Ontology-based Information Extraction from Cultural Heritage Digital Representations: A Case Study in Portuguese Archives

Journal Title
XX(X):1–18

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Mariana Dias[1] and Carla Teixeira Lopes[1]

Abstract

Linked Data enables cultural heritage institutions to refine archival descriptions and improve access, but manually creating such descriptions remains labor-intensive. Automating information extraction from digitized records for Linked Data descriptions generation addresses this problem. However, complex or resource-intensive information extraction pipelines are often impractical for resource-constrained archival institutions. This paper identifies and evaluates the most effective methods for ontology-aligned information extraction from Portuguese archival collections under low-resource conditions. We compare early sequence labeling architectures (fine-tuned) with transformer-based zero-shot models for entity and relation extraction aligned with ArchOnto, an ontology designed for the archival domain and based on CIDOC-CRM, an ontology for the cultural heritage domain. Models were evaluated on general-domain Portuguese datasets and domain-specific 20th-century Portuguese archival texts consisting of Optical Character Recognition (OCR)-extracted text and corresponding human-made transcriptions. Our results highlight the challenges of extracting information from noisy archival texts, where BiLSTM-CRF-based named entity recognition models achieved solid performance, while GLiREL produced poor relation extraction results, limiting reliable ontology-guided triple extraction.

Keywords

Information extraction, Cultural heritage, Ontology, Archives

1 Introduction

Linked Data emerged as a way to structure and connect data, enabling interoperability across various domains, including Cultural Heritage (1). Galleries, Libraries, Archives, and Museums (GLAM) and other cultural heritage institutions have increasingly invested in publishing their data as Linked Open Data (LOD) (2). The EPISA (Entity and Property Inference for Semantic Archives) project pursued a similar goal of integrating the Portuguese National Archives into the LOD network by developing an ontology for archives, ArchOnto (3).

Linked Data enables navigation across archival data sources by querying large volumes of data and supporting semantic reasoning (4). It also promotes interoperability between cultural heritage data sources and data from other domains. As a result, users can gain deeper and more comprehensive insights into collections, enriching their understanding and engagement with cultural resources. Figure 1 illustrates the Linked Data representation of a 20th-century document from the Portuguese National Archives[1] using the ArchOnto ontology. However, manually creating Linked Data descriptions for cultural heritage objects remains labor-intensive and time-consuming, making describing documents or collections in finer detail challenging. Automating the extraction of relevant information from digitized archival records to create Linked Data descriptions can alleviate this burden for cultural heritage professionals.

Despite the investment in infrastructures such as the European Collaborative Cloud for Cultural Heritage (ECCCH) (5), archival institutions typically process large volumes of digitized archival documents using legacy infrastructures with limited computation resources and technical expertise. The implementation of complex or resource-intensive information extraction pipelines remains impractical for scalable deployment. Thus, our objective is to identify and evaluate the most effective strategies for ontology-aligned information extraction from Portuguese archival texts under low-resource conditions.

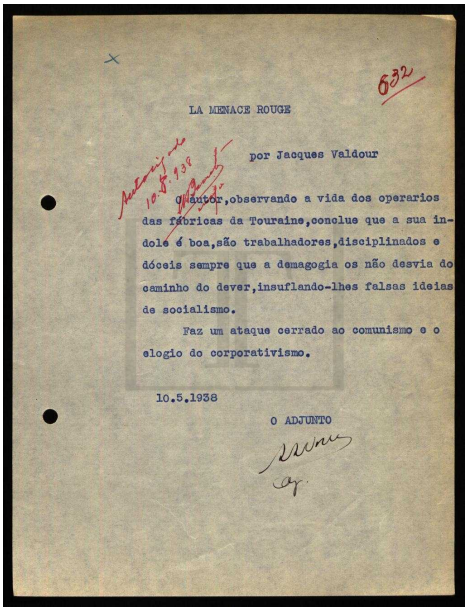
We aim to address the following research questions:

- RQ1: How do early sequence labeling architectures compare with newer transformer-based sequence labeling models for ontology-aligned entity and relation extraction from 20th-century Portuguese archival texts?
- RQ2: How does model performance vary between domain-specific archival texts and general-domain Portuguese documents?

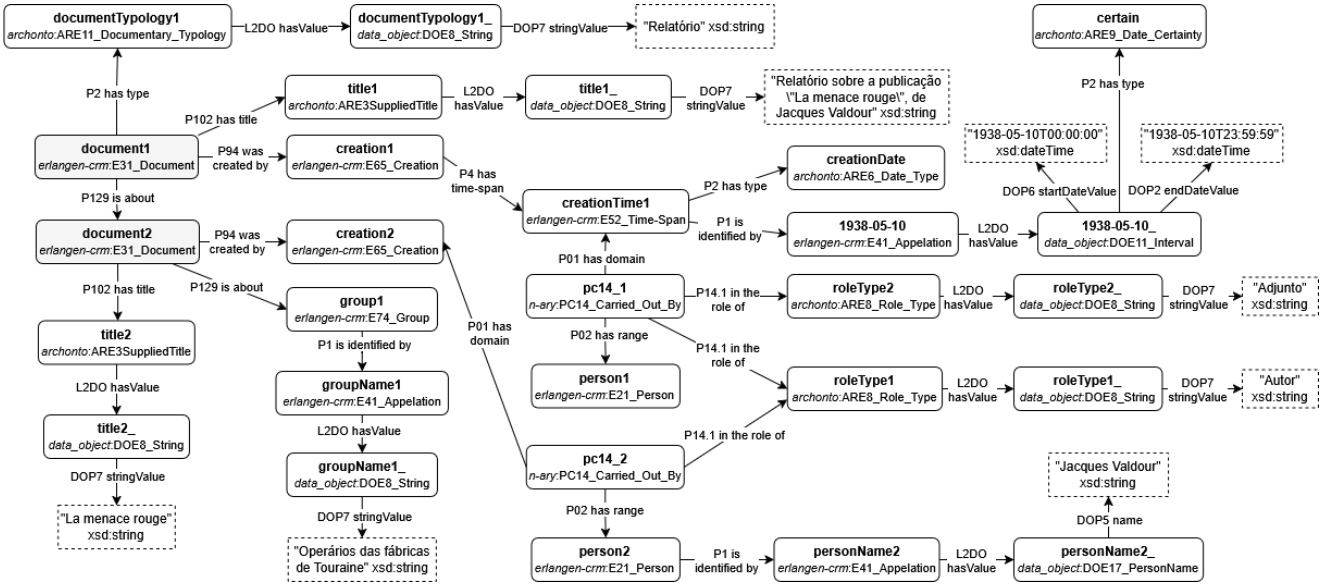
[1]INESC TEC, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

Corresponding author:
Mariana Dias
Email: up201606486@fe.up.pt

[1]Archival record available at: https://digitarq.arquivos.pt/en/documentDetails/384fae40b3aa4b339da5c63a9dede352.



(a) Digitized book censorship report.



(b) ArchOnto representation of digital representation curated by archivists.

Figure 1. Linked Data representation of a 20th-century Portuguese digitized archival document using the ArchOnto ontology.

- RQ3: What is the impact of Optical Character Recognition (OCR)-induced noise on ontology-based entity and relation extraction performance?
- RQ4: How does entity extraction quality affect downstream ontology-aligned relation extraction?
- RQ5: How does the end-to-end extraction pipeline perform when combining the best-performing entity and relation extraction models, and how does this performance vary between clean and OCR-degraded archival texts?

Our approach explores early sequence labeling architectures, such as BiLSTM-CRF with embedding models, and transformer-based sequence labeling models like bidirectional encoders for ontology-aligned entity and relation extraction using ArchOnto. We trained the early sequence labeling models on a general-domain Portuguese dataset. All models were evaluated on the general-domain Portuguese dataset and a domain-specific dataset of 20th-century Portuguese archival texts consisting of OCR-extracted text and corresponding human-made transcriptions.

The main contributions of this work are: (1) annotated general-domain and domain-specific archival documents in Portuguese with entities and relations, and (3) an information extraction pipeline designed for ontology-aligned information extraction from Portuguese archival texts. Datasets and fine-tuned models are publicly available[2], with information extraction model inference code published on GitHub[3].

