

From MARC to Linked Open Data: AI-Driven Entity Extraction from Hebrew Manuscript Metadata Using Distant Supervision

Alexander Goldberg¹  

¹Department of Information Science and Applied AI, Bar-Ilan University, Israel

Abstract

Cultural heritage institutions preserve invaluable provenance information in Machine-Readable Cataloging (MARC) records, yet much of this knowledge remains trapped in unstructured note fields, inaccessible to computational analysis. Transforming these legacy catalogs into Linked Open Data (LOD) requires extracting structured person-role relationships—identifying authors, scribes, owners, and censors—from cataloger narratives. This study presents an AI-driven system that uses *distant supervision from MARC metadata itself* to automatically generate training data, eliminating the prohibitive cost of manual annotation for specialized cultural heritage domains. By exploiting the dual structure of catalog records, where structured fields provide authoritative labels and unstructured notes provide context, we achieve 85.70% F1 for person extraction and 100% role classification accuracy, outperforming general Hebrew NER models by +55.55% F1. Our approach demonstrates how existing metadata can be leveraged to train AI systems that align with the values of cultural heritage preservation: accuracy, provenance tracking, and semantic enrichment. The extracted entities populate ontology instances based on CIDOC-CRM and IFLA-LRM, enabling computational analysis of scribal networks and manuscript circulation at scale.

1. Introduction

Hebrew manuscript catalogs document centuries of cultural transmission, preserving information about the persons who created, copied, owned, and regulated historical texts. The National Library of Israel maintains over 120,000 MARC records describing these manuscripts, containing rich provenance narratives that connect manuscripts to their human contexts. However, this valuable information remains largely inaccessible to computational analysis—trapped in free-text notes written by catalogers in natural language.

Our work demonstrates that *metadata itself can supervise AI training*, ensuring that automated systems align with the semantic precision that catalogers have invested in their records. Rather than imposing external annotation schemes, we leverage the implicit expertise already encoded in structured MARC fields to train models that understand domain-specific person-role relationships and preserve and reflect those values.

This approach addresses two critical gaps in current practice: (1) general-purpose NER systems fail on manuscript catalogs due to historical naming conventions and domain-specific terminology, and (2) supervised learning requires expensive manual annotation, prohibitive for under-resourced cultural heritage institutions. Our innovative distant supervision method eliminates annotation costs entirely while achieving high extraction accuracy.

Short Paper

Available Aug 01, 2026

Correspondence to
Alexander Goldberg
shvedbook@gmail.com



Copyright © 2026 Goldberg. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license, which enables reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator.

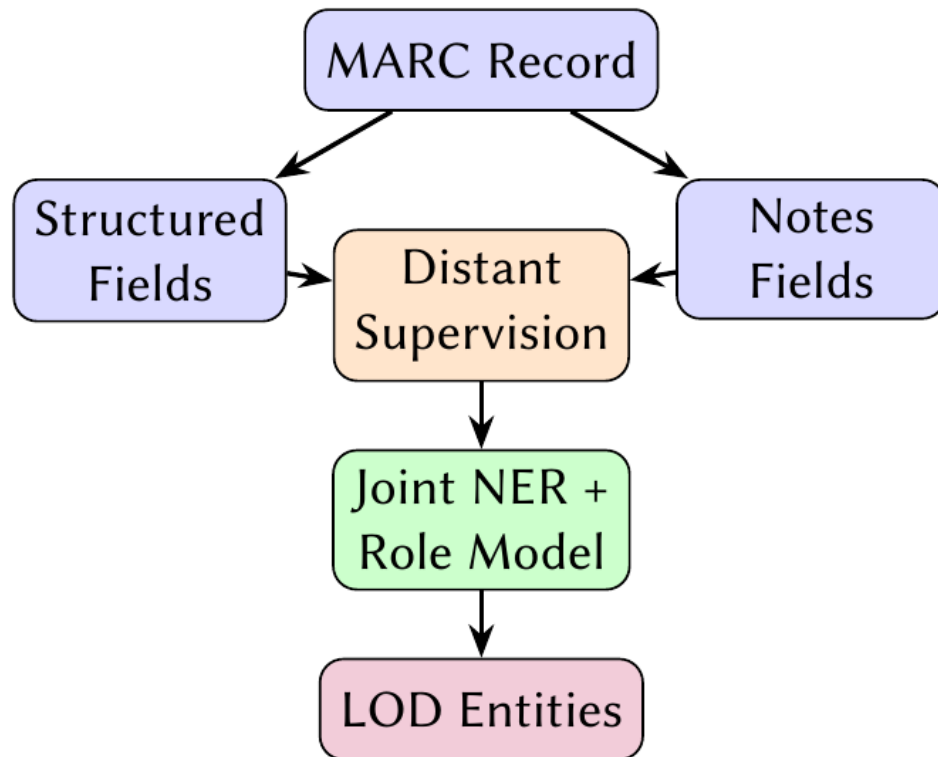


Figure 1. Pipeline: MARC dual structure enables automated training data generation for LOD entity extraction.

