

BibNER-Werk: Bibliographic NER for Historical German Catalogs

Mary Ann Tan
FIZ Karlsruhe – Leibniz Institute for
Information Infrastructure
Eggenstein-Leopoldshafen, Germany
Karlsruhe Institute of Technology
Karlsruhe, Germany
ann.tan@fiz-karlsruhe.de

Genet Asefa Gesese
FIZ Karlsruhe – Leibniz Institute for
Information Infrastructure
Eggenstein-Leopoldshafen, Germany
Karlsruhe Institute of Technology
Karlsruhe, Germany
genet-asefa.gesese@fiz-karlsruhe.de

Harald Sack
FIZ Karlsruhe – Leibniz Institute for
Information Infrastructure
Eggenstein-Leopoldshafen, Germany
Karlsruhe Institute of Technology
Karlsruhe, Germany
harald.sack@fiz-karlsruhe.de

Abstract

General-purpose named entity recognition does not identify the title and responsible person needed to link historical German bibliographic catalog strings in the Deutsche Digitale Bibliothek (DDB) to GND¹. Work authorities, since the task's label set and centuries-spanning historical register have no counterpart in standard training corpora. This paper uses ISBD punctuation as a tiered silver labeling signal and evaluates a deterministic rule-based baseline alongside zero-shot, few-shot, and fine-tuned GLiNER-multi and fine-tuned XLM-R-large on a stratified, human-annotated gold set. The rule-based baseline beats both fine-tuned systems on untrained data; on trained data, only GLiNER-multi's fine-tuning gains a real, confirmed precision advantage, and XLM-R-large's never surpasses the rule.

Keywords

named entity recognition, weak supervision, bibliographic metadata, digital libraries, entity linking, FRBR

ACM Reference Format:

Mary Ann Tan, Genet Asefa Gesese, and Harald Sack. 2026. BibNER-Werk: Bibliographic NER for Historical German Catalogs. In *The 2026 ACM/IEEE Joint Conference on Digital Libraries (JCDL '26)*, October 13–16, 2026, Frisco, TX, USA. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3805696.3846060>

1 Introduction

Bibliographic catalog strings in the Deutsche Digitale Bibliothek (DDB)² are semi-structured: `dc:title` is the field present in nearly every record, but its values range from a bare title to an International Standard Bibliographic Description (ISBD) [11]-formatted string packing subtitle, statement of responsibility (SoR), and imprint into the same field, and `dc:creator` is frequently absent [21]. Identifying the Work a catalog record describes, the prerequisite for querying a Work authority entry in GND (Gemeinsame Normdatei), Germany's integrated authority file, requires a title and its responsible person, and neither field supports this reliably.

¹GND, https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html

²DDB, <https://www.deutsche-digitale-bibliothek.de>



This work is licensed under a Creative Commons Attribution 4.0 International License. *JCDL '26, Frisco, TX, USA*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2597-5/2026/10

<https://doi.org/10.1145/3805696.3846060>

General-purpose named entity recognition is not applicable to this task, and for three independent reasons: the target label set (TITLE, OTHER_TITLE, and PERSON, defined for Work-level identification) has no counterpart in standard training corpora such as CoNLL [22] or OntoNotes [18]; the domain is GLAM bibliographic metadata rather than newswire or web text; and the register alternates between ISBD's formulaic punctuation grammar and unstructured, historically variable prose, including Early Modern orthography predating 1750. Even GLiNER-multi, a model purpose-built for zero-shot span extraction against arbitrary label descriptions, achieves only $F_1 = 0.087$ on this label set out of the box, evidence that the difficulty is intrinsic to the label set and domain rather than an artifact of an ill-suited off-the-shelf model; supplementing its zero-shot labels with in-context example spans, a natural remediation short of fine-tuning, does not close the gap, failing for every label at every tested k (Section 6).

The same punctuation that defeats general NER is, where present, a deterministic segmentation signal: a colon (":") separates title from subtitle, and a slash ("/") introduces an SoR: for example, "Die Befreyung von Theben : Ein Trauerspiel in fünf Aufzügen / Christian Felix Weiße" segments into TITLE (*Die Befreyung von Theben*), OTHER_TITLE (*Ein Trauerspiel in fünf Aufzügen*), and PERSON (*Christian Felix Weiße*). The silver labeling pipeline presented here treats that grammar as a labeling source rather than a nuisance, with confidence tiered by pattern completeness. The signal is corroborated enough to make a record eligible for silver training in only 7.56% of records (Section 3); whether existing systems trained on that subset nonetheless generalize to the rest, which falls below the promotion threshold without necessarily lacking punctuation entirely, is evaluated directly in Section 6.

This paper makes two contributions:

- (1) A training-free ISBD rule-based baseline that matches or exceeds fine-tuning on the untrained majority of the corpus (92.44%), with fine-tuning's benefit on trained records narrower than assumed (Section 6).
- (2) An ISBD-guided silver labeling pipeline with tiered confidence, comparing zero-shot, few-shot, and fine-tuned GLiNER-multi against fine-tuned XLM-R-large and a rule-based baseline via k -fold cross-validation (Section 5).

2 Related Work

HIPE [9], the chief historical NER benchmark, reports a German strict-regime F1 ceiling of 0.703–0.794 under exact span match; its person/location/organization types (plus one dataset's citation-scoped literary-work type) have no counterpart in the label set used

here (TITLE, OTHER_TITLE, PERSON). The closest silver-labeling prior work, weakly supervised historical NER [20], differs on every relevant axis: newspaper prose, standard entity types, and general-purpose weak supervision, versus, in this paper, ISBD-punctuated catalog strings, FRBR Work-level labels, and a deterministic punctuation grammar.