[2] https://figshare.com/s/cde1ccdfffbae587945d
[3] https://github.com/MarianaFerrDias/ArchExtract

The article is organized as follows. Section 2 details Linked Data models applied in the archival domain and information extraction tools. Section 3 presents relevant related research on information extraction for Portuguese and cultural heritage. Section 4 outlines the workflow and the implementation details of the information extraction experiments. Section 5 describes the creation of datasets used to train and test information extraction models. Section 6 presents the experimental results. Section 7 illustrates the pipeline for mapping extracted information to Linked Data through population mapping rules. The results are discussed in Section 8, and Section 9 concludes with directions for future work.

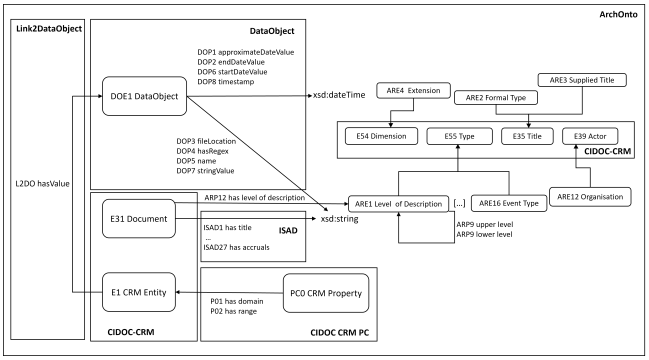


Figure 2. ArchOnto ontology architecture (14).

2 Background

This section describes and compares ontologies used in the archival domain, as well as an overview of information extraction tools.

2.1 Archival Linked Data Models

The International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM) developed CIDOC-CRM[4] (Conceptual Reference Model) (6), an ontology designed to promote semantic interoperability among heterogeneous sources of cultural heritage data. It is recognized as a robust and well-established cultural heritage CRM and ISO standard (ISO 21127:2006, renewed in 2014) (7), and has since served as a foundation for several studies that extend or adapt it to develop new ontologies (8).

The International Council on Archives (ICA) introduced RiC-O[5] (Records in Contexts Ontology), an ontology to represent archival resources (9), with its first stable version released in 2023, which has been adopted in projects and initiatives led by the National Archives of France (10).

Still, various other ontologies have been developed for the archival domain (3; 11; 12; 13). Among them, the ArchOnto ontology was developed within the EPISA project[6] (3). Archonto[7] is a modular ontology for the archival domain consisting of five ontologies, with CIDOC-CRM as the core ontology, and four other satellite ontologies: CIDOC-CRM PC (Property Class) to represent non-binary relationships, DataObject for data validation and data properties, ISAD to include textual information from ISAD(G) (General International Standard for Archival Description) descriptions, and Link2DataObject to connect DataObject and CIDOC-CRM. It extends CIDOC-CRM by defining archival entities and properties with the prefixes *ARE* and *ARP*, respectively. Figure 2 illustrates ArchOnto's modular architecture.

There are have been several studies comparing these three ontologies for archival modeling (15; 16; 17). Oliveira et al. (16) analyzed CIDOC-CRM and RiC-O by comparing their taxonomies and mapping equivalences. Their study found that RiC-O presents a more straightforward data representation, while CIDOC-CRM operates at a higher level of abstraction, with more general classes and finer granularity in its distinction between temporal and physical entities, although having some gaps in the representation of corporate data. Giagnolini et al. (17) compared the RiC-O and ArchOnto ontologies by modeling a dataset of baptism records spanning several centuries. Their analysis showed that RiC-O offers greater flexibility and more straightforward applicability, while ArchOnto presents a more complex model built on a robust foundation that ensures interoperability across diverse GLAM domains through its use of CIDOC-CRM as a core ontology, aligning with the observations of Oliveira et al. (16).

2.2 Information Extraction Tools

Information extraction involves identifying and structuring information, such as entities, relations, and events, from unstructured or semi-structured information (18). Pipelines may decompose it into subtasks, such as named entity recognition (NER) and relation extraction (RE), or perform both in a joint task (19).

Early approaches relied on rule-based methods with gazetteers and patterns to identify entities and relations (18), as implemented in tools like GATE[8] (20). Machine learning methods followed, using feature-based statistical models, such as Hidden Markov Model (HMM) (21) and Conditional Random Fields (CRF) (22; 23), that framed NER and RE as sequence labeling tasks.

Advances in deep learning established bidirectional Long Short-Term Memory (BiLSTM) architectures as state-of-the-art NER models, particularly when combined with CRFs and embeddings, due to their capacity to capture bidirectional context and learn representations (24). Later developments on transformer-based architectures introduced BERT models fine-tuned for NER with token classification (25), for RE using entity-marked sequence inputs (26), and joint NER and RE approaches (27), outperforming prior methods.

Large language models have transformed information extraction from a sequence labeling task into a generation task (28), enabling zero-shot information extraction through prompting engineering (29), as well as approaches like in-context learning, fine-tuning, and knowledge distillation (30), all achieving strong results.

[4] IRI of CIDOC-CRM ontology: http://www.cidoc-crm.org/cidoc-crm/.
[5] IRI of RiC-O ontology: https://www.ica.org/standards/RiC/ontology.
[6] https://episa.inesctec.pt/
[7] IRI of ArchOnto ontology: https://purl.org/episa/archonto.
[8] https://gate.ac.uk/

Beyond large language models, alternative zero-shot approaches for information extraction have emerged. Relik (31) advances open information extraction by jointly performing entity linking and relation extraction through a retrieval-based approach. GLiNER (32) employs a bidirectional transformer for NER that can identify entity types defined at inference time without relying on task-specific retraining by matching them to textual spans in latent space. GLiREL (33) extends this approach to relation extraction using a bidirectional encoder to model entity pair representations and scorers that generalize to unseen relation types.

3 Related Work

This section reviews prior work on entity and relation extraction from archival documents and Portuguese text.

3.1 NER in Portuguese Cultural Heritage

Recent advances in machine-learning-based NER for Portuguese cultural heritage have been driven by neural architectures[9]. Table 1 summarizes key studies, detailing their approaches, annotated corpora, NER datasets outcomes, and evaluation results. The approaches varied in entity scope, although all included person, location, and organization entities.

Vieira et al. (35) and Zilio et al. (37) employed models trained on contemporary Portuguese corpora but applied them to historical 18th century text, while Cunha et al. (41) and Santos et al. (43) trained their models directly on digitized archival data from the 18th century.

Although the identified studies used similar architectures, performance varied significantly across datasets. For instance, CNNs achieved F1-scores ranging from 38.4% in Vieira et al. (35) to 68.4% in Cunha et al. (41). Moreover, BiLSTM-CRF performed best in Vieira et al. (35) and had optimal results in Santos et al. (43), but underperformed in Cunha et al. (41). This performance difference may be due to Cunha et al.'s BiLSTM-CRF architecture not leveraging pre-trained embeddings, unlike Vieira et al. and Santos et al., who employed Flair embeddings (FlairBBP) with their BiLSTM-CRF implementation.

Zilio et al. (37) and Santos et al. (43) evaluated transformer architectures with LLama2, mT5, and masked language models, such as BERT-based models and XLM-R for NER in Portuguese digitized archives. The results of their work report that BERT-based and XLM-R models achieved the best results. In contrast, LLama2 and mT5 yielded F1-scores below 50%.

3.2 RE in Cultural Heritage

Relation extraction enables the identification and categorization of meaningful relationships between cultural artifacts, historical events, and other entities within the cultural heritage domain. Efremova et al. (46) compared the performance of a Support Vector Machine (SVM) model to a HMM to extract family relationships in historical notary acts. Chantaraj et al. (47) implemented a rule-based approach to extract relationships in Thai Buddhist temple documents. Christou et al. (48) employed a distantly supervised BERT-based model to extract semantic relationships in 19th-century Greek Literature documents. Recently, the CLEF Evaluation Lab introduced HIPE-2026 (49), a shared task for person-place relation extraction from multilingual OCR-derived historical documents.

To our knowledge, RE has not been explored in the domain of Portuguese CH. However, some works have applied RE to the broader domain of the Portuguese language (50; 51). Portuguese evaluation contests, such as HAREM (52) and IberLEF 2019 (53), included RE tasks. Santos et al. (50) created a rule-based system to extract family relations. Collovini et al. (51) evaluated a CRF model for open-relation extraction using the HAREM corpora and reported a 26% F1-score increase in extracting the placement relation compared to other RE systems. Still, a gap in research remains, as no prior work has investigated an end-to-end NER+RE pipeline for Portuguese cultural heritage archival documents using existing models for low-resource conditions.

