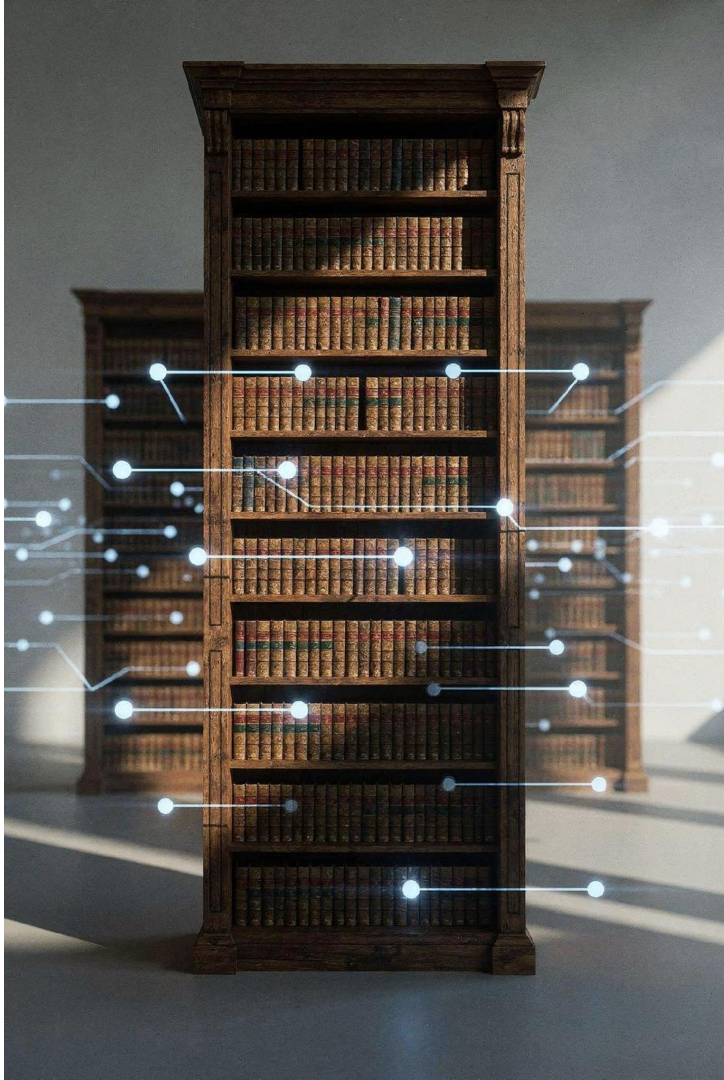
Tiziana Possemato - @Cult & Casalini Libri