Bibliographic NLP on catalog records typically operates downstream of field segmentation, clustering or linking already-separated fields at the FRBR Work level: Manguinhas et al. [14], for instance, presents a MARC FRBRization approach that assumes title and responsibility role are already segmented. BookReconciler [15] clusters Manifestations and Expressions into Works via external authority services, achieving near-perfect results on well-documented English works but performing substantially worse on multilingual contemporary fiction; like the FRBRization work above, it operates downstream of recovering TITLE, OTHER_TITLE, and PERSON spans from a single unstructured `dc:title` string. Prior work by the current authors applies NER and extractive question answering to DDB title strings on a 108K-record subset [21]. This paper extends that NER baseline into a full silver labeling and fine-tuning pipeline evaluated at corpus scale.

GLiNER [23] reformulates NER as span classification against free-text label descriptions; its multilingual checkpoint (`gliner_multi-v2.1`) underlies this paper’s zero-shot, few-shot, and fine-tuned conditions. Batista et al. [3] report a parallel finding elsewhere in structured text: fine-tuned transformers outperform LLM baselines on formulaic administrative text. LLMs have also been applied to historical NER [24] and bibliographic metadata enrichment [16]; this paper instead uses the smaller, non-generative GLiNER/XLM-R family, more practical to fine-tune and deploy at corpus scale.

3 Corpus and Task

The corpus is a reprocessed snapshot of the DDB’s library sector, comprising 9.2M catalog records tagged or confirmed as German, Middle High German, Low German, or Latin, contributed by several hundred distinct libraries, archives, and museums, so `dc:title` values reflect each provider’s own cataloging practice rather than a uniform style. A non-negligible share of these records, particularly pre-1900, may originate from retroconversion of scanned catalogs rather than direct keying, introducing OCR noise as a possible confound; Appendix A gives detail, and Section A.1 returns to this risk.

Title length varies sharply by publication era in a U-shaped pattern: longest in the earliest era, shortest in the 19th century, then rising again in the modern stratum, consistent with concise Early Modern conventions giving way to richer digital-born metadata. Appendix A gives the word-count methodology and the full per-era trend in Figure A.1. The corpus is stratified into four dated eras plus an undated residual, split at 1700 and 1900 rather than uniform century cuts. Appendix A gives the rationale for both boundaries and a language/era confound caveat on the resulting strata.

Each input to the task is a single `dc:title` string drawn from a DDB library record; the output is a set of character-offset span triples, each consisting of a start offset, an end offset, and a label. The ISBD colon and slash markers that make this segmentation possible, where present, are introduced with a worked example

previously. This punctuation grammar is the labeling signal the silver pipeline uses, and it is corroborated enough to promote only a minority of the corpus to training; the remainder, tier-0, falls below that promotion threshold without necessarily lacking punctuation entirely, as Section 6’s rule-based baseline demonstrates directly.

Records are assigned to one of three confidence tiers based on how much of this structure a per-title ISBD field-detection pass recovers, detailed in Section 4:

- **Tier-0** (92.44% of the corpus): neither the area separator nor enough detected fields; excluded from silver training.
- **Tier-1** (7.48%; 49,999 sampled for training): at least three detected fields, or an SoR marker co-occurring with a detected year; drawn as a stratified subsample balanced across the four eras above.
- **Tier-2** (0.08%; 2,384 validated for training): additionally requires the area separator, at least one detected Manifestation-level field (edition, place, publisher, year, or series), and an SoR validated to name a person rather than a corporate body, editor, or family collective.

Three labels are defined: TITLE (proper title, before `:` or `/`), OTHER_TITLE (subtitle or other title information, after `:`), and PERSON (SoR, after `/`); PERSON’s semantic scope is the broader SoR role, not just the `/`-heuristic used to generate its silver label (Section 4).

4 Silver and Gold Dataset

This section describes the two datasets used throughout the paper: a small, human-annotated gold set used to evaluate every system in Section 6, and the larger silver-labeled pool built directly from ISBD punctuation that the fine-tuned systems train on. The gold set is presented first; the silver pool used to train the fine-tuned systems (Section 5) follows.

Gold Annotation Set. A gold standard of 395 records was compiled from the DDB library sector, stratified by publication era and silver labeling tier to ensure coverage of the full range of orthographic and structural variation present in the corpus (composition detail in Appendix A, Table A.1). This stratification is deliberately non-proportional to the corpus’s own tier distribution (Section 3): tier-0 is 92.44% of the full corpus but only 45% of the gold set, since the gold set targets coverage of the rarer tier-1/tier-2 strata rather than corpus-representative proportions.

The size follows a Wilson interval sample-size calculation [4] targeting a ± 10 pp half-width on TITLE F1 per era, met in three of four eras but infeasible for PERSON given its low prevalence in the two historical eras, so PERSON is reported only as a secondary metric with wider, explicit CIs (derivation detail in Appendix A).

Three annotators, with backgrounds spanning computational linguistics, library science, and DDB metadata practice, independently labeled each record in Doccano [17], producing character-offset span triples in JSONL format consistent with the silver output (Section 4), restricted to the three Work-level entities defined in Section 3 (TITLE, OTHER_TITLE, PERSON). Fewer than five records per annotator were left unreviewed, covered by the adjudication pass (Appendix A); records with no labelable span were explicitly marked as negative judgments and retained in the data.