Table 1. NER approaches in Portuguese Cultural Heritage (F1=F1-score, FM=rate of full matches).

Approach	Testing data	Model	Evaluation
Vieira et al. (35)	Parish Memories (1758–1761) (36)	BiLSTM-CRF +FlairBBP*	F1=45.8
		CNN**	F1=38.4
Zilio et al. (37)	18th-century medical texts (38; 39; 40)	BERT-CRF*	FM=83.7
		CNN (spaCy_lg)**	FM=55.5
		CNN (spaCy_sm)**	FM=63.7
Cunha et al. (41) ***	Dataset from Portuguese Archives (42)	BiLSTM-CRF	F1=53.0
		CNN	F1=68.4
		Maximum Entropy	F1=66.6
Santos et al. (43) ***	Manually annotated subset of the Parish Memories dataset	BERTimbau-Large	F1=70.5
		BiLSTM-CRF +FlairBBP +W2V-SKPG	F1=67.5
		BiLSTM-CRF +FlairBBP +Glove	F1=66.3
		LLama2 (8bit) + LoRa	F1=49.0
		mT5-large	F1=42.8
		XLM-R-Large	F1=70.8

*Trained on First HAREM dataset (44); **Trained on WikiNER (45); ***Trained on datasets from Portuguese archives.

[9] See Ehrmann et al. (34) for a literature review of NER in historical documents.

4 Methodology

The proposed experimental workflow for ontology-based information extraction from archival documents is shown in Figure 3.

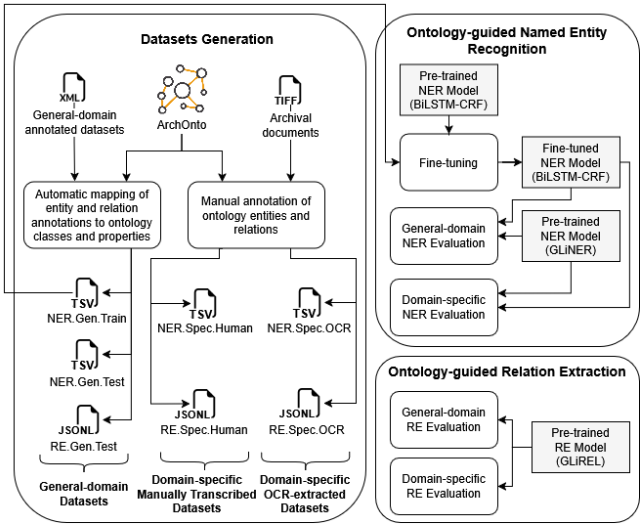


Figure 3. Experimental workflow for ontology-guided information extraction from archival documents.

We first generate three dataset types: (1) general-domain NER datasets (NER.Gen.Train and NER.Gen.Test) from existing annotations mapped to ArchOnto for pre-trained NER models fine-tuning and to baseline named entity recognition on contemporary Portuguese, (2) general-domain RE dataset (RE.Gen.Test) to baseline relation extraction on contemporary Portuguese, and (3) domain-specific archival datasets from human-made transcriptions (NER.Spec.Human and RE.Spec.Human) and OCR-extracted texts (NER.Spec.OCR and RE.Spec.OCR) to evaluate target-domain performance. Dataset generation is detailed in Section 5, with ontology-aligned NER and RE experiments described below. All experiments were executed on a personal computer with an Intel® Core™ i7-8750H CPU @ 2.20GHz (2.21 GHz) and 16 GB of RAM, using only CPU-based processing.

4.1 Ontology-aligned NER

We evaluated two NER approaches: (1) an early neural architecture for sequence labeling, a BiLSTM-CRF model (54) requiring fine-tuning on data annotated with specified entity types, and (2) a recent transformer-based model for sequence labeling, GLiNER(32), which does not require any training nor predefined entity types.

4.1.1 Implementation Details We used Flair NLP[10] (55), an open-source library for sequence labeling, to fine-tune BiLSTM-CRF models with word and contextual embeddings on the NER.Gen.Train dataset. We employed two pre-trained contextual language models: FlairBBP[11] (56), bidirectional Flair embeddings trained on 4.9B tokens from three large Portuguese corpora, and FlairEL[12] (57), Flair embeddings trained on 0.9B tokens of Portuguese CommonCrawl data. We also used a pre-trained word embedding model, Skip-gram Word2Vec (300 dimensions)[13]. We adopted the training hyperparameters set by Santos et al. (56) to avoid the computational cost of optimizing hyperparameters. We set the initial learning rate to 0.1 with an annealing factor of 0.5 on every three epochs without improvement. The training was stopped after 150 epochs or if the learning rate was below 0.0001.

We used GLiNER[14] to run three models from the v2.1 release: small (50M parameters), medium (90M), and multi. We specified eight entity types: Person, Organization, Date, Place, Role, Event, Title, and Dimension. The Group label was replaced with Organization, as the latter is a less ambiguous entity type in NER tasks. A threshold of 0.5 was applied to filter predicted entities, with data processed in batches of sentences for the general-domain dataset and per document for the domain-specific dataset.

4.1.2 Evaluation Metrics Following SemEval 2013 (58) guidelines, we evaluated models using strict matching and type matching, reporting F1-scores. Strict matching requires exact named entity matches (59), while type matching considers the correctness of named entity terms regardless of boundaries.

4.2 Ontology-aligned RE

We used GLiREL (33), a transformer-based zero-shot sequence labeling model for relation extraction, under two configurations: (1) unconstrained relation extraction using only relation labels, and (2) constrained relation extraction with entity type restrictions.

4.2.1 Implementation Details We employed the GLiREL[15] (glirel-large-v0) model, experimenting with different top-k retrieval and thresholding hyperparameters for accepting identified triples. We also tested two configurations to assess the impact of entity type restrictions. Table 2 lists the constraints for head and tail entity types for each relation from Section 5.3.

4.2.2 Evaluation Metrics We evaluated models using strict matching, reporting precision, recall, and F1-score. Moreover, we assessed ontology compatibility by calculating the percentage of predicted properties satisfying defined head and tail entity type constraints:

$$\text{Score} = \frac{\text{Number of valid properties}}{\text{Total number of predicted properties}} \quad (1)$$

5 Datasets Generation

This section describes the creation of general-domain training and test datasets for NER and RE, as well as a domain-specific evaluation dataset of archival documents.

[10] https://github.com/flairNLP/flair

[11] Repository available online at: https://github.com/jneto04/ner-pt.

[12] Repository available online at: https://github.com/ericlief/language_models.

[13] http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc.

[14] https://github.com/urchade/GLiNER

[15] Repository available online at: https://github.com/jackboyla/GLiREL.

Table 2. Head and tail entity constraints for relation extraction.

Relation	Head	Tail
P4 has time-span	E5 Event	E52 Time-Span
P7 took place at	E5 Event	E53 Place
P14.1 in the role of	E21 Person, E74 Group	ARE2 Formal Title
P53 has current or former location	E74 Group	E53 Place
P74 has current or former residence	E74 Group	E53 Place
P89 falls within	E53 Place	E53 Place
P107 has current or former member	E74 Group	E21 Person, E74 Group
author of	E21 Person	ARE2 Formal Title
date of creation	ARE2 Formal Title	E52 Time-Span
director of	E21 Person	E74 Group
editor of	E21 Person	ARE2 Formal Title
has owner	E74 Group	E21 Person, ARE8 Role Type
participant in	E74 Group	E21 Person, ARE8 Role Type
produced by	ARE2 Formal Title	E21 Person, ARE8 Role Type

5.1 General-domain NER Dataset

Given the scarcity of manually curated annotated archival collections in Portuguese, we adapted contemporary general-domain NER corpora to align with ArchOnto's structure. We adapted the HAREM Golden Collections (GC), which consist of two datasets: the First HAREM GC (44), and the Second HAREM GC (52). The First HAREM GC includes 1,202 documents and 5,132 named entities, while the Second HAREM GC contains 1,040 documents and 7,847 named entities. The collections cover a range of entity categories, such as Person, Location, Organization, Time, Value, Abstraction, Event, Thing, Work, and Other. The pipeline for the creation of the training and test datasets, NER.Gen.Train and NER.Gen.Test, adapted from the HAREM GCs, is shown in Figure 4.