DCMI 2026
International Conference on Dublin Core and Metadata Applications

Meaning-Driven AI: Using Metadata to Align Systems with Human Values

2026 August 3rd to 7th

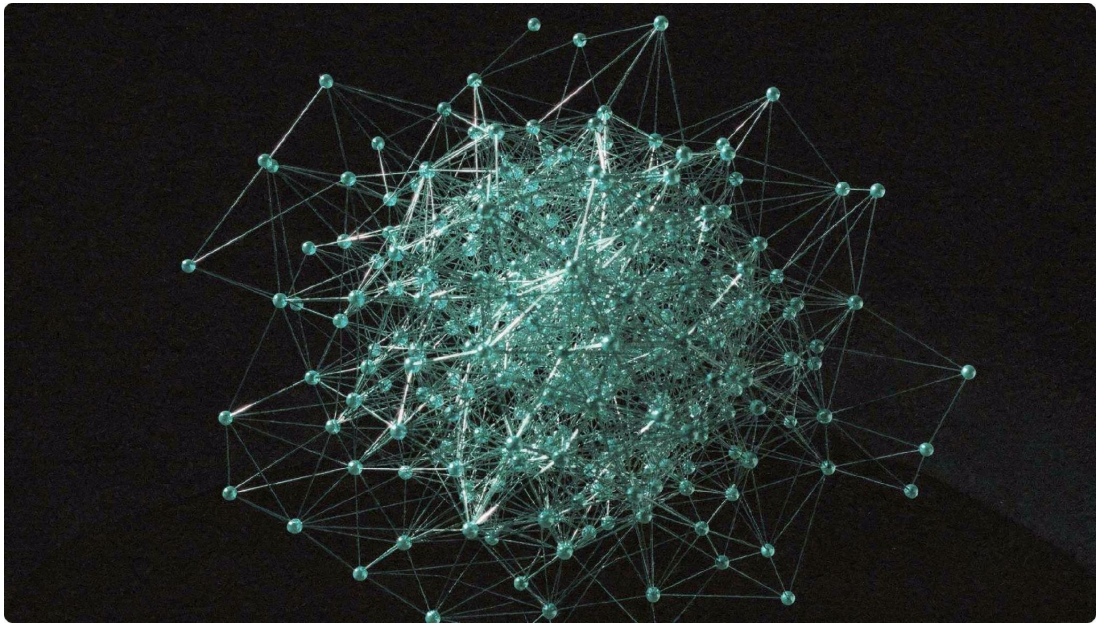
National Library of Korea, Seoul



From Legacy Metadata to Knowledge Infrastructures

MARC and TEI-XML encode extensive intellectual work — but much knowledge remains **implicit**: embedded in free text, local conventions, or relationships understandable to humans but not machines.

The transition to **Linked Open Data** reorganizes information around identifiable entities — agents, works, places, events — and the relationships connecting them.



The central question: How can AI make implicit knowledge explicit while preserving transparency, provenance, and cataloguing control?

Two Complementary Applications

This presentation explores two applications developed within entity-based and collaborative knowledge infrastructures, both following the same general principle: AI interprets controlled evidence and proposes structured semantic information without obscuring the origin or status of the resulting assertions.

Part I — **Tech4You**

Enrichment of sparse entity descriptions derived from TEI-XML documents within a cultural heritage knowledge graph.

Part II — **Contributor Roles**

Automatic identification of contributor roles from MARC 21 bibliographic records through a hybrid deterministic and AI-assisted pipeline.

PART I — TECH4YOU

From TEI-XML Documents to a Knowledge Graph

The project **Tech4You** preserves and disseminates cultural heritage of the Basilicata region.

Documentary sources encoded in TEI-XML are transformed into RDF via the **Share LOD Platform** customised within the **Aracne framework**, modelled with BIBFRAME, and loaded into a triplestore. Entity resolution via **fuzzy matching** and **semantic similarity** groups equivalent occurrences.

Entities are linked to VIAF, Wikidata, and GeoNames — turning encoded documents into a navigable, interoperable knowledge graph.

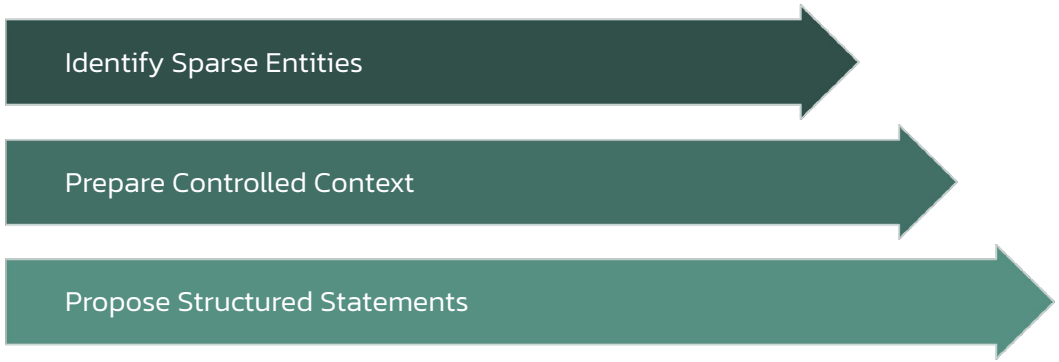
PART I — TECH4YOU

Enriching Incomplete Entities with AI Agents

TEI-XML conversion often yields sparse entity descriptions — a label, a type, a document link. Technically valid, but insufficient for discovery or disambiguation.

The AI enrichment process identifies entities below **completeness thresholds** and prepares a controlled context: mentions, existing attributes, candidate identifiers, and related evidence from the knowledge base.

The agent proposes additional statements — descriptions, attributes, relationships — acting as an **interpretative layer** between textual evidence and structured metadata.



PART I — TECH4YOU

Provenance and Human Control

Full Traceability

Every AI-generated statement records its source, evidence, process, date, model, and validation status.

Separation from Human Metadata

Machine-generated statements are never silently merged with human-authored records.

Reversibility

Enrichments can be reviewed, accepted, corrected, or rejected — keeping cataloguers in authority.

The result is enrichment that is **explainable, auditable,** and **reversible**.