Inter-coder agreement and the reconciliation pipeline that produced the evaluated gold set (used in all evaluations in Section 6) are detailed in Appendix A. The evaluation scripts are released under Apache 2.0; the gold set and silver spans are released for research use, with per-record rights inherited from DDB’s provider-declared statements (Section 1).

Silver Labeling Pipeline. Tier-1 and tier-2 assignment (Section 3) uses a per-title ISBD field-detection pass: *structural* detection applies when the area separator (. -) is present, splitting the title into a title-area segment (searched for other-title/SoR markers) and a following segment (searched for the Manifestation-level fields); more conservative whole-string *heuristic* pattern matching applies otherwise. The combined tier-1/tier-2 training pool totals 52,383 records.

The : and / cues and the labels they trigger follow the definitions in Section 3. The SoR following / is a heterogeneous field in practice, not a clean personal-name slot (validation breakdown in Appendix A). The pipeline therefore classifies each candidate SoR with the same general-purpose German NER tagger used in the failure analysis (flair/ner-german-large [1], Section 6), restricting PERSON to tagger-identified persons; editor/family-collective statements are filtered separately by keyword before the tagger runs and left unlabeled (detail in Appendix A). Not every candidate ISBD-adjacent signal survived validation: a trailing period, present in 19.5% of the corpus, was tested as a possible area-close marker and rejected as unreliable (validation detail in Appendix A).

Each silver record is stored as a character-offset span-triple list over the original dc: title string, the same schema used by the gold annotation set (Section 4, whose 395 records are excluded here too) and consumed by both fine-tuned systems (format handling detail in Appendix A). Limitations of the pipeline, including a qualitative review of its date-contamination behavior and an OCR-noise risk, are discussed in Appendix A.

5 Experimental Setup

Systems. Five systems form the comparison, evaluated identically against the held-out gold set (Section 4). Table 1 reports four of them; few-shot is omitted there, since its tier-level F1 is uniformly zero (Section 6).

- **Rule-based:** a deterministic ISBD punctuation rule (rule_ baseline), training-free; isolates fine-tuning’s contribution over directly parsing the silver-labeling heuristics.
- **GLiNER-multi** (gliner_multi-v2.1), evaluated three ways, two in Table 1: **GLiNER-ZS** (zero-shot, the difficulty floor); few-shot (not tabulated, $k = 1, 3, 5$), the ceiling reachable without fine-tuning; and **GLiNER-FT**, fine-tuned on the silver pool (Section 4).
- **XLM-R-FT:** xlm-roberta-large fine-tuned (BIO), a supervised BIO sequence tagger on the same silver pool.

Rule-based applies the same heuristics directly at inference, using the raw / marker rather than the tagger-based disambiguation that builds the silver labels (Section 4). Both fine-tuned systems train on the same silver pool, excluding the gold set (Section 4); data-format handling differs between them (detail in Appendix A). Hyperparameters and training configuration follow the source implementations’ defaults, unchanged for this setting (detail in Appendix A).

Table 1: F1 by silver tier, mean across ten folds (Section 5); Rule-based needs no training, the other three train only on tier-1/tier-2. Bold marks each tier’s confirmed-best system (paired per-fold significance test); tier-2 has none. Precision/recall detail in Appendix A (Table A.3).

System	Tier-0	Tier-1	Tier-2
Rule-based	0.347	0.384	0.240
GLiNER-ZS	0.050	0.040	0.275
GLiNER-FT	0.249	0.332	0.262
XLM-R-FT	0.295	0.377	0.242

Evaluation Protocol. A prediction counts as correct only under exact span match (character boundaries and label both agreeing with gold), following HIPE-2022’s strict-regime convention [9]. F1 is reported per label and per era stratum with its support, never pooled into micro or macro averages, since either would misrepresent GND-linking utility: micro would be dominated by TITLE, present in nearly every record, while macro would weight rare PERSON spans equally with it. The tier-level comparison (Table 1; full precision/recall in Appendix A) also reports precision and recall, since these can diverge from F1 ranking in ways relevant to GND linking, where a false-positive link is costlier than a false-negative coverage gap.

The gold set (Section 4) is split into $k = 10$ stratified folds over the composite era $\times$ silver-tier key (tier-balanced), with $k = 10$ fixed by a timing pilot confirming both fine-tuned systems complete within a practical per-run budget. Fold assignment applies to evaluation only: both fine-tuned systems always exclude the entire gold set from training regardless of fold, so the ten per-system checkpoints differ only in training-time randomness, and each is scored only against its own assigned gold subset.³ Each system’s F1 is aggregated as mean $\pm$ std across the ten folds, per label and era (strata with fewer than five gold spans in a fold excluded rather than diluted), alongside a 95% bootstrap confidence interval [7] (1,000 resamples) for the same cell, separating fold-composition and training-run variance from sampling uncertainty. Table 1’s “confirmed-best” markers use a paired bootstrap test, uncorrected for multiple comparisons (methodology and caveat in Appendix A).

6 Results

Results proceed in four parts: zero-shot, few-shot, fine-tuned GLiNER-multi/XLM-R-large, and PERSON failure analysis, with Rule-based (Section 5) compared throughout. Per-label F1 pooled across era is summarized in Appendix A (Table A.2); Table 1 instead reports F1 by silver tier, the paper’s primary rule-vs-fine-tuning comparison (full precision/recall in Appendix A, Table A.3; per-era PERSON detail in Table A.4).