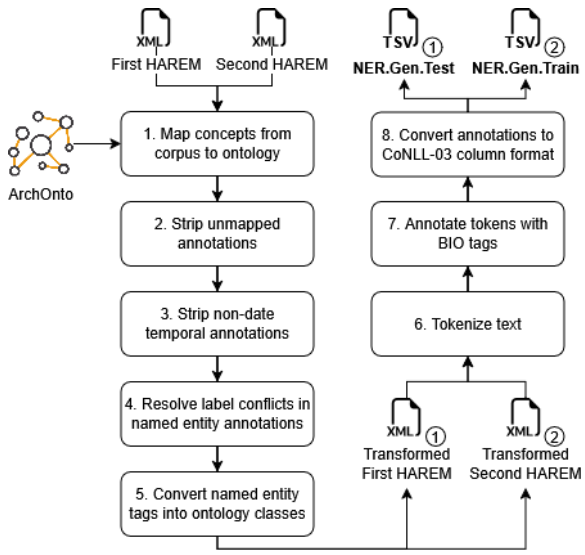


Figure 4. Pipeline for creating training and testing general-domain datasets for the NER task.

First, we aligned ArchOnto concepts with the HAREM GCs' labels and manually defined mapping rules by consulting the HAREM dataset documentation and drawing upon established knowledge of ArchOnto's structure and existing entity representations (task 1 in Figure 4). We detail the mapping between HAREM GCs concepts and ArchOnto classes in Table 3, using CIDOC-CRM and its archival extension. We then excluded unmapped and non-date temporal annotations (tasks 2 and 3), such as "passada sexta-feira" (last Friday), "mais tarde" (later), "ontem" (yesterday), and "há quinze anos" (fifteen years ago), as they fall outside the scope of ArchOnto's representation of dates.

Table 3. Mapping between concepts in the HAREM GCs and ArchOnto classes.

ArchOnto Class	HAREM Mapping (category→type→subtype)
ARE2 Formal Title	Obra (Work)→ Arte (Art) → Pintura (Painting)
	Obra (Work) → Plano (Plan)
	Obra (Work) → Reproduzida (Reproduced) → Livro (Book), Program (Program), Musica (Music), Outro (Other)
ARE8 Role Type	Pessoa (Person) → Cargo (Position), GrupoCargo (GroupRole)
E5 Event	Acontecimento (Event) → Efemeride (Anniversary), Evento (Event), Organizado (Organized), Outro (Other)
E21 Person	Pessoa (Person) → Individual, Membro (Member)
E52 Time-Span	Tempo (Time) → Data (Date), Periodo* (Period)
	Tempo (Time) → Tempo_Calend** (Time_Calend) → Data (Date), Intervalo (Interval)
E53 Place	Local (Place) → Humano (Human) → Rua (Street), Pais (Country), Divisao (Division), Regiao (Regiion), Outro (Other)
	Local (Place) → Fisico (Physical) → Ilha (Island), Aguacurso (Water course), Planeta (Planet), Regiao (Region), Relevo (Relief), Aguamassa (Water mass), Outro (Other)
E54 Dimension	Valor (Value) → Quantidade (Quantity)
E74 Group	Organizacao (Organization) → Administracao (Administration), Empresa (Company), Instituicao (Institution), Outro (Other)
	Pessoa (Person) → GrupoInd (GroupInd), GrupoMember (GroupMember)

* Mapping with First HAREM GC; ** Mapping with Second HAREM GC

The GCs contained 450 ambiguous entities with overlapping classifications that needed to be disambiguated (e.g., "Twin Towers" was classified as *E5 Event* and *E53 Place*). As shown in Table 4, *E74 Group* was the most frequent among ambiguous entities (310 occurrences), followed by *E53 Place* (224), *E5 Event* (132), and *ARE8 Role Type* (113), with all others below 100 occurrences. We manually resolved these conflicts using context clues while ensuring consistency among related entities (task 4). For example, if an ambiguous entity shared inclusion

relationships with other named entities, such as a place nested within a place and a group within a group, all were classified under the same ontology class to ensure hierarchical coherence.

Table 4. Frequency of ambiguous entity classifications per ArchOnto class in pre-processed HAREM datasets.

Class	HAREM I	HAREM II	Total
ARE2 Formal Title	5	61	66
ARE8 Role Type	3	110	113
E5 Event	22	110	132
E21 Person	2	62	64
E52 Time-Span	9	35	44
E53 Place	16	208	224
E54 Dimension	5	14	19
E74 Group	61	249	310

Subsequently, HAREM tags were systematically converted to ArchOnto labels based on the previously described alignment, resulting in the Transformed First and Second HAREM datasets (task 5). Following the CoNLL-03 (60) (Computational Natural Language Learning) standard, we converted the transformed HAREM corpora by tokenizing XML content and implementing BIO tagging (Beginning, Inside, and Outside) (tasks 6, 7, and 8).

The distribution of named entities in the NER datasets is shown in Table 5, revealing class imbalances between entity types.

Table 5. Corpus statistics for general-domain NER datasets.

Class	Entity Count	
	NER.Gen.Train	NER.Gen.Test
ARE2 Formal Title	503 (7.5%)	267 (4.5%)
ARE8 Role Type	178 (2.7%)	92 (1.6%)
E5 Event	313 (4.7%)	156 (2.7%)
E21 Person	1,863 (27.7%)	1,333 (22.6%)
E52 Time-Span	644 (9.6%)	462 (7.8%)
E53 Place	1,390 (20.7%)	1,639 (27.8%)
E54 Dimension	247 (3.7%)	666 (11.3%)
E74 Group	1,591 (23.6%)	1,281 (21.7%)
Total	6,729	5,896

5.2 General-domain RE Dataset

The Second HAREM contest produced the ReRelEM (Recognition of Relations between Mentioned Entities) GC, a Portuguese relation-annotated dataset. It includes 4,803 relations of 38 relation labels, both symmetric and direct/indirect, such as identity (*ident*), inclusion (*inclui/incluido*), and location (*ocorre_em/sede_de*). We used a pre-processed subset of the ReRelEM dataset to develop an ArchOnto-aligned test dataset for RE models, as outlined in Figure 5.

We first analyzed ReRelEM GC's relation labels to identify which could be mapped to ArchOnto's structure, selecting 14 relevant relationships (task 1 in Figure 5. After applying the dataset pre-processing steps described in the previous section of removing unmapped annotations (task 2), resolving ambiguous label annotations (task 3), and converting HAREM tags to ArchOnto class and property labels (task 4), we mapped the most frequent relations to ArchOnto property assertions, as shown in Table 6: "autor_de/obra_de" (authorship), "inclui/incluido" (inclusion), "participante_em/ter_participacao_de" (participation), and "vinculo_inst" (affiliation).

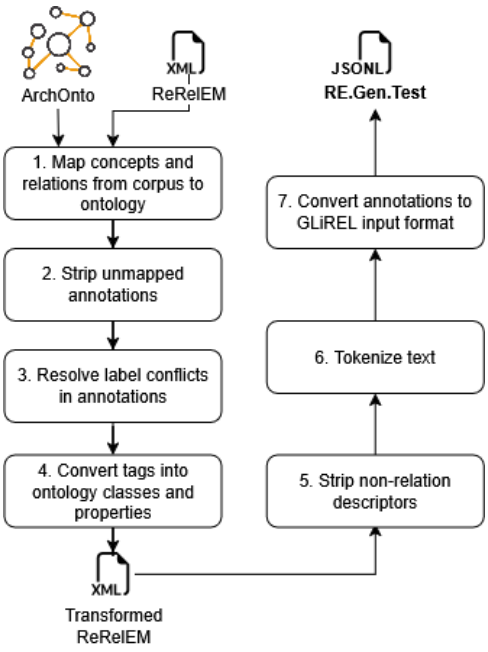


Figure 5. Pipeline for creating general-domain testing dataset for the RE task.

Table 6. Mapping between relations in the HAREM GC and ArchOnto property assertions.