PART II — CONTRIBUTOR ROLES

The Problem of Implicit Roles in MARC Records

In MARC 21, contributor roles should appear in field 700 subfield $4 as relator codes (*aut, edt, trl, ill*).

In many legacy records, the code is **missing** — the role is expressed only in the free-text statement of responsibility (field 245 $c): *"translated by Mario Rossi; edited by Anna Bianchi."*

A human cataloguer interprets this easily; a machine cannot without analysis.

In BIBFRAME, explicit role representation is essential for semantic precision, disambiguation, and interoperability.

A Hybrid and Conservative Approach

Principle: rules first, AI second — abstain rather than guess.

What System Reads

Primary Evidence

245 $c — Statement of responsibility (segmented by ISBD punctuation)

700 — Contributor names to resolve

100 — Main author (when present — triggers names-only rule)

Contextual Evidence

Leader — Resource type, bibliographic level

336/337/338 — Content, media, carrier types

655 — Genre/form (film, documentary, essays...)

508 — Credits note ("Director, Colin Low")

511 — Participant note ("Narrated by D. Attenborough")

505/520 — Contents and summary notes

The system collects structured evidence before reasoning — not just free text.

Documenting How the Role Was Assigned

Lookup

Explicit expression matched unambiguously to LC vocabulary. E.g., "translated by" → trl. Also includes names_only_rule (names without role → aut) and extraction from fields 508/511.

Rule-based disambiguation

Multiple candidates from vocabulary; contextual rules (resource type, genre) select the correct code. E.g., "produced by" in a film → pro, not fmp.

Inference

No direct lexical match. LLM interprets full bibliographic context: resource type, genres, credits, content notes, subject names. Includes not_determinable when evidence is insufficient.

Not Determinable

Not a separate method — it is the LLM's verdict when evidence is insufficient. A deliberate reliability feature: better no answer than a wrong one.

PART II — EVALUATION

Evaluation Results — 724 MARC Records

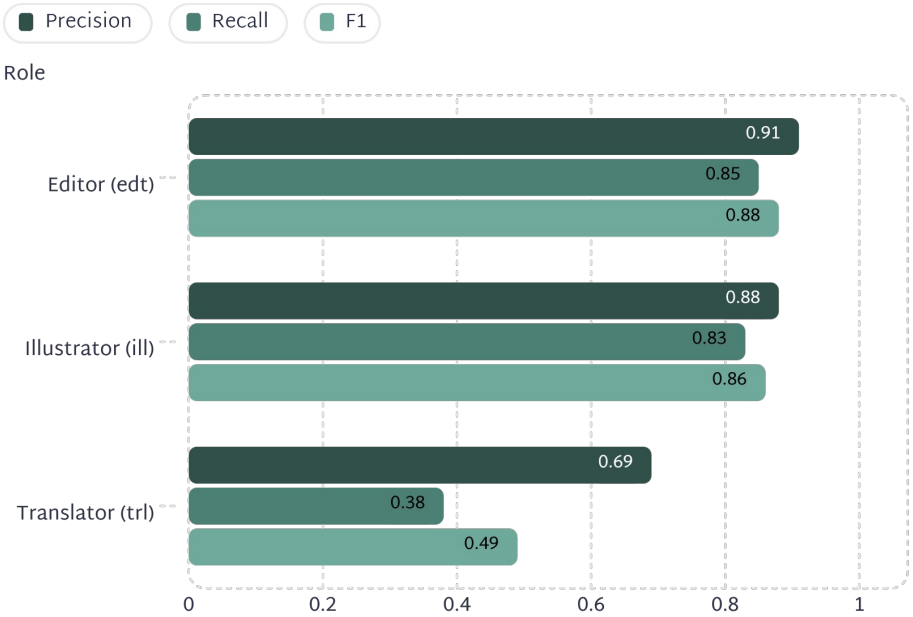
73%	0.67	54%
Overall Precision	Micro F1 Score	Abstention Rate
~3 in 4 proposed assignments were correct	Balanced precision and recall across all roles	The system abstains when evidence is insufficient — preventing unsupported assertions

The system correctly identified **898 of 1,764** ground-truth role assignments. High abstention reflects genuine gaps in source metadata — a cataloguer would also need external sources in these cases.

Abstention is part of the reliability model, not a weakness.

PART II — EVALUATION

Performance by Role



Editors and illustrators use **stable linguistic patterns** — the system excels here. Translator detection is harder: statements are more variable, indirect, or incomplete in legacy records.

The overall F1 of **0.67** represents a **+24 percentage point improvement** over a trivial baseline that always assigns the most frequent code.