A paired per-fold significance check confirms Rule-based beats both fine-tuned systems on tier-0 and beats GLiNER-FT on tier-1 (XLM-R-FT ties it there); tier-2 has no confirmed winner among the four. GLiNER-FT still shows a confirmed precision advantage over the rule on both trained tiers, traded against a confirmed recall

³This differs from conventional k -fold cross-validation in a way discussed in Appendix A.

deficit at tier-1 only (full precision/recall in Appendix A, Table A.3); XLM-R-FT shows no such advantage at any tier.

Zero-Shot Results. GLiNER-multi, zero-shot with the bare label set, achieves an overall support-weighted F1 of 0.087 against the held-out gold set (Section 4); since the model is not retrained per fold, the ten-fold split (Section 5) applies only at scoring time, for consistency with the fine-tuned systems’ reporting. OTHER_TITLE is entirely absent (F1 = 0.000, zero recall in all 395 records) and TITLE is weak but not absent (0.039); PERSON is the exception, at 0.339 exceeding both fine-tuned systems’ PERSON scores (0.149 GLiNER-FT, 0.120 XLM-R-FT) in every era stratum, not only on average (Table A.4, Appendix A). Section 6 examines this reversal alongside the fine-tuned systems’ own PERSON failure modes.

Few-Shot Results. GLiNER-multi, evaluated with each label’s free-text description supplemented by k domain-specific example spans drawn from tier-1 silver (Section 5), achieves an overall support-weighted F1 of 0.000 for every label at every tested k (1, 3, 5): the model predicts no span, for any label, in any of the 395 gold records (implementation checks ruling out scoring-threshold and example-quality artifacts in Appendix A). This is a stronger finding than a simple few-shot-underperforms result: the zero-shot PERSON score of 0.339 (Section 6) already exceeds every few-shot condition tested, so supplementing the label with example spans removes a signal the bare label already carried rather than merely failing to add one.

Fine-Tuned Results. GLiNER-FT, fine-tuned on the silver pool (Section 4) and evaluated per Section 5, reaches an overall support-weighted F1 of 0.278 (pooled; full per-fold results in the released repository, Appendix A): TITLE strongest (0.362), PERSON weakest (0.149), with two of five era strata scoring exactly zero on samples too large to attribute to small-sample noise.

XLM-R-FT, fine-tuned as a BIO sequence tagger on the same silver pool and folds, reaches an overall F1 of 0.333, exceeding GLiNER-FT’s 0.278; the gain concentrates in TITLE (0.414 vs. 0.362) and OTHER_TITLE (0.324 vs. 0.203), while PERSON is instead weaker under XLM-R-FT (0.120) than GLiNER-FT (0.149), continuing rather than reversing the zero-shot-exceeds-fine-tuned pattern from Section 6. Section 6 returns to PERSON’s weakness across both fine-tuned architectures.

Failure Analysis. PERSON’s weakness admits two candidate explanations: a data-scarcity effect, since tier-2 supplies few training records with an explicit SoR (Section 3), or a label-definition mismatch, since the silver pipeline restricts PERSON to personal names while the gold guideline places no such restriction. The zero-shot comparison in Section 6 bears on which explanation is more likely: a pure scarcity account predicts fine-tuning helps a little or not at all, not that it underperforms an untrained zero-shot baseline in every era stratum, which points more toward the training signal itself being misleading than toward scarcity alone.

The scarcity account sharpens at the era level: the two zero-F1 strata, Pre-1700 and Unknown, are precisely the two eras with zero tier-2 gold records (Table A.1), a near-total absence of PERSON training signal there specifically.

The label-mismatch hypothesis was tested and ruled out, finding no systematic person-versus-organization confusion (test detail in Appendix A).

The false negatives instead separate into two patterns: a general recall failure spanning both short and long names, consistent with scarcity, and a pattern specific to a monarch named under a full titlature rather than a bare name (e.g. “*Georg der Dritte, von Gottes Gnaden König von Groß-Britannien [...]*”), substantially longer on average than a true positive (length comparison in Appendix A). An independent general-purpose German NER tagger (flair/ner-german-large [1]) recovers the short names but, for the titlature case, tags only the bare name, correctly separating the surrounding place names and institution as LOC/ORG, a labeling-convention effect distinct from scarcity.

The same analysis replicates under XLM-R-large (Section 6, sharing no architecture, loss function, or label-assignment mechanism with GLiNER-multi) within measurement noise (1 of 121 false negatives, 37.6 vs. 15.6 characters, identical missed spans) – evidence the scarcity and labeling-convention accounts are properties of the data and guideline, not of one fine-tuning procedure.

False positives were examined separately: both fine-tuned systems and the rule-based baseline (Section 5) share the same PERSON false-positive pattern: over-extension by a leading lowercase “von” (“by”), consistent with the guideline’s own boundary convention rather than any one system’s weakness, left for future correction (full breakdown, and PERSON’s complete blind spot at tier-0, in Appendix A).