HAREM Relation	ArchOnto Triples
authorship	E31 Document → P102 has title → E35 Title* E31 Document → P94 was created by → E65 Creation PC14 Carried Out By → P14.1 in the role of → ARE8 Role Type ARE8 Role Type → L2DO hasValue → DOE8 String DOE8 String → DOP7 stringValue → "Autor xsd:string" PC14 Carried Out By → P01 has domain → E65 Creation** PC14 Carried Out By → P02 has range → E39 Actor***
inclusion	E74 Group → P107 has current or former member → E39 Actor***
	E53 Place → P89 falls within → E53 Place
participation	PC14 Carried Out By → P01 has domain→E7 Activity** PC14 Carried Out By → P02 has range → E39 Actor***
affiliation	E74 Group → P107 has current or former member → E21 Person

*Includes individuals that are subclasses of *E35 Title* (e.g., *ARE2 Formal Title*, *ARE3 Supplied Title*); **Includes individuals of E7 Activity and its subclasses (e.g., *E5 Event*, *E65 Creation*, etc.); ***Includes individuals that are subclasses of *E39 Actor* (e.g, *E21 Person*, *E74 Group*)

The distribution of these relations in the RE dataset is displayed in Table 7, where bidirectional relations are aggregated.

Table 7. Corpus statistics for RE.Gen.Test dataset.

Property	Relation Count
autor_de / obra_de (authorship)	91 (17.0%)
inclui / incluido (inclusion)	273 (50.8%)
participante_em / ter_participacao_de (paticipation)	72 (13.4%)
vinculo_inst (affiliation)	101 (18.8%)
Total	537

5.3 Domain-specific Archival Datasets

We developed domain-specific datasets to evaluate information extraction in Portuguese archival documents. First, we manually developed a dataset comprised of ArchOnto representations of 13 Portuguese 20th-century archival records extracted from the Torre do Tombo National Archive (61). Two archivists from the General Directorate for Book, Archives and Libraries (DGLAB) described the typewritten textual content visible in the digital representations using the ArchOnto ontology. The descriptions were then harmonized through a consensus process and converted into OWL format. These representations serve as the ground truth for assessing the potential of ontology-aligned information extraction.

Using these 13 records, we built four domain-specific evaluation datasets for NER and RE: OCR-extracted text (NER.Spec.OCR, RE.Spec.OCR) and manually transcribed text (NER.Spec.Human, RE.Spec.Human). The original text in both dataset types originates from prior work on OCR for cultural heritage typewritten documents (62). We then manually annotated these texts with entity-relation triples. Table 8 presents domain-specific archival dataset statistics: per-document (number of triples, entities, and words), plus entity and relation distributions.

The domain-specific datasets mostly contain *E74 Group*, *ARE8 Role Type*, and *E52 Time-Span*, and *E53 Place* entities. NER.Spec.OCR has fewer entities than NER.Spec.Human due to OCR errors in low-quality documents and handwritten text. Moreover, *E74 Group*, *E52 Time-Span*, *E53 Place* and *ARE2 Formal Title* entities average over 3 words per mention.

Relations in the datasets primarily include the properties *P107 has current or former member*, *P14.1 in the role of*, and *P74 has current or former residence*, aligning with the most frequent entities. Relations with property *E52 Time-Span* are scarce, likely creating dangling entities.

Document-level statistics show fewer triples than ArchOnto gold standard representations, likely due to lack of explicit relation descriptors, ArchOnto's abstract modeling (e.g., implicit *E7 Activity* entities), and limited coverage of named entity recognition for activity-related entities. NER.Spec.OCR generally has fewer words, entities, and triples than NER.Spec.Human, depending on the document's OCR quality. OCR-extracted documents 6, 7, and 8 contain zero entities and triples due to gibberish OCR output. In contrast, OCR-extracted document 1 contains more entities than its transcription, because it recognized image regions that were not included in the manual transcription.

Figure 6 demonstrates an excerpt of the annotation of entities and relations in manually transcribed and OCR-extracted text from document 13, shown in Figure 1.

6 Results

This section presents the results of named entity recognition and relation extraction on the general-domain and domain-specific archival datasets.

6.1 Ontology-aligned NER

This section presents the NER results on the general-domain and domain-specific archival datasets.

6.1.1 General-domain Evaluation We evaluated the models on a general-domain dataset, NER.Gen.Test, to establish a baseline for comparison with archival data. Table 9 presents F1-scores for the BiLSTM-CRF NER models, organized by entity label, embedding models, and evaluation scenario in the NER.Gen.Test dataset.

In the strict matching evaluation scenario, models that employed a Skip-gram Word2Vec word embedding model generally outperformed others on nearly all named entity categories, with the BBP+W2V combination achieving the highest overall score. As expected, the most frequently annotated entities, *E21 Person*, *E52 Time-Span*, *E53 Place*, and *E74 Group*, achieved the best results. However, the models struggled to recognize events, titles, and roles, achieving F1-scores below 40%.

Type matching substantially improved recognition for *E52 Time-Span* and *E54 Dimension* by resolving boundary-related errors. For events, titles, and roles, however, the difference between type matching and strict matching is less apparent, indicating that the models have more difficulty in identifying the correct entity types, rather than simply making boundary errors.

Table 10 presents F1-scores for the GLiNER models, organized by entity label, model size, and evaluation scenario in the NER.Gen.Test dataset.

GLiNER models showed noticeably low performance on the NER.Gen.Test dataset, achieving an overall F1-score of 4.63% with the multi model under the strict match evaluation. Entities ARE2 Formal Title, *ARE8 Role Type*, *E5 Event*, and *E54 Dimension* obtained F1-scores below or close to 1%, likely because they appear less frequently (see Table 5) and have higher contextual variability, making them harder to detect in a zero-shot setting. For instance, the multi model misclassified *Role Type* entities as *Title*, likely interpreting the *Title* label as forms of address rather than as references to document or work titles.

Type matching did not substantially improve entity recognition, indicating that the low performance was not primarily related to boundary identification errors but rather to the presence of spurious entity predictions. High false positive error rate might signify a combination of language mismatch, as the NER.Gen.Test dataset is in Portuguese, while the models were predominantly trained on English, and contextual misinterpretation, even for common entity

Table 8. Summary statistics of entity, relation, and document-level annotations for NER.Spec.Human, RE.Spec.Human, NER.Sec.OCR, and RE.Spec.OCR datatsets.

(a) Document-level statistics (T=triples, E=entities, W=words).

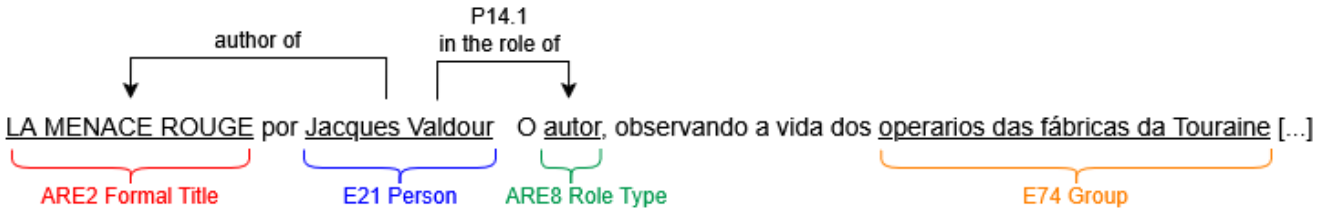
ID	Title	Spec.Human			Spec.OCR		
		|T|	|E|	|W|	|T|	|E|	|W|
1	Letter from Ramalho Ortigão requesting personal papers found on his desk at the Ajuda Library	3	18	144	3	15	163
2	Authorization for the delivery of goods to the University of Coimbra Museum	3	13	103	3	14	112
3	Work order and authorization for delivery of goods related to Request 12	2	16	169	0	4	41
4	Clarifications regarding the delivery of musical instruments	3	12	107	1	5	67
5	Request for delivery of chairs for a party at the Park of Necessidades	3	15	95	1	13	93
6	"A flor do bairro" (The flower of the neighborhood)	0	2	9	0	0	7
7	"Actas das Sessões 1961" (Minutes of the 1961 Sessions)	2	3	7	0	0	4
8	"O Norte Desportivo" (The Sporting North)	6	15	84	0	11	62
9	Case concerning the analysis of the publication of the book "O Assalto ao Santa Maria" (The Assault on the Santa Maria), authored by Henrique Galvão, to be published by Delphos, with its sale prohibited	2	8	27	0	5	28
10	Ban on the book "Vagão J"	0	3	23	0	0	9
11	"A Marinha Grande"	1	9	67	0	8	64
12	"Petróleo no Mundo" (Oil in the World)	4	10	48	4	10	66
13	Report No. 00632 "La menace rouge"	2	6	57	2	5	63

(b) Entity class distribution.