PART II — EXAMPLES

Three Examples of Role Identification

1	2	3
Deterministic — "*edited by Ahmed Abdulkadir…*" Explicit phrase governs all 7 names. Rules assign *edt* to each. No LLM needed. Ideal: explicit evidence, high precision, low cost.	Contextual Inference — "Yuji Nawata, Hans Joachim Dethlefs (eds.)" The 245 $c names only two curators with "(eds.)". Eight additional 700 contributors are absent from the text. AI reasons by exclusion: if curators are explicitly marked, remaining contributors are individual essay authors. Assigns edt to 2, aut to 8. All 10 correct.	Semantic Disambiguation — "produced by Stefan Gravesen" "Produced by" is ambiguous. AI considers audiovisual resource type and selects the correct film-producer relator. All 3 roles correct. This example comes from an audiovisual subset evaluated separately.

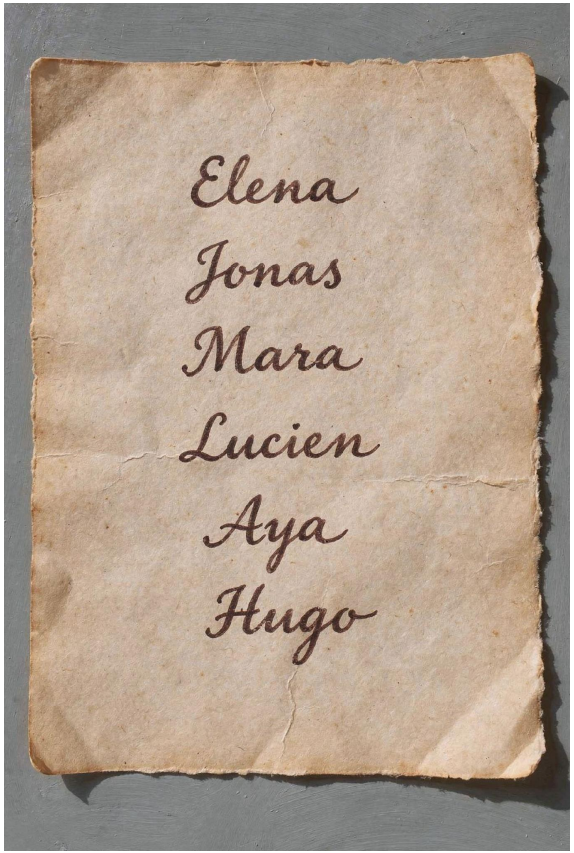
Limitations: Abstention

Correct Abstention

56% of missed assignments reflected missing source evidence, not detection errors.

Generic roles such as *ctb* or *oth* could not be reliably inferred without information external to the MARC record.

Limitations: Default-to-Author Bias



When field 245 $c contains only a list of names with no explicit role expression, the model tends to infer *aut* for all — correct for joint monographs, but wrong when editors or generic contributors are unlabelled.

In one case, the system assigned *aut* to 16 names; ground truth showed 2 editors and 14 generic contributors. This pattern generated **~74% of false positives** (242 incorrect assignments out of 329 total).

⚠ Fix: make the system more conservative — prefer *not_determinable* over unsupported defaults.

PART III — WORKFLOW

From Proof of Concept to Cataloguing Workflow

The role-detection system is not designed to replace the cataloguer. Its purpose is to **redistribute cataloguing effort**.

01

Batch Processing

Backlogs of records without $4 relator codes are processed automatically through scheduled operations.

02

Automatic Acceptance

Assignments based on explicit, unambiguous expressions are accepted automatically when institutional policy and validation thresholds permit.

03

Cataloguer Review Queue

Uncertain cases are presented with the relevant 245 $c segment, contributor name, candidate codes, proposed role, resolution method, confidence score, and evidence considered. The cataloguer receives a motivated proposal — not an empty field.

04

Isolation of Unresolved Cases

Cases marked `not_determinable` are directed to workflows requiring external research or direct examination of the resource.

Costs, Scalability, and Next Steps

Estimated Costs

Approximately US$5 per 1,000 records, at about 2 seconds per record.

The approach appears scalable for large backlogs, but production costs also include infrastructure, quality control, provenance, security, governance, and long-term maintenance.