7 Conclusion

This paper presents an ISBD-guided silver labeling pipeline for FRBR Work-level NER over historical German catalog strings, and evaluates a rule-based baseline against zero-shot GLiNER-multi and two independently fine-tuned architectures (GLiNER-multi, XLM-R-large) on a held-out, human-adjudicated gold set. The rule-based baseline outperforms both fine-tuned systems on untrained data; only GLiNER-FT gains a confirmed precision advantage on trained data, which XLM-R-FT never matches. Fine-tuning’s benefit over the rule is narrower than assumed, confined to one architecture and one axis (precision). The fine-tuned models’ overall F1, best under XLM-R-large, falls well short of HIPE’s F1 ceiling for a different, more mainstream entity set. A second finding concerns silver label quality: for TITLE/OTHER_TITLE, a qualitative review found the silver heuristic disagreeing with independent re-annotation on every sampled tier-1 record, yet fine-tuning still lifts both labels well over the zero-shot baseline since silver volume helps where the structural signal is reliable. For PERSON, by contrast, fine-tuning drives F1 below the zero-shot baseline in every era stratum under both architectures, best explained by silver label quality constraining PERSON specifically, though not independently confirmed. Three directions extend this work: a learned strategy for tier-0, which received no silver training data despite the rule performing reasonably there; a boundary-fix ablation for leading “von” and date/volume clauses; and decoder-based (LLM) models, not evaluated here (Section 2). In practice, the rule can be deployed broadly, reserving costlier fine-tuning for the tier-1/2 minority, where GLiNER-FT’s precision gain offsets GND-linking’s false-positive risk.

Acknowledgments

The authors are grateful to Frank Pöhlmann, Kunigunde Binder, and Maria Huvar, whose expertise and careful annotation work were essential to this study.

Declaration on Generative AI During the preparation of this work, the authors used Claude (Anthropic) for drafting and editing under substantial human review and editorial oversight.

CRedit authorship contribution statement. **Mary Ann Tan:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, review, & editing. **Genet Asefa Gesese:** Writing – review & editing. **Harald Sack:** Project administration, Supervision, Writing – review & editing.

References

- [1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1638–1649. <https://aclanthology.org/C18-1139/>
- [2] Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34, 4 (2008), 555–596.
- [3] Rodrigo Batista, Luis Filipe Cunha, Purificação Silvano, Nuno Guimarães, Alípio Jorge, Evelin Amorim, and Ricardo Campos. 2026. *MiNER: A Two-Stage Pipeline for Metadata Extraction from Municipal Meeting Minutes*. Springer Nature Switzerland, 427–435. doi:10.1007/978-3-032-21300-6_33
- [4] Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. 2001. Interval Estimation for a Binomial Proportion. *Statist. Sci.* 16, 2 (2001), 101–133. doi:10.1214/ss/1009213286
- [5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 8440–8451. doi:10.18653/v1/2020.acl-main.747
- [6] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The Hitchhiker’s Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 1383–1392. doi:10.18653/v1/P18-1128
- [7] Bradley Efron and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap* (1 ed.). Chapman and Hall/CRC, New York. doi:10.1201/9780429246593
- [8] Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named Entity Recognition and Classification in Historical Documents: A Survey. *Comput. Surveys* 56, 2 (Sept. 2023), 1–47. doi:10.1145/3604931
- [9] Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022. Extended Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. In *Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum (CEUR Workshop Proceedings, Vol. 3180)*, Guglielmo Faggioli, Nicola Ferro, Allan Hanbury, and Martin Potthast (Eds.). Bologna, Italy, 1038–1063. <https://ceur-ws.org/Vol-3180/>
- [10] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. doi:10.5281/zenodo.1212303
- [11] International Federation of Library Associations and Institutions. 2011. *ISBD: International Standard Bibliographic Description, Consolidated Edition*. Technical Report. IFLA. <https://repository.ifla.org/items/d4fd3e2c-8609-4738-8cfd-259432895813>
- [12] Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage, Beverly Hills, CA.
- [13] Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology* (2 ed.). Sage, Thousand Oaks, CA.
- [14] Hugo Miguel Álvaro Manguinhas, Nuno Miguel Antunes Freire, and José Luis Brinquete Borbinha. 2010. FRBRization of MARC records in multiple catalogs. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries (Gold Coast, Queensland, Australia) (JCDL '10)*. Association for Computing Machinery, New York, NY, USA, 225–234. doi:10.1145/1816123.1816157
- [15] Matthew Miller, Dan Sinykin, and Melanie Walsh. 2025. BookReconciler: An Open-Source Tool for Metadata Enrichment and Work-Level Clustering. In *2025 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 1–4. doi:10.1109/JCDL67857.2025.00046
- [16] Ashwin Nair, Ee Min Hoon, and Robin Dresel. 2024. Metadata Enrichment with Named Entity Recognition using GPT-4. In *Proceedings of the International Conference on Dublin Core and Metadata Applications (Dublin Core Conference, Vol. 2024)*. Dublin Core Metadata Initiative, Dublin, OH, USA. doi:10.23106/dcmi.952445840
- [17] Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text Annotation Tool for Human. Software available from <https://github.com/doccano/doccano>. <https://github.com/doccano/doccano>
- [18] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards Robust Linguistic Analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, Julia Hockenmaier and Sebastian Riedel (Eds.). Association for Computational Linguistics, Sofia, Bulgaria, 143–152. <https://aclanthology.org/W13-3516/>
- [19] Lance Ramshaw and Mitch Marcus. 1995. Text Chunking using Transformation-Based Learning. In *Third Workshop on Very Large Corpora*. <https://aclanthology.org/W95-0107/>
- [20] Marco Sorbi, Laurent Moccozet, and Stephane Marchand-Maillet. 2026. Weakly Supervised Named Entity Recognition for Historical Texts. In *Proceedings of the 10th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature 2026*. Association for Computational Linguistics, Rabat, Morocco, 48–65. doi:10.18653/v1/2026.latechclfl-1.6
- [21] Mary Ann Tan, Shufan Jiang, and Harald Sack. 2024. How to Turn Card Catalogs into LLM Fodder. In *Proceedings of the Workshop on Deep Learning and Linked Data (DLnLD) @ LREC-COLING 2024*, Gilles Sérasset, Hugo Gonçalo Oliveira, and Giedre Valunaite Oleskeviciene (Eds.). ELRA and ICCL, Torino, Italia, 53–65. <https://aclanthology.org/2024.dlnld-1.5/>
- [22] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 142–147. <https://aclanthology.org/W03-0419/>
- [23] Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 5364–5376. doi:10.18653/v1/2024.naacl-long.300
- [24] Shibingfeng Zhang and Giovanni Colavizza. 2026. Named Entity Recognition of Historical Texts via Large Language Model. arXiv:2508.18090 [cs.DL] <https://arxiv.org/abs/2508.18090>

A Reproducibility

The repository⁴ provides the silver labeling, fine-tuning, and evaluation code; the per-fold result files underlying every number in Section 6; the gold set and annotation guideline; the regenerable silver pool (Section 4); and hyperparameter/sample-size derivation detail (Sections 4, 5); trained checkpoints are public.

A.1 Corpus and Task

Prior work by the same authors established named entity recognition baselines on a 108K-record subset of an earlier DDB extract, filtered further by title length and restricted to monographs [21]; the present corpus (Section 3) applies no such length or item-type restriction. The 1700 boundary separates dense Early Modern orthographic markers from a largely marker-free transitional eighteenth century, and tracks where dc:creator availability shifts (absent in 67–71% of pre-1900 records, rarely so after), motivating in-title PERSON extraction in the two historical strata (Section 3). The 1900 boundary matches the HIPE historical NER benchmark [9] (Section 2) rather than any title-length discontinuity, letting the 19th-c. stratum be discussed against HIPE’s German strict-regime F1 range. Language and era are not independently varied: Early Modern and Middle High German forms cluster in the earliest strata, so era-stratified results (Section 6) do not isolate orthographic period from

⁴<https://github.com/anntanp/bibner-werk>

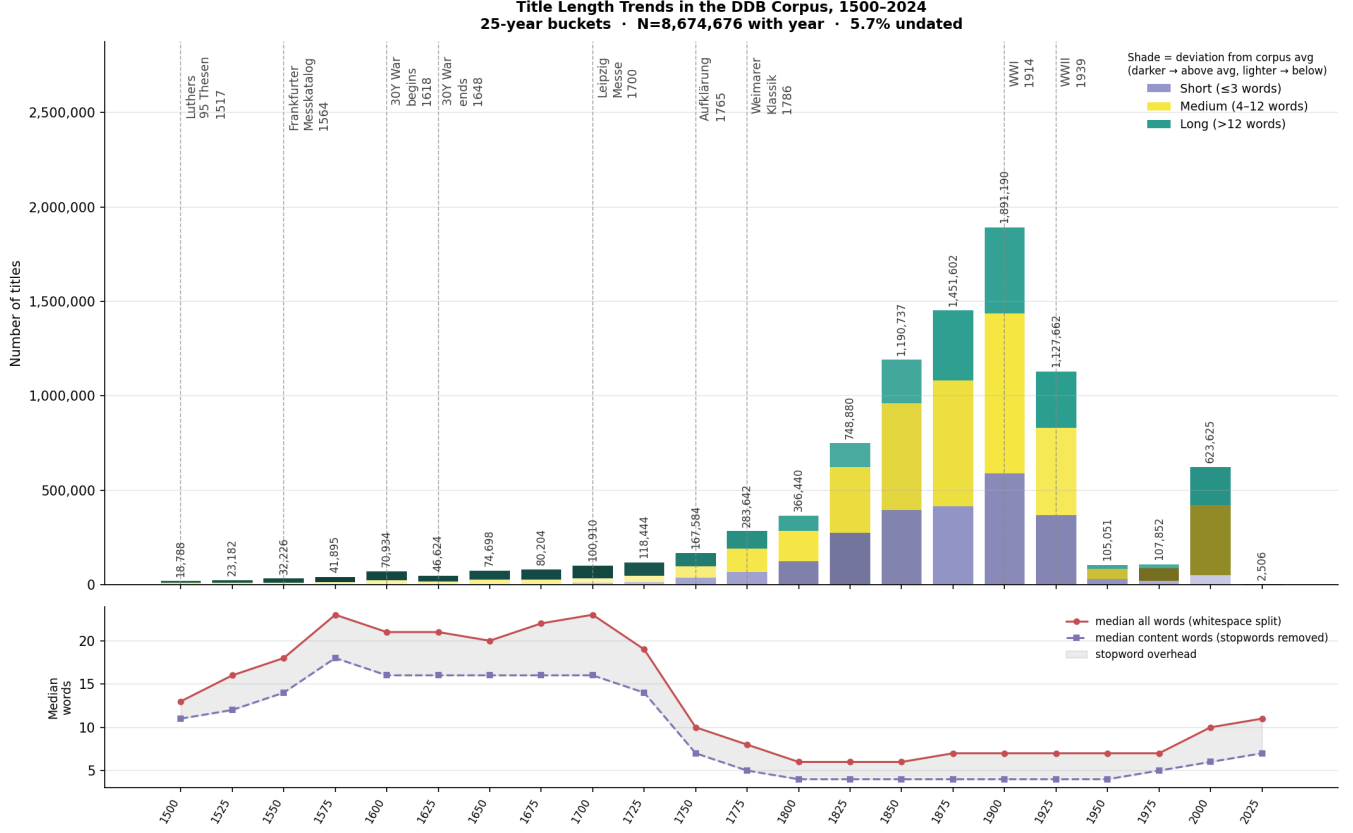


Figure A.1: Median title length (whitespace word count) by 25-year bucket, 1500–2024 (Section 3).

language variety. The retroconversion-driven OCR-noise confound (Section 3) stands alongside genuine historical orthography as a source of noise [8], and is not separately flagged anywhere in the corpus’s metadata schema. Word counts underlying Figure A.1 are whitespace-based, the only length measure comparable across this line of work’s two incompatible subword tokenizers (spaCy [10] vs. xlm-roberta byte-pair encoding): median title length declines from 19 words pre-1700 to 6 in the 19th century before rising again to 8 in the modern stratum, and long-title share follows the same U-shaped pattern, highest in the two earliest strata and lowest in the 19th century.

Gold Annotation. The Wilson-interval sample-size target (Section 4) is met in three of four eras; the 19th century falls two records short, a negligible gap.

Inter-coder agreement was measured using Krippendorff’s α [12, 13] with the nominal distance function, following Artstein and Poesio [2]’s recommendations for a multi-coder, mutually-exclusive-label setting, computed on pre-adjudication annotations to reflect independent labeling. No label reaches the $\alpha \geq 0.800$ threshold recommended by Artstein and Poesio [2] at this pre-adjudication stage: $\alpha = 0.202$ (TITLE), 0.387 (OTHER_TITLE), 0.364 (PERSON); OTHER_TITLE is closest, TITLE is weakest.

Disagreements were resolved through a tiered pipeline: unanimous and clean-majority documents were accepted automatically;

Table A.1: Gold-set composition by era and silver tier (395 records). Pre-1700 has no tier-2 records: the SoR marker defining tier-2 (Section 4) is rare in this era’s title-page conventions.

Era	Tier-0	Tier-1	Tier-2	Total
Pre-1700	96	34	0	130
1700–1800	37	33	10	80
19th-c	15	30	15	60
Modern	20	40	20	80
Unknown	10	35	0	45
Total	178	172	45	395

the remainder was adjudicated by the lead author. This produced the reconciled gold set used in all evaluations reported in Section 6.

Because Krippendorff’s α measures only whether a label is assigned to a record, not where its boundaries fall, span-level agreement was assessed separately as an exact-match rate and mean Jaccard overlap on character-position sets, computed where both annotators agreed a label was present: pooled across labels, exact-match rates ranged from 38.9% to 50.3% and mean Jaccard from 0.726 to 0.792 across the three annotator pairs, with the ordering

only partially mirroring the α results, suggesting whether a label applies at all and where its boundaries fall are two distinct disagreement sources, not one. Of the 395 gold documents, the reconciliation pipeline accepted 42 unanimously and 50 by clean majority; the remaining 303 went to adjudication.

Silver Labeling. A 100-record validation sample found only 35% of statements following the / marker (Section 4) actually name an author, with the remainder split between corporate bodies (19%), editors (5%), and non-responsibility false positives (41%); an earlier keyword-only classifier reached only 41% precision on the same sample, motivating the tagger-based classification used in Section 4. Editor and family-collective statements are filtered from this classification by keyword (*Hrsg.*, *hg.*, *bearb.*, *Familie*, *Nachlass*) before the tagger runs, since a general-purpose tagger would still tag these role-qualified statements as a person; they are consequently left unlabeled rather than assigned PERSON.

Tested as an ISBD area-close marker (Section 4), the trailing period was rejected after showing a 93% false-positive rate on a 200-record sample, dominated by ordinary sentence-final punctuation, ordinal numbers, and date expressions, with genuine area-closes already captured by other signals.

A qualitative review of 16 tier-1 records, drawn from the fine-tuned XLM-R model’s held-out development split and independently re-annotated by one of the paper’s authors without access to the silver labels, disagreed with the silver span on every record. The disagreements are not uniform: eleven of the sixteen share a single pattern, in which a trailing or leading date or issue/volume number is folded into a TITLE or OTHER_TITLE span rather than left unlabeled. A representative case is “*Jeversches Wochenblatt : Friesisches Tageblatt ; gegr. 1791*”, where silver’s OTHER_TITLE span runs through “; *gegr. 1791*”, a founding-date clause the independent annotation excludes entirely. The review covered tier-1 exclusively; tier-2’s reliability, defined by the more restrictive SoR criterion in Section 3, was not independently checked. A related, unquantified risk sits upstream of this date-contamination pattern: silver tier assignment depends entirely on the literal presence of ISBD punctuation markers, principally the area separator (. -) and the : and / field delimiters, so an OCR-introduced substitution or deletion of any of these characters, plausible for the retroconverted subset of records noted in Section 3, would silently reassign a record to a lower tier or corrupt a span boundary rather than raise a visible error. Ehrmann et al. [8] report NER performance degrading sharply even under mild character-level noise (F-score drops 24 points for person entities between good and poor OCR quality); whether this corpus’s OCR-derived subset reaches that range is not established here and is left as an open verification, not a resolved limitation.