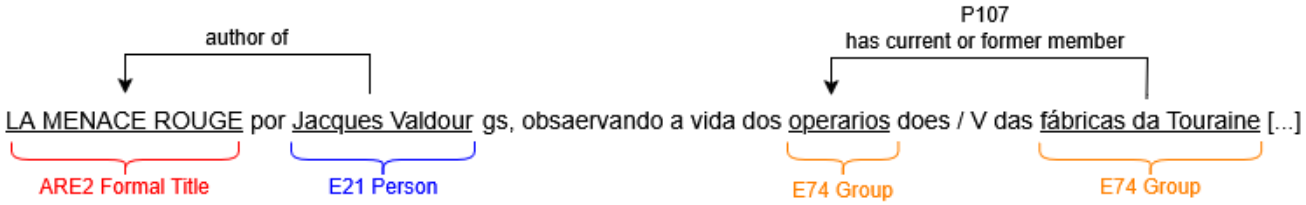
Class	Spec.Human		Spec.OCR	
	E	W / E	E	W / E
E74 Group	46 (35.4%)	4	31 (34.4%)	3
ARE8 Role Type	30 (23.1%)	1	22 (24.2%)	1
E52 Time-Span	20 (15.4%)	4	12 (13.2%)	4
E53 Place	15 (11.6%)	3	13 (14.4%)	2
E21 Person	10 (7.7%)	3	9 (9.9%)	2
ARE2 Formal Title	7 (5.4%)	5	3 (3.3%)	4
E5 Event	1 (0.8%)	1	0 (0.0%)	-
E54 Dimension	1 (0.8%)	1	0 (0.0%)	-
Total	130	3	73	3

(c) Relation property distribution.

Property	Relation Count	
	Spec.Human	Spec.OCR
P107 has current or former member	10 (34.5%)	6 (50.0%)
P14.1 in the role of	3 (10.3%)	1 (8.3%)
P74 has current or former residence	2 (6.9%)	1 (8.3%)
P4 has time-span	1 (3.4%)	0 (0.0%)
P7 took place at	1 (3.4%)	0 (0.0%)
P53 has current or former location	1 (3.4%)	0 (0.0%)
P89 falls within	1 (3.4%)	0 (0.0%)
director of	3 (10.3%)	1 (8.3%)
editor of	3 (10.3%)	1 (8.3%)
author of	2 (6.9%)	1 (8.3%)
has owner	2 (6.9%)	1 (8.3%)
date of creation	1 (3.4%)	0 (0.0%)
produced by	1 (3.4%)	0 (0.0%)
Total	29	12



(a) Human-made transcription.



(b) OCR-extracted text.

Figure 6. Gold-standard annotations for Document 13 from domain-specific dataset. **(a)** Human-made transcription: *ARE2 Formal Title*, *E21 Person*, *ARE8 Role Type*, and *E74 Group* entities with *P14.1 in the role of* relations for authorship (*E21→ARE2*) and role assignment (*E21→ARE8*). **(b)** OCR-extracted text: *ARE8 Role Type* entity extraction failure and *E74 Group* entity splitting connected by *P107 has current or former member* relation (*E74→E74*).

Table 9. F1-score (%) performance of BiLSTM-CRF models using different embedding models on the NER.Gen.Test dataset (EL=FlairEL, W2V=Skip-gram Word2Vec, BBP=FlairBBP).

Class	Strict Match				Type Match			
	EL	EL+W2V	BBP	BBP+W2V	EL	EL+W2V	BBP	BBP+W2V
ARE2 Formal Title	31.26	**33.40**	29.71	30.71	**47.70**	44.40	39.75	45.23
ARE8 Role Type	26.61	30.70	**32.03**	29.79	36.29	41.23	**45.89**	38.30
E5 Event	35.54	**38.46**	35.82	35.62	46.39	**52.07**	48.36	48.12
E21 Person	75.97	76.47	78.27	**79.62**	88.64	90.23	88.82	**91.45**
E52 Time-Span	**76.87**	75.03	75.22	74.47	90.28	**91.97**	89.66	89.86
E53 Place	84.67	**86.49**	84.48	85.13	87.48	**88.92**	86.58	87.05
E54 Dimension	**54.02**	49.04	45.48	50.73	**75.86**	75.40	71.52	68.38
E74 Group	67.11	69.40	68.28	**71.49**	73.15	75.27	73.88	**76.82**
Overall	70.19	71.05	69.98	**71.93**	79.68	**81.24**	79.17	80.42

Bold values indicate the best F1-score per entity type and evaluation scenario.

Table 10. F1-score (%) performance of GLiNER models (sm=small, md=medium, mt=multi) on the NER.Gen.Test dataset.

Class	Strict Match			Type Match		
	sm	md	mt	sm	md	mt
ARE2 Formal Title	**1.29**	0.54	0.00	3.86	**4.08**	3.24
ARE8 Role Type	0.00	0.00	0.00	0.00	**0.40**	0.00
E5 Event	0.63	0.85	**0.90**	1.56	1.71	**2.54**
E21 Person	**6.46**	3.45	3.19	3.06	**7.06**	6.44
E52 Time-Span	5.61	5.04	**5.63**	**9.35**	9.01	8.78
E53 Place	7.71	6.18	**8.76**	11.36	9.90	**12.49**
E54 Dimension	0.21	**0.57**	0.39	1.24	2.44	**3.30**
E74 Group	4.77	4.50	**4.89**	11.52	11.57	**12.12**
Overall	4.49	3.82	**4.63**	8.46	7.69	**8.70**

types, such as *E21 Person*, *E74 Group*, and *E53 Place*. For example, the GLiNER models frequently classified generic references to people as instances of *E21 Person* (e.g., "você"/"you", "a minha mãe"/"my mum").

6.1.2 Domain-specific Evaluation We evaluated the models on domain-specific archival datasets to assess their ability to extract entities from Portuguese 20th-century digitized archival records. Tables 11 and 12 report BiLSMT-CRF performance, while Table 13 presents GLiNER results on OCR-extracted and human-transcribed datasets.

BiLSTM-CRF-based models outperformed GLiNER across both datasets. The BiLSTM-CRF variant combining

Table 11. F1-score (%) performance of BiLSTM-CRF models using different embedding models on the NER.Spec.OCR dataset (EL=FlairEL, W2V=Skip-gram Word2Vec, BBP=FlairBBP).

Class	Strict Match				Type Match			
	EL	EL+W2V	BBP	BBP+W2V	EL	EL+W2V	BBP	BBP+W2V
ARE2 Formal Title	0.00	**50.00**	0.00	0.00	0.00	**50.00**	0.00	0.00
ARE8 Role Type	48.65	48.65	48.28	**50.00**	54.05	**59.46**	48.28	50.00
E5 Event	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E21 Person	44.44	**52.17**	37.50	47.06	55.56	52.17	62.50	**70.59**
E52 Time-Span	47.62	50.00	47.62	**52.63**	**85.71**	80.00	**85.71**	73.68
E53 Place	15.38	**48.00**	9.09	28.57%	46.15	**72.00**	27.27	47.62
E54 Dimension	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E74 Group	26.09	**28.95**	22.64	28.07	55.07	50.00	**56.60**	52.63
Overall	33.33	**40.86**	29.93	36.60	56.32	**58.06**	53.06	53.59

Table 12. F1-score (%) performance of BiLSTM-CRF models using different embedding models on the NER.Spec.Human dataset (EL=FlairEL, W2V=Skip-gram Word2Vec, BBP=FlairBBP).

Class	Strict Match				Type Match			
	EL	EL+W2V	BBP	BBP+W2V	EL	EL+W2V	BBP	BBP+W2V
ARE2 Formal Title	0.00	**54.55**	0.00	0.00	0.00	**54.55**	0.00	0.00
ARE8 Role Type	**56.52**	48.89	50.00	37.21	**60.87**	57.78	50.00	41.86
E5 Event	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E21 Person	28.57	36.36	**52.17**	27.59	40.00	54.55	**69.57**	55.17
E52 Time-Span	44.44	**57.14**	46.67	51.85	66.67	71.43	**73.33**	66.67
E53 Place	14.81	29.63	24.00	**33.33**	29.63	29.63	**48.00**	33.33
E54 Dimension	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E74 Group	32.99	29.41	31.25	**34.88**	63.92	64.71	58.33	**65.12**
Overall	34.71	**37.75**	36.28	34.39	53.72%	**57.83**	55.75	52.49

Table 13. F1-score (%) performance of GLiNER models (sm=small, md=medium, mt=multi) on the domain-specific datasets.