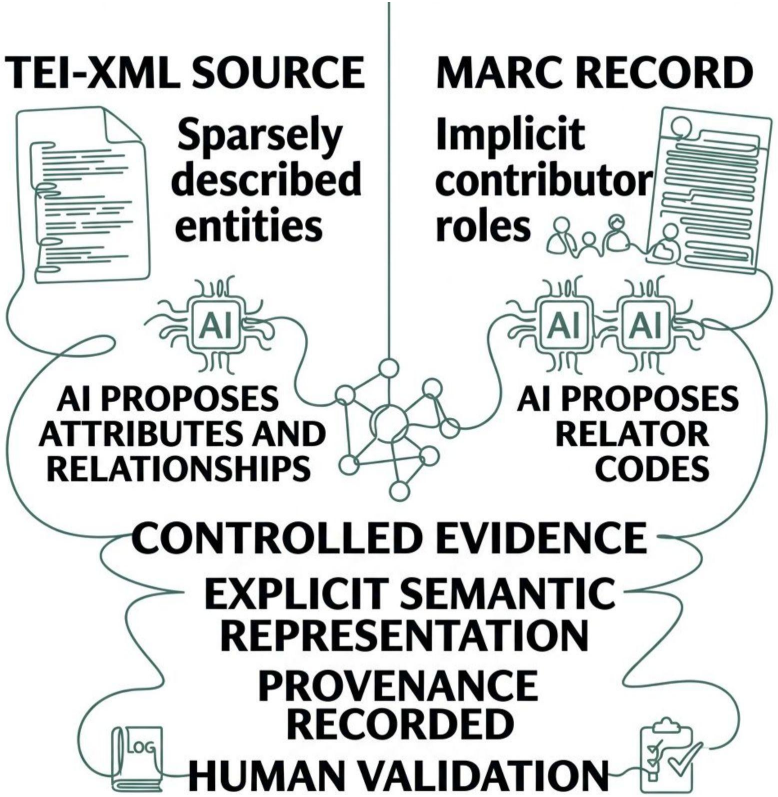
Future Development

Next steps

- Expand multilingual vocabularies and cross-institutional validation
- Improve testing, confidence calibration, and policy profiles
- Strengthen catalogue integration and feedback loops

Until validation thresholds are agreed, outputs should remain recommendations.