2. Methodology

2.1. Distant Supervision from MARC Metadata

Our key insight is that catalogers have already implicitly annotated their data. When a cataloger enters a person name in a structured MARC field (e.g., 100\$a for main entry, 700\$a for added entry) and describes that person in a notes field (e.g., 500\$a for general notes, 561\$a for provenance), they create a natural training example. The structured field provides the authoritative entity label; the notes field provides the natural language context.

From 123,621 MARC records from the National Library of Israel, we automatically generated 7,568 training samples by aligning names appearing in both structured and unstructured fields. For example, a structured access field with an owner role aligned with the same name in a notes field supplies both label and context. This approach captures domain-specific naming patterns (historical Hebrew names, patronymics, geographic epithets) that general NER systems miss.

2.2. Joint Entity-Role Model

We employ a multi-task learning architecture based on DictaBERT [1], a Hebrew language model, with dual classification heads: one for entity extraction (BIO tagging) and one for role classification (author, scribe, owner, censor, etc.). The architecture uses a shared encoder for entity-boundary and role predictions; the performance gain below is primarily from in-domain MARC supervision, not joint learning alone.

Table 1. External reference baselines. Results reflect in-domain distant supervision and role-aware training, not a joint-learning ablation.

Model	NER F1	Role Acc.
Base DictaBERT (zero-shot)	1.40%	51.84%
DictaBERT-NER (general Hebrew)	30.15%	0.0%
Proposed MARC-supervised model	85.70%	100.0%

3. Results

Using 5-fold cross-validation on our automatically generated dataset, evaluation proceeded in five rotations. In each rotation, a fresh model was fitted on all samples outside the current held-out fold and scored on that fold; scores were then averaged.

Because the baselines were not trained on the MARC distant-supervision data, the +55.55% F1 gap should be interpreted mainly as in-domain supervision, not joint learning alone. The role-aware head adds the semantic labels needed for LOD conversion.

4. LOD Integration and Impact

Extracted person-role relationships populate ontology instances based on CIDOC-CRM [2] and IFLA-LRM [3]. A manuscript transcription event becomes an F27 Work Creation instance linking a person (E21 Person), their functional role, and the manuscript (F4 Manifestation Singleton). This event-based modeling enables:

- **Provenance queries:** tracing ownership chains across collections.
- **Network analysis:** mapping scribal schools and master-apprentice relationships.
- **Temporal-spatial analysis:** visualizing manuscript production and circulation patterns.

This work is one component of a larger National Library of Israel project that publishes the full MARC metadata for Hebrew manuscripts as LOD; our contribution focuses on extracting people and their roles, providing essential enrichment for historical and cultural analysis.

5. Transferability and Limitations

Transferability depends on reliable structured authority fields, notes that repeat the same names, and consistent role coding. Collections with sparse notes, OCR noise, other languages, or fewer repeated names need local validation, so exact scores are collection-specific.

6. Conclusion

We demonstrate that MARC records' dual structure—authoritative structured fields paired with descriptive notes—provides an ideal source for distant supervision, eliminating manual annotation costs while achieving high extraction accuracy. This

paradigm extends to any cultural heritage domain with structured catalog metadata, offering a meaning-driven approach where AI systems are trained on, and aligned with, the values embedded in existing professional metadata practices.

Acknowledgments

We thank the National Library of Israel for providing access to their MARC catalog data.

References

- [1] A. Shmidman, S. Shmidman, M. Koppel, and Y. Goldberg, "DictaBERT: A state-of-the-art BERT model for modern Hebrew," 2023. doi: [10.48550/arXiv.2305.04096](https://doi.org/10.48550/arXiv.2305.04096).
- [2] M. Doerr, "The CIDOC conceptual reference model: An ontological approach to semantic interoperability of metadata," *AI Magazine*, vol. 24, no. 3, pp. 75–92, 2003, doi: [10.1609/aimag.v24i3.1720](https://doi.org/10.1609/aimag.v24i3.1720).
- [3] P. Riva, P. Le Boeuf, and M. Žumer, "IFLA library reference model: A conceptual model for bibliographic information," International Federation of Library Associations and Institutions, 2017. [Online]. Available: <https://repository.ifla.org/handle/20.500.14598/40>

UNCORRECTED PROOF