Class	Spec.Human						Spec.OCR					
	Strict Match			Type Match			Strict Match			Type Match		
	sm	md	mt	sm	md	mt	sm	md	mt	sm	md	mt
ARE2 Formal Title	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ARE8 Role Type	0.00	**4.76**	0.00	0.00	**4.76**	0.00	0.00	0.00	0.00	**7.69**	7.14	0.00
E5 Event	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E21 Person	7.69	**9.09**	8.70	15.38	18.18	**26.09**	9.52	8.33	**10.00**	19.05	16.67	**30.00**
E52 Time-Span	6.25	0.81	**10.26**	25.00	32.43	**35.90**	0.00	**10.53**	0.00	**42.11**	31.58	31.58
E53 Place	13.79	**15.38**	**15.38**	20.69	**30.77**	**30.77**	8.00	8.70	**9.09**	**24.00**	17.39	18.18
E54 Dimension	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E74 Group	**8.82**	5.33	0.00	**38.24**	32.00	33.80	4.44	**11.76**	4.17	26.67	23.53	**29.17**
Overall	6.90	**7.31**	4.69	21.67	22.83	**24.41**	4.26	**7.84**	4.32	**22.70**	18.30	21.58

FlairEL and Word2Vec embeddings achieved the best performance, with nearly 50% F1-score on NER.Spec.OCR. In contrast, GLiNER models exhibit poor results, with maximum F1-scores below 10% F1-score on either dataset. No model successfully identified E5 Event or E54 Dimension entities, consistent with their underrepresentation in the corpus (see Table 8).

BiLSTM-CRF-based models demonstrated robust performance on both datasets despite OCR-induced text degradation in NER.Spec.OCR. Strict matching performance remained low, especially for titles, events, and groups. Type matching exceeded strict matching, emphasizing the models' greater proficiency at entity type classification rather than entity boundary detection, amplified by the presence of multi-word entities (groups, place, and person average 3-4 words per mention).

Performance comparisons between NER.Spec.OCR and NER.Spec.Human show marginal gains for non-Word2Vec BiLSTM-CRF variants and substantial improvement for GLiNER-small on human-transcribed text. Other models either perform the best on OCR data or show comparable results across the two datasets. These findings indicate Word2Vec embeddings' robustness to OCR noise and the influence of reduced gold-standard annotations in degraded OCR text.

6.1.3 Execution Time We analyzed inference times for all NER models across the general-domain and domain-specific datasets by executing three runs for each model and dataset, as shown in Figure 7.

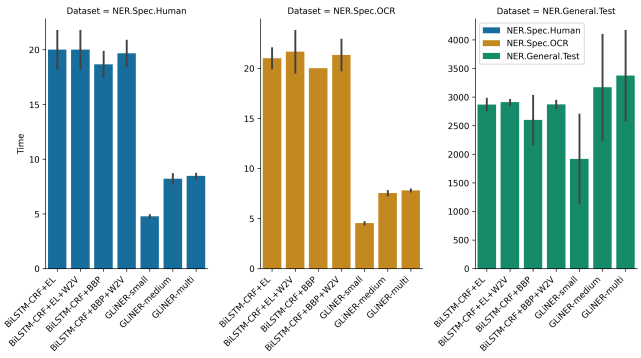


Figure 7. Inference times of NER models on general-domain and domain-specific datasets.

GLiNER models processed NER.Gen.Test (7,986 sentences) in batches, with speeds of 2.84 sent/s (small), 1.89 sent/s (medium), and 1.87 sent/s (multi), compared to BiLSTM-CRF models that processed datasets as documents. For domain-specific datasets, both model types processed the documents without batching. While GLiNER models were slower than BiLSTM-CRF models on the general-domain dataset with high variability, they were substantially faster on domain-specific data due to smaller data volumes.

6.2 Ontology-aligned RE

This section presents the RE results on the general-domain and domain-specific archival datasets.

6.2.1 General-domain Evaluation We evaluated the model on a general-domain dataset to assess their capability of extracting relation between named entities from Portuguese 20th-century digital archival records, as in the previous Section 6.1.2. Table 14 reports GLiREL performance on the RE.Gen.Test dataset.

Table 14. Performance of GLiREL model on the RE.Gen.Test dataset.

Approach	P	R	F1
Baseline GLiREL	0.49%	0.74%	0.59%
+ constraints	0.52%	0.74%	0.61%

The low strict matching scores likely stem from a high number of unconnected entities that generated false positives during relation extraction. The RE.Gen.Test dataset contains many named entities that lack corresponding gold relations, leading GLiREL to produce predictions across these dangling entities. Some documents also exceed token limits and get truncated, further degrading performance. The minimal improvement from ArchOnto-based entity type constraints (+0.02% F1) suggests that the model struggles mainly with entity pairing rather than type compatibility.

6.2.2 Domain-specific Evaluation We evaluated the model on domain-specific archival datasets to assess their capability of extracting relation between named entities from Portuguese 20th-century digital archival records, as in the previous Section 6.1.2. Table 15 report GLiREL performance with constraints on human-transcribed datasets with varying top-k and thresholding configurations.

Table 15. GLiREL performance on RE.Spec.Human for varying top-k and confidence thresholds.

Top-k	Threshold	Precision	Recall	F1-score
1	0.1	0.95%	19.35%	1.82%
	0.3	1.34%	19.35%	2.51%
	0.5	3.43%	19.35%	5.83%
	0.7	4.76%	9.68%	6.38%
	0.8	4.00%	3.23%	3.57%
6	0.1	0.43%	29.03%	0.85%
	0.3	1.18%	22.58%	2.24%
	0.5	3.06%	19.35%	5.29%
	0.7	4.76%	9.68%	6.38%
	0.8	4.00%	3.23%	3.57%

Results show that top-k=1 consistently outperforms top-k=6 on RE.Spec.Human, achieving the highest F1-core at threshold=0.7. Low thresholds maximize recall but drop precision with the introduction of false positives. Our next experiments use top-k=1 and thresholding=0.7

Results show that adding entity type constraints to relation extraction did not yied performance improvement for GLiREL on domain-specific data. The F1-scores remain below 10%, highlighting strict matching limitations. OCR-extracted text slightly dropped performance (-1.10% F1).

Relation extraction in the human-transcribed data achieved a low compatibility score of 4.65%. OCR-extrated text yielded a higher compatibility score of 10.72%, but with only two documents contributing non-zero scores (records 12, 13: 50% and 14.29%). This difference likely stems from OCR-based analysis covering just six documents instead of eleven documents. Low compatibility scores suggest that GLiREL might not be sufficiently reliable for relation extraction on archival documents.

7 Mapping Extracted Information to Linked Data

From the extracted entities and properties, we define mapping rules to populate the ArchOnto ontology by creating class, object and datatype property assertions from the identified entities (see Table 17) and their relations (see Table 18). We also establish naming conventions for creating individuals to ensure consistency (for example, instances use camel case), as shown in Table 19.

We present a SPARQL query template that generates assertions from the extraction of a person's name in Listing 1 using the example from Figure 1 (document 13). The variables *?personId=2* (an auto-incremented identifier) and *?personName="Jacques Valdour"* would require an additional counter query.

Using the best configuration of our top-performing BiLSTM-CRF+EL+W2V NER model paired with the GLiREL RE model (top-k=1, threshold=0.9), both pipelines correctly identified that the document titled "LA MENACE ROUGE" (*ARE2 Formal Title*) was authored by "Jacques

Table 16. Performance of GLiREL model on the domain-specific datasets.

Approach	Spec.Human			Spec.OCR		
	P	**R**	**F1**	**P**	**R**	**F1**
Baseline	6.35%	12.90%	8.51%	7.69%	7.14%	7.41%
+ constraints	6.35%	12.90%	8.51%	7.69%	7.14%	7.41%

Table 17. Mapping of extracted entities to property assertions in ArchOnto.