Two Sources, One Enrichment Principle

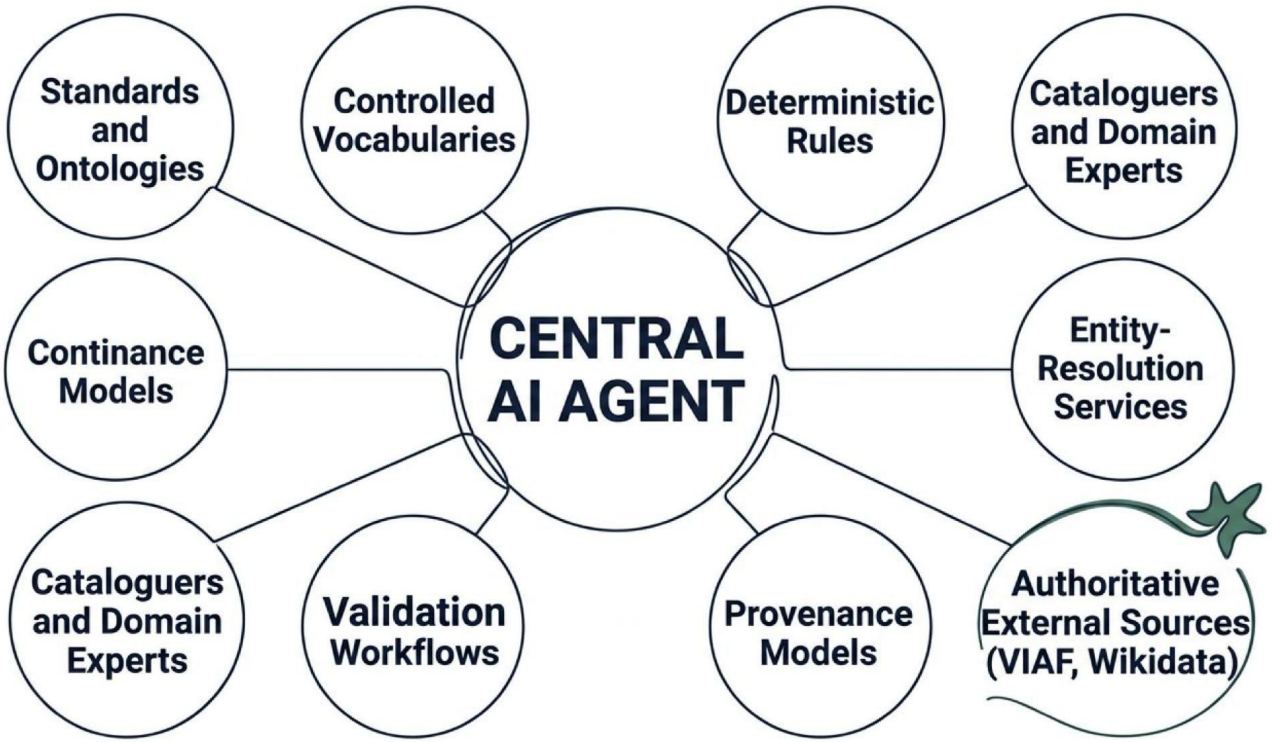


Both applications follow the same principle: AI receives **controlled evidence**, identifies implicit or incomplete information, proposes an explicit semantic representation, records how the proposal was generated, and allows human validation.

The transition in both cases is the same:

From textual evidence to explicit semantic relationships.

AI Agents Within a Governed Metadata Ecosystem



AI-Assisted Metadata Enrichment: Automation with Human Authority

Machines Excel At

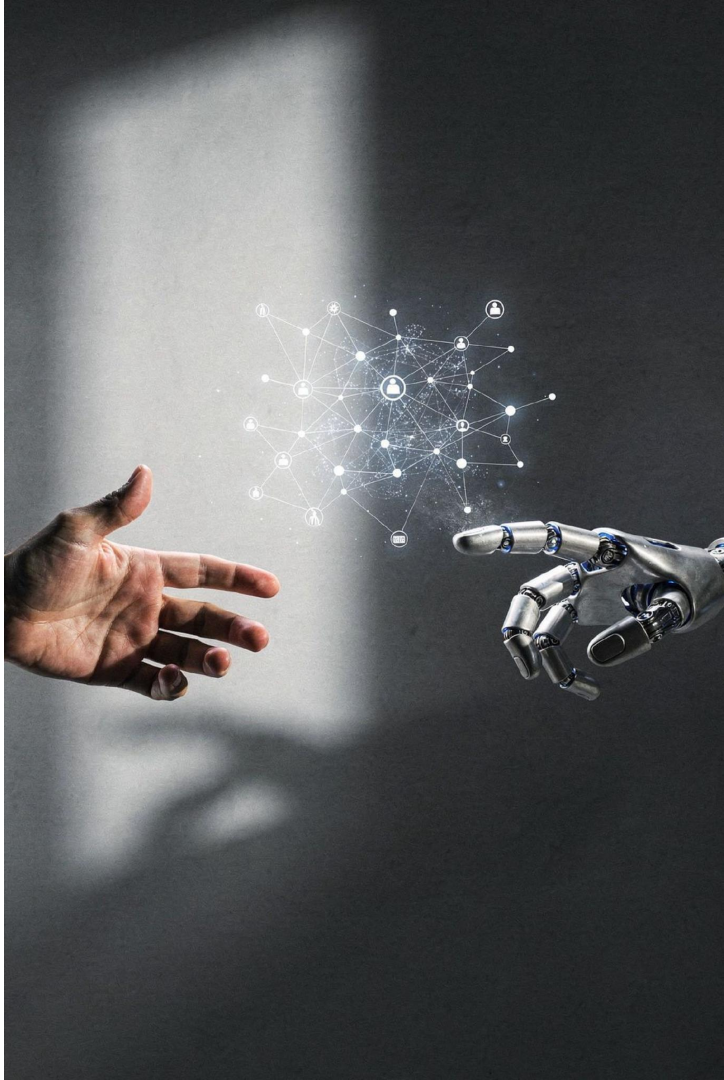
Processing large record sets, recognizing patterns, applying vocabularies, proposing interpretations, flagging uncertain cases.

Cataloguers Remain Essential For

Interpreting ambiguity, validating relationships, correcting biases, setting policy, governing the knowledge base.

The Governing Principle

Evidence preserved. Methods documented. Uncertainty represented. Machine statements identifiable. Human expertise retains authority.



Thank You!

Acknowledgements

With special thanks to **Matteo Morelli, Alessandra Moi**, and all colleagues from the **@Cult-Casalini Libri** and **Share Family** teams who contributed to the development, testing, and validation of this work.

Learn More

SHARE Family

share-family.org

SHARE Wiki

wiki.share-family.org

Contact Us

info@share-family.org