A.2 Experimental Setup

Training configuration. GLiNER consumes the silver pool’s character-offset span triples directly, while XLM-R-large [5] requires converting them to BIO tags [19]. The class-weighting scheme reweights each BIO tag by the inverse square root of its training-split frequency to counter PERSON’s scarcity (Section 6); the 256-subword truncation cap drops only 0.41% of entity spans overall and leaves PERSON almost untouched (0.20%), ruling it out as a contributor to PERSON’s weakness. GLiNER’s negative-type sampling

(max_neg_type_ratio) is retained at its package default (1.0), not returned here, since Zaratiana et al. [23]’s ablation targets a 13,000-type pretraining distribution rather than a small fixed label set.

Few-shot verification. Two checks rule out implementation error in the few-shot $F_1 = 0.000$ result (Section 6): entity scores below the 0.5 reporting threshold never reach it either (highest raw score 0.452 across all 3,555 record-label- k combinations), and repeating the run with a hand-curated, artifact-screened example set changed nothing, ruling out both a candidate-generation failure and example quality as the cause.

Fold design caveat. Fold assignment (Section 5) differs from conventional k -fold cross-validation, where each fold’s training data is the complement of its held-out fold; the term is retained here only for the resulting mean-and-std reporting format, not for a genuine train/test partitioning. Dror et al. [6]’s caution that cross-validation folds are not independent test sets applies with extra force to this shared-training-pool design.

Significance test methodology. Table 1’s “confirmed-best” markers come from a paired bootstrap: per-fold score differences between two systems are resampled 1,000 times, and an advantage is confirmed only if the resulting 95% CI excludes zero. This threshold is not adjusted for the many comparisons made across systems, tiers, and metrics, so closer comparisons should be read as suggestive rather than guaranteed; those with the largest observed gaps are least sensitive to this caveat.

A.3 Results

Per-label summary. Table A.2 gives the pooled per-label F1 discussed in Section 6’s prose, included here for convenient reference, with Rule-based omitted since it is training-free and has no zero-shot/few-shot/fine-tuned analog (Section 5). Bold marks the highest F1 per row.

Table A.2: Support-weighted F1 by label, pooled across era strata.

Label	GLiNER-ZS	Few-shot	GLiNER-FT	XLM-R-FT	n
TITLE	0.039	0.000	0.362	0.414	360
OTHER_TITLE	0.000	0.000	0.203	0.324	176
PERSON	0.339	0.000	0.149	0.120	130
Overall	0.087	0.000	0.278	0.333	666

Table A.3 gives the full precision/recall detail underlying the F1-only Table 1 in Section 6, mean across ten folds (Section 5), with Rule-based needing no training and the other three training only on tier-1/tier-2. Bold marks each tier’s confirmed-best system (paired per-fold significance test), tier-2 having none.

Table A.3: Precision, recall, and F1 by silver tier, mean across ten folds.

System	Metric	Tier-0	Tier-1	Tier-2
Rule-based	P	0.335	0.381	0.268
	R	0.392	0.402	0.226
	F1	0.347	0.384	0.240
GLiNER-ZS	P	0.196	0.078	0.310
	R	0.058	0.062	0.289
	F1	0.050	0.040	0.275
GLiNER-FT	P	0.326	0.535	0.380
	R	0.220	0.254	0.213
	F1	0.249	0.332	0.262
XLM-R-FT	P	0.254	0.368	0.277
	R	0.363	0.403	0.228
	F1	0.295	0.377	0.242

A.4 Failure Analysis

Label-mismatch test. The label-mismatch hypothesis (Section 6) was tested by checking every false-negative PERSON span against a list of German institutional nouns: only 1 of 120 matched (a personal name with an appended office title, not a corporate body), ruling out systematic person-versus-organization confusion.

False-negative length. The PERSON false negatives discussed in Section 6 average 38 characters, against 16 for true positives, consistent with the titlature-style span (a monarch named under a full

title, not a bare name) being the harder, longer case the fine-tuned model fails to recover.

Table A.4 reports the per-era breakdown for PERSON, the one label whose Section 6 claim (zero-shot exceeds fine-tuned in every era) needs stratum-level evidence; TITLE/OTHER_TITLE per-era detail is in the released per-fold results instead, since no claim needs it at that granularity.

Table A.4: Per-era F1 for PERSON: GLiNER-ZS, GLiNER-FT, and XLM-R-FT (mean $\pm$ std across ten folds). Bold marks the best system per era.

Era	GLiNER-ZS	GLiNER-FT	XLM-R-FT	<i>n</i>
Pre-1700	0.065 $\pm$ 0.092	0.000 $\pm$ 0.000	0.000 $\pm$ 0.000	14
1700–1800	0.190 $\pm$ 0.140	0.100 $\pm$ 0.200	0.083 $\pm$ 0.171	33
19th-c	0.471 $\pm$ 0.235	0.083 $\pm$ 0.171	0.050 $\pm$ 0.150	25
Modern	0.541 $\pm$ 0.266	0.304 $\pm$ 0.323	0.252 $\pm$ 0.251	46
Unknown	0.015 $\pm$ 0.046	0.000 $\pm$ 0.000	0.000 $\pm$ 0.000	12

False positives and tier-0 blind spot. Both fine-tuned systems’ predicted PERSON spans are frequently over-extended by a leading lowercase “von” (“by”) rather than starting at the name itself (59%/33% of GLiNER-FT’s/XLM-R-FT’s false positives), a pattern the rule-based baseline (Section 5) shares too, as its majority tier-2 false-positive type. At the opposite tier, PERSON is a blind spot for every system: Rule-based, GLiNER-FT, and XLM-R-FT all score exactly zero P/R/F1 against tier-0 PERSON spans (per-fold results in the repository), which carry no punctuation cue.