Entity	ArchOnto Triples
ARE2 Formal Title	E31 Document→P102 has title→ARE2 Formal Title ARE2 Formal Title→L2DO hasValue→DOE8 String DOE8 String→DOP7 stringValue→*value (xsd:string)*
ARE8 Role Type	PC14 Carried Out By→P02 has range→E21 Person* PC14 Carried Out By→P14.1 in the role of→ARE8 Role Type ARE8 Role Type→L2DO hasValue→DOE8 String DOE8 String→DOP7 stringValue→*value (xsd:string)*
E5 Event	E5 Event→P1 is identified by→E41 Appellation E41 Appellation→L2DO hasValue→DOE8 String DOE8 String→DOP7 stringValue→*value (xsd:string)*
E21 Person	E21 Person→P1 is identified_by→E41 Appellation E41 Appellation→L2DO hasValue→DO17 PersonName DO17 PersonName→DOP5 name→*value (xsd:string)*
E52 Time-Span	E52 Time-Span→P1 is identified by→E41 Appellation E41 Appellation→L2DO hasValue→DOE11 Interval DOE11 Interval→DOP6 startDateValue→*value (xsd:dateTime)* DOE11 Interval→DOP2 endDateValue→*value (xsd:dateTime)*
E53 Place	E53 Place→P1 is identifed by→E41 Appellation E41 Appellation → L2DO hasValue → DOE8 String) DOE8 String → DOP7 stringValue → *value (xsd:string)*
E54 Dimension	E54 Dimension → P90 has value → *value (xsd:string)* E54 Dimension → P91 has unit → E58 Measurement Unit
E74 Group	E74 Group → P1 is identified by → E41 Appellation E41 Appellation→L2DO hasValue→DOE8 String DOE8 String→DOP7 stringValue→*value (xsd:string)*

*Individuals can be E21 Person or E74 Group, but E39 Actor should not be directly instantiated, so we default to creating E21 Person

Table 18. Mapping of extracted relations to property assertions in ArchOnto.

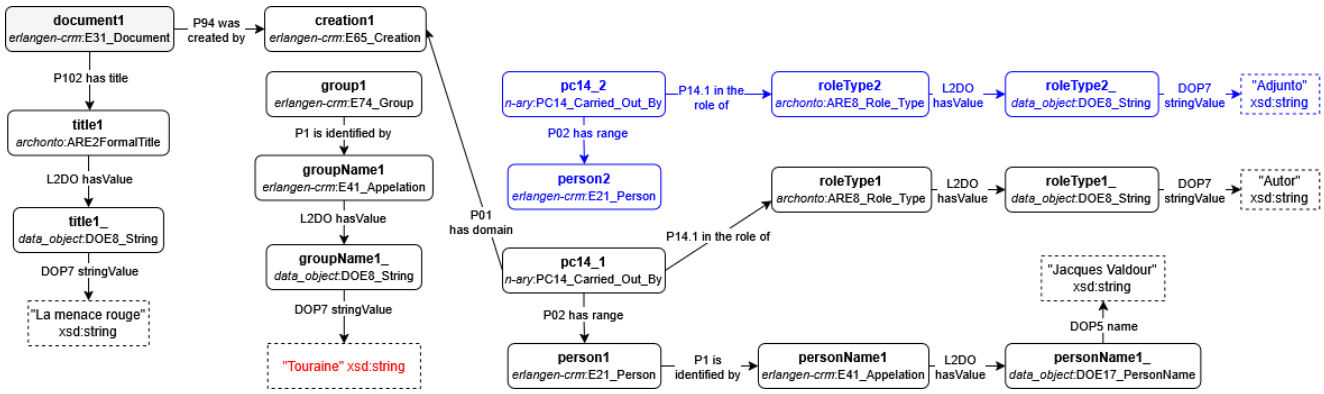
Relation	ArchOnto Triples
P4 has time-span	E7 Activity*→P4 has time-span→E52 Time-Span
P7 took place at	E7 Activity*→P7 took place at→E53 Place
P14.1 in the role of	E31 Document→P102 has title→E35 Title** E31 Document→P94 was created by→E65 Creation PC14 Carried Out By→P14.1 in the role of→ARE8 Role Type PC14 Carried Out By→P01 has domain→E65 Creation* PC14 Carried Out By→P02 has range→E39 Actor***
P53 has current or former location	E74 Group→P53 has current or former location→E53 Place
P74 has current or former residence	E74 Group→P74 has current or former residence→E53 Place
P89 falls within	E53 Place→P89 falls within→E53 Place
P107 has current or former member	E74 Group→P107 has current or former member→E39 Actor***
author of, director of, editor of, produced by	E31 Document→P102 has title→E35 Title** E31 Document→P94 was created by→E65 Creation PC14 Carried Out By→P14.1 in the role of→ARE8 Role Type ARE8 Role Type→L2DO hasValue→DOE8 String DOE8 String→DOP7 stringValue→"role xsd:string"**** PC14 Carried Out By→P01 has domain→E65 Creation* PC14 Carried Out By→P02 has range→E39 Actor***
date of creation	E74 Group→P107 has current or former member→E21 Person
has owner	E22 Human-Made Object→P52 has current owner→E74 Group E22 Human-Made Object→P128i is carried by→E31 Document E31 Document→P102 has title→ARE2 Formal Title

*Includes individuals of E7 Activity and its subclasses (e.g., E5 Event, E65 Creation, etc.); **Includes individuals that are subclasses of E35 Title (e.g., ARE2 Formal Title, ARE3 Supplied Title); ***Includes individuals that are subclasses of E39 Actor (e.g, E21 Person, E74 Group); ****Can be "autor", "editor", "diretor", or "produtor" depending on relation type

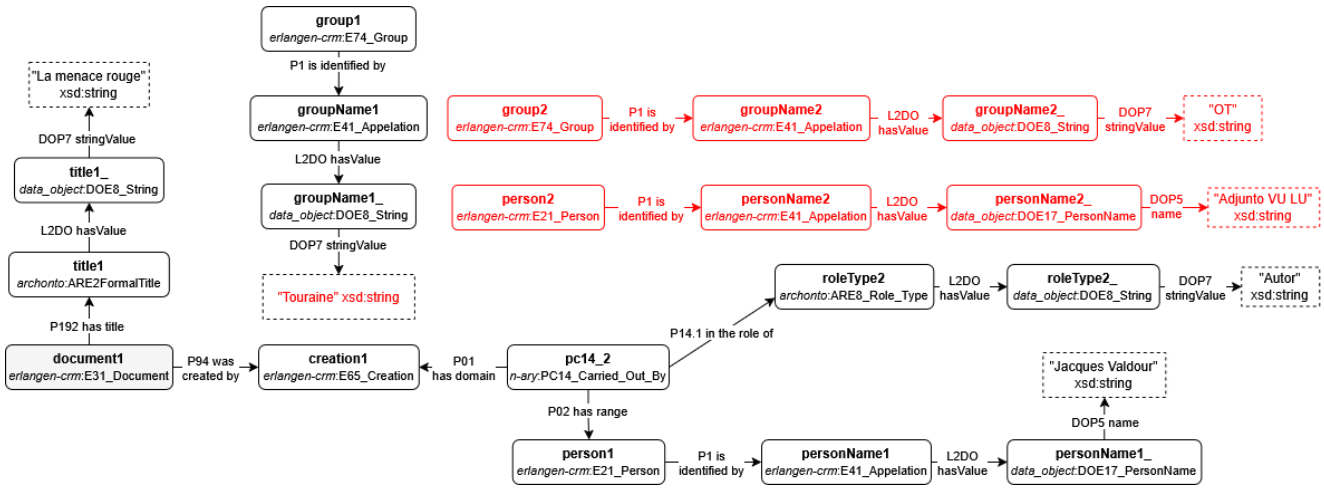
Valdour" (*E21 Person*). Human-made transcription failed to correctly identify the "ADJUNTO" (*ARE8 Role Type*) entity, while the OCR-extracted text incorrectly classified it as *E21 Person*. Figure 8 shows the ArchOnto representations for document 13 using extracted entities and relations, after post-processing to remove unrepresentable triples. Both pipelines captured the essential document authorship relationship despite NER missing the explicit "author" entity.

8 Discussion

Early sequence labeling architectures, such as BiLSTM-CRF with various embedding model variants, outperformed newer transformer-based sequence labeling models like GLiNER on NER tasks involving archival Portuguese texts



(a) Using human-made transcription.



(b) Using OCR-extracted text.

Figure 8. Automatic ArchOnto representations from digital representation of archival document in Figure 1. Red underlines extraction errors, and blue underlines missed predictions.

Table 19. Naming conventions applied to instantