Schema-Driven Information Extraction from Medieval Latin Regesten: A Four-Way Evaluation of GLiNER2 for Ontology Population*

Luana Moraes Costa[1], Bärbel Kröger[1] and Christian Popp[1]

[1]Göttingen Academy of Sciences and Humanities in Lower Saxony, Germany

Abstract

The *Repertorium Germanicum* (RG), a critical edition of papal registers for the Holy Roman Empire (1378–1484), presents a paradigm case for automated knowledge extraction from historical sources: entries written in dense abbreviated medieval Latin, structured by a domain ontology under active development within the HisQu project (HisQu — Forschungsdateninfrastruktur Historische Quellen - Research Data Infrastructure „historical sources"). This paper evaluates GLiNER2—a schema-driven zero-shot information extraction framework built on a DeBERTa-v3 backbone—across four experimental configurations, varying model variant (large vs. multilingual) and preprocessing strategy (abbreviation expansion vs. raw text), for the task of populating the *Ablass* (papal indulgence) branch of this ontology. Our results reveal a fundamental tension between act-level detection, which both variants perform with reasonable coverage, and relation-level instantiation, which collapses across all configurations. We document a systematic subtype conflation bias, near-zero extraction of recipient church institutions despite explicit schema instructions, and entity boundary errors specific to abbreviated Latin tokenisation. Counterintuitively, abbreviation expansion degrades archival source reference extraction. We argue that the structural formula of the RG—not the abbreviated Latin per se—is the primary challenge, and conclude that generic zero-shot NER, while useful as a feasibility probe, is insufficient for the semantic depth required by ontology-based indexing of this source type.

Keywords

Information Extraction, GLiNER2, DeBERTa, Ontology Population, Medieval Latin, Repertorium Germanicum, Knowledge Graphs, Digital Humanities

1. Introduction

The construction of knowledge graphs from historical sources sits at the intersection of NLP, knowledge representation, and Digital Humanities [6, 16]. While automated ontology population has been an active research area for decades [5], its application to pre-modern, non-standardised documentary corpora remains largely unsolved [4]. Medieval administrative documents such as RG regests present this challenge acutely: their content is semantically structured—each record encodes a specific legal act with defined participants, institutions and dates within the framework of contemporary canon law—yet their surface form is radically compressed, formulaic, and language-specific in ways that defeat general-purpose NLP.

The *Repertorium Germanicum*, published by the German Historical Institute in Rome [18, 10], indexes the contents of papal registers from 1378 to 1484. The edition now comprises more than 200,000 regests that summarise, according to established editorial conventions, the legally relevant content of late medieval curial administrative activity. Each regest records one or more ecclesiastical acts in tightly abbreviated medieval Latin. A typical regest of an indulgence reads :

> *Aschersleve - eccl. domus fr. min. op. Halberstad. dioc.: indulg. ad instar eccl. s. Marie in Portiuncula 5 iul. 1401 L 94 199.* [1]

SemDH 2026: Third International Workshop on Semantic Digital Humanities, Co-located with ESWC 2026, June 2026, Dubrovnik, Croatia

*Work in progress. Pipeline code and output TTL files will be released as open resources under CC BY 4.0 at https://github.com/luanamoraescosta/GLINER2RG.

author@institution.de (L. M. Costa); baerbel.kroeger@adwgoe.de (B. Kröger); christian.popp@adwgoe.de (C. Popp)

[1]https://rg-online.dhi-roma.it/RG/2/645

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-06-11

This regest encodes: an institution that receives the indulgence (*Aschersleve - eccl. domus fr. min.* = Franciscan friary in Aschersleben), its diocese (*Halberstad. dioc.* = Diocese of Halberstadt), the type of papal grant (ad-instar Ablass = indulgence whose defining feature is its reference character: the grant is structured according to the pattern, scope, or privilege of another recognized indulgence or church), the model church (= the referenced church; in this case the chapel of Portiuncola, Santa Maria degli Angeli, Assisi), the date (5 July 1401), and the archival source reference (L = Archivio Segreto Vaticano, Registri Lateranensi, vol. 94, fol. 199). The domain ontology developed within the HisQu project formalises these elements into an OWL class hierarchy. The research question is whether current zero-shot NLP models can extract these elements and instantiate them as ontology individuals without RG-specific training data.

We contribute: (1) a systematic evaluation of GLiNER2 [1] across four experimental conditions; (2) an error taxonomy specific to abbreviated medieval Latin; (3) a schema design aligned 1:1 with the RG ontology; and (4) empirically grounded recommendations for hybrid extraction pipelines in Digital Humanities.

2. Background

2.1. Information Extraction and Ontology Population

Ontology population—instantiating a pre-existing schema with named individuals derived from a corpus—requires not only named entity recognition but precise semantic typing and relation extraction [5, 12]. For historical corpora, the task has been approached through pattern-based methods [14], conditional random fields [13], and transformer-based models [3, 4]. Named entity recognition in historical documents is dominated by three error sources: abbreviation handling, entity boundary detection, and domain shift [4, 9]. Our experiments confirm and extend these findings to the relation-extraction layer. Prior attempts to structure RG entries using grammar-based parsing (ANTLR) have yielded valuable insights, but have also shown the limits of rule-based approaches for this material [17], suggesting that the challenge is not specific to any single NLP paradigm.

Medieval Latin has received comparatively little NLP attention. Unlike classical Latin [7], medieval administrative Latin combines regularised formulaic structures with dense abbreviation systems and the absence of conventional sentence boundaries. The RG adds a further layer: its entries are not sentences but structured slots separated by punctuation conventions that carry structural meaning distinct from their prose usage [15].

2.2. GLiNER2: DeBERTa Backbone and Schema-Driven Extraction

GLiNER2 [1] is built on a DeBERTa-v3 encoder [2]. Unlike autoregressive LLMs, it uses a span-scoring mechanism that encodes input text and label descriptors jointly in a shared embedding space and scores candidate spans via a bilinear matching function. This performs *label-conditioned span classification*: ontological constraints and disambiguation instructions provided in natural language participate directly in scoring, meaning the schema itself influences extraction at inference time—the theoretical justification for using GLiNER2 as an ontology population tool.

DeBERTa's key architectural advance over BERT [3] is *disentangled attention*: token content and positional information are encoded in separate vectors and combined during attention computation. For abbreviated Latin—where a token such as *eccl.* (= church) carries different semantic weight depending on position (before colon: institution receiving an indulgence; after the string *ad instar*: a church serving as a point of reference or model)—this mechanism is theoretically advantageous: it can learn that positional context modulates semantic interpretation independently of lexical identity.

Three GLiNER2 API methods structure our pipeline. `classify_text()` assigns the subentry to an act-type label (Stage 1). `extract_json()` accepts a typed field schema with per-field natural language instructions and returns a structured record (Stage 2). This combination allows a single model to

perform classification, relation extraction, and semantic typing conditioned on an externally provided schema, without fine-tuning [1].

The multilingual variant (GLiNER2-multi) was trained on a corpus including Italian, Portuguese, Spanish, and French. The vocabulary overlap with medieval Latin is non-trivial(*ecclesia* → *chiesa/iglesia*, *monachus* → *monaco/monje*, *indulgentia* → *indulgenza/indulgencia*) , and Italian notarial and papal chancery documents share date formulas and register references with the RG tradition. The hypothesis that Romance-language training provides cross-lingual transfer to medieval Latin is grounded in distributional genre overlap [8], though untested on abbreviated administrative registers.

2.3. The RG Ontology and the HisQu Project

The RG domain ontology is being developed as part of HisQu[2], an interdisciplinary research project funded by the German Research Foundation (DFG) and conducted by the Göttingen Academy of Sciences and Humanities in Lower Saxony, Friedrich Schiller University Jena, the German Historical Institute in Rome (DHI), and the Gotha Research Centre at the University of Erfurt.

The experiments with the zero-shot NER system GLiNER should be understood as case studies within the research data infrastructure pursued by HisQu. HisQu aims at a domain-specific, ontology-based enrichment of historical source texts, in which heterogeneous and semi-structured data are integrated into an iterative, transparent, and reproducible digital workflow. Against this background, the objective of the GLiNER tests is not to directly model the complex RG ontology, but rather to evaluate the potential contribution of a generic zero-shot approach to such an infrastructure.

The RG ontology is under active development. It follows an event-based approach. It is therefore guided by the thematic focus of the Repertorium Germanicum, which is primarily concerned with documenting the so-called Gnadenerweise (grants issued in response to petitions to the papal curia). In total, more than thirty distinct types of such papal acts of grace can be distinguished, one of which is the indulgence. The present paper focuses on the ontology branch for the ecclesiastical indulgences, as it is one of its more fully elaborated branches. The classes belonging to the *Ablass* branch are distinguished by controlled abbreviation markers encoded as *stringInRG* ontology annotations: *Plenarablass* (*plen. indulg.*), *Ad-instar-Ablass* (*indulg. ad instar*), *Jubilaeumsablass* (*indulg. iubilei*). Key relational properties include *hat_Ablassempfaenger* (recipient institution), *hat_Vorbild* (model church), *Dauer_gewaehrt* (duration granted), and *hat_Fundstelle* (archival source reference).

3. Experimental Design

3.1. Corpus and Pipeline

As source material, we use a JSON file comprising the complete corpus of all RG records. Due to the size and complexity of the RG, we decided to work with a small subset of the overall corpus concerning ecclesiastical indulgences. Entries relating to this topic were semi-automatically extracted for five dioceses. A domain expert in our team (historian) verified that each selected regest centrally concerns the granting of an indulgence. The resulting dataset consists of 167 regests. The pipeline uses a two-stage architecture: classify_text() determines the *Gnadenerweis* types that are mentioned in the regests (Stage 1); extract_json() with a 17-field schema aligned 1:1 with the ontology extracts slot values for the indulgences (Stage 2). Non-Ablass entries receive minimal triples with the appropriate *Gnadenerweis* subclass via *hat_Weitere_Aktivitaet*. Output is serialised to OWL/Turtle following the ontology's individual naming conventions.

3.2. Four Experimental Conditions

We vary two independent factors: model variant and preprocessing strategy (Table 1). Abbreviation expansion substitutes each Latin abbreviation with all possible expansions from the abbreviation list

[2] https://adw-goe.de/germania-sacra/hisqu

provided by the editors of the RG [3] (733 entries, 191 ambiguous), producing expanded text such as *eccl. [ecclesia] domus fr. [frater] min. [minor / minutus].*

Table 1
Experimental configurations.

Label	Model	Preprocessing
O1	GLiNER2-large	Abbreviation expansion
O2	GLiNER2-multi	Abbreviation expansion
O3	GLiNER2-multi	Raw abbreviated text
O4	GLiNER2-large	Raw abbreviated text

3.3. Evaluation

Evaluation at this stage is structural and qualitative, we analyse the class distribution of extracted individuals against the ontology's expected hierarchy and identify systematic error patterns through manual inspection.

4. Results

4.1. Quantitative Overview

Table 2 presents the class distribution across all four outputs. Semantic divergence concentrates in the extraction layer, particularly in the properties most critical for ontology population (bold rows).

4.2. The Ad-instar Conflation Bias

The most striking finding is the near-total collapse of Ablass subtype discrimination. Across all four configurations, *Plenarablass* and *Jubilaeumsablass* are never instantiated. Both model variants converge on *Ad-instar-Ablass* as the dominant subtype (53–134 instances). O4 (large-raw) collapses entirely to Ad-instar (97 instances, 1 generic Ablass).

Inspection reveals the mechanism: the model treats any Latin phrase following *indulg.* as evidence of an ad-instar reference. Thus *indulg. 5 dec. 1400* (a plain indulgence with a date) is misread as *indulg. [something] [something]* and classified as ad-instar. The actual discriminating signal—the phrase *ad instar* followed by a named church—is low-frequency and is not privileged over the dominant class prior [11].

4.3. The Kirchliche_Institution Collapse

The *hat_Ablassempfaenger* relation—identifying which church received the indulgence—is extracted with near-zero recall across all configurations (1–7 instances out of 97–149 central acts). Despite explicit schema instructions, the model either omits the field, maps institution names to *Ort* (place), or constructs malformed individuals concatenating institution name, abbreviations, and diocese. O3 produces only 1 *hat_Ablassempfaenger* triple across 138 central acts. The institution-level structure that makes the RG valuable for studying the geography of the papal indulgence system is essentially invisible to the model, consistently across all four configurations.

4.4. The Ort Explosion and Entity Boundary Failure

Ort is the most populated class across all runs (81–126 individuals), yet URIs such as `wp:Beygendorpe_Razeburg_Razeburgensis_dioc_diocesis_Ort` show the model ingesting

[3] https://rg-online.dhi-roma.it/denqRG/index.php?view=doc_abkuerzungen_layout

Table 2
Class and property distribution across four experimental configurations. Bold: most critical properties for ontology population.

OWL Class / Property	O1	O2	O3	O4
Sublemma_RG (total)	167	173	168	167
hat_Zentrale_Aktivitaet	114	149	138	97
hat_Weitere_Aktivitaet	53	24	30	70
Ablass (generic)	63	21	3	1
Ad-instar-Ablass	**53**	**126**	**134**	**97**
Plenarablass	0	0	0	0
Jubilaeumsablass	0	0	0	0
Kirchliche_Institution	**5**	**2**	**3**	**10**
hat_Ablassempfaenger	**7**	**2**	**1**	**6**
hat_Vorbild	0	0	4	4
Fundstelle	**0**	**22**	**44**	**2**
Dioezese	15	8	5	7
Dauer	18	23	20	21
Petent	12	54	59	21
Person_explizit	4	25	24	4
Ort	93	126	114	81
Gnadenerweis (other acts)	43	15	15	66
Dispens	5	2	10	2

entire multi-token sequences—toponym, diocese abbreviation, title—as a single place individual. The original regest reads:

> *Beygendorpe Razeburg. dioc. - par. eccl. in B.: de indulg. 12. sept. 1437 S 340 149r.* [4]

Rather than isolating the place name Beygendorpe, the model appears to absorb the broader phrase, conflating locative and ecclesiastical information within one entity. This problem seems particularly pronounced in abbreviated Latin, where punctuation marks abbreviation rather than syntactic or token boundaries. [9]. In this setting, boundary detection is made difficult not simply by the presence of dots, but more generally by unconventional tokenisation and dense contextual packaging. Abbreviation expansion aggravates this for the large model (93 vs. 81 *Ort*), as it produces longer noun phrases and introduces new ambiguities (*eccl.* expanded to *[ecclesia]* may be treated as a separate named entity rather than a structural marker).

4.5. Fundstelle (Source) Extraction: Where Multi-Raw Wins

The archival source reference (*Fundstelle*) shows the sharpest cross-configuration contrast: O3 (multi-raw) extracts 44 correct instances with well-formed URIs (e.g., `wp:10_dec_1391_L_24_242v`), while O1 (large+expansion) produces 0. The source reference follows a highly regular formula—[date] [register-letter] [volume] [folio]—that is legible in raw abbreviated text and disrupted by abbreviation expansion, which breaks the token sequence with parenthetical month expansions (*iul. [iulius]*). One possible explanation is that the multilingual model is better able to generalise over formulaic archival reference patterns, perhaps due to broader exposure to similar conventions in its pretraining data.

4.6. The Petent Confusion

The multilingual variant produces more *Petent* individuals (54–59 vs. 12–21), but inspection suggests systematic difficulties in identifying this role.

[4] https://rg-online.dhi-roma.it/RG/5/678

In the RG context, a petitioner is not necessarily an individual person: petitioners may equally be clerics, nobles, convents, communities, or ecclesiastical institutions—any entity submitting a request to the papal curia. This role must be distinguished from that of the indulgence recipient (*hat_Ablassempfaenger*), since petitioner and beneficiary institution may coincide, but often do not. The analysed regests, however, do not follow a single syntactic pattern. In simpler entries, a place name heads the record and the receiving institution follows directly; in others, a procedural layer introduced by the colon marks the underlying petition (*supplic.*); in still others, persons or families occupy the header position while the beneficiary church appears only later in a *pro* construction. The same structural slot—the pre-colon header—can thus be filled by a place name, an institution, a person with titles, or a combination thereof.

This variability appears to be a primary source of confusion for both model variants. The multilingual model, trained on text where grammatical subjects are typically persons, tends to classify header-position entities as petitioners even where the header contains a place name or institutional designation. The large model shows the inverse tendency, absorbing header strings—including those containing personal names and titles—into the *Ort* class. Neither variant reliably captures the structural distinction between petitioner and beneficiary that is encoded in the regest formula but not overtly marked by consistent surface-level cues.

5. Discussion

5.1. Act Detection vs. Relation Instantiation

The results show a systematic asymmetry. At the act level—determining that a subentry contains an *Ablass* rather than a *Dispens* or provision—both models achieve reasonable coverage (97–149 *hat_Zentrale_Aktivitaet* triples out of 167 sublemmas). At the relation level, however, the model fails to instantiate the structure of the act. This matters critically: a knowledge graph populated only with act nodes supports only impoverished queries. Answering "which churches in the diocese of Halberstadt received ad-instar indulgences modelled on Portiuncula?" requires populated *hat_Ablassempfaenger*, *hat_Dioezese*, and *hat_Vorbild* simultaneously.

5.2. The Abbreviation Expansion Paradox

Abbreviation expansion was designed to reduce input opacity. The results are counterintuitive: it is harmful for *Fundstelle* extraction, marginal or neutral for entity boundary detection, and irrelevant for subtype discrimination. The abbreviations in the RG are drawn from a controlled vocabulary that is also encoded in the ontology itself through stringInRG annotations—a model with Romance-language exposure may partially decode them without expansion. Structural slot-boundary markers (tagging the colon as institution/act separator) may be more effective than lexical expansion.

5.3. Does Romance-Language Training Help?

The hypothesis receives mixed support. GLiNER2-multi outperforms on *Fundstelle* (22–44 vs. 0–2) and *Person_explizit* (24–25 vs. 4). This may be consistent with overlap with Italian archival genres; the advantage could also reflect broader multilingual pretraining or more robust pattern recognition [8]. However, it shows worse subtype discrimination and worse *Petent* precision. The net effect is a precision/recall trade-off: more entity recall, lower semantic precision.

5.4. Structural Formula as the Primary Challenge

The consistent failure to populate *hat_Ablassempfaenger* across all configurations suggests the core problem is structural, not linguistic. The RG formula [institution] [diocese]: *indulg.* [subtype] [model church] [date] [source] violates assumptions of text models trained on natural language: no conventional verbs, institutional rather than personal grammatical subjects, and punctuation conventions carrying

structural meaning distinct from prose [15]. Explicitly labelling structural slots—using the colon as a hard boundary, date patterns as terminal anchors, and source reference formulas as closing signals—would transform the problem from span detection to semantic typing of pre-segmented spans, which should be substantially more tractable.

6. Conclusion

This work-in-progress evaluates GLiNER2 across four configurations for populating the HisQu RG ontology. Schema-driven extraction with DeBERTa is viable for act-level detection but insufficient for relation-level instantiation. Key findings: (1) a systematic Ad-instar conflation bias in all configurations; (2) near-zero *hat_Ablassempfaenger* recall; (3) entity boundary failures from abbreviated Latin tokenisation; (4) a negative effect of abbreviation expansion on source reference extraction; (5) a multilingual/large trade-off between entity recall and semantic precision.

The experiment shows that generic zero-shot NER systems, within the RG context, provide only a heuristic pre-annotation and are not capable of capturing the semantic depth required for ontology-based indexing. In particular, the structural formula of the RG regests—where the same syntactic position can encode places, institutions, or persons, and where punctuation carries structural rather than syntactic meaning—defeats the assumptions of models trained on conventional prose. The GLiNER tests should therefore be interpreted as controlled feasibility studies aimed at identifying the limits of such methods for this type of source material. Within the broader HisQu infrastructure, where domain-specific data modelling workflows already achieve substantially more precise results, the zero-shot approach evaluated here does not offer a viable path for production-level ontology population. All pipeline code, schema definitions, and output TTL files will be released as open resources.

Acknowledgments

The authors thank the Göttingen Digital Academy for supporting this project.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) to assist with pipeline code scaffolding and text drafting. All content was reviewed, edited, and verified by the authors, who take full responsibility for the publication's content.

References

[1] U. Zaratiana, G. Pasternak, O. Boyd, G. Hurn-Maloney, A. Lewis. GLiNER2: An Efficient Multi-Task Information Extraction System with Schema-Driven Interface. *arXiv:2507.18546*, 2025.

[2] P. He, J. Gao, W. Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv:2111.09543*, 2021.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL*, pp. 4171–4186, 2019.

[4] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, A. Doucet. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, 56(2):1–47, 2023.

[5] P. Cimiano, J. Völker. text2onto: A Framework for Ontology Learning and Data-Driven Change Discovery. In *Proc. NLDB*, pp. 227–238, 2005.

[6] J. Flanders, F. Jannidis (eds.). *The Shape of Data in the Digital Humanities.* Routledge, 2019.

[7] D. Bamman, P. J. Burns. Latin BERT: A Contextual Language Model for Classical Philology. *arXiv:2009.10053*, 2020.

[8] T. Pires, E. Schlinger, D. Garrette. How Multilingual Is Multilingual BERT? In *Proc. ACL*, pp. 4996–5001, 2019.

[9] A. Hamdi, E. L. Pontes, M. Ehrmann, A. Doucet. A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers. In *Proc. ACL-Findings*, 2021.

[10] J. Hörnschemeyer. Repertorium Germanicum online. In M. Matheus (ed.), *Friedensnobelpreis und historische Grundlagenforschung*, pp. 605–615. De Gruyter, Berlin, 2012. https://doi.org/10.1515/9783110259551.605.

[11] J. Li, A. Sun, J. Han, C. Li. A Survey on Deep Learning for Named Entity Recognition. *IEEE Trans. Knowledge and Data Engineering*, 34(1):50–70, 2020.

[12] N. F. Noy, D. L. McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report, 2001.

[13] J. Lafferty, A. McCallum, F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. ICML*, pp. 282–289, 2001.

[14] C. Grover, S. Dingare, J. Herd, M. Nissim, M. Toolan. Named Entity Recognition for Digitised Historical Texts. In *Proc. LREC*, 2004.

[15] L. Burnard. *What Is the Text Encoding Initiative?* Open Edition Press, 2014.

[16] L. Ehrlinger, W. Wöß. Towards a Definition of Knowledge Graphs. In *Proc. SEMANTiCS*, 2016.

[17] P. Stahl, D. Motz, J. Mitschunas, C. Beckstein. Paredros: Eine interaktive Entwicklungsumgebung zur grammatikbasierten Analyse semi-strukturierter Quellen. Zenodo, 2026. https://doi.org/10.5281/zenodo.18703014.

[18] Deutsches Historisches Institut Rom (ed.). *Repertorium Germanicum. Verzeichnis der in den päpstlichen Registern und Kameralakten vorkommenden Personen, Kirchen und Orte des Deutschen Reiches, seiner Diözesen und Territorien.* Niemeyer / De Gruyter, Berlin/Tübingen, 1916–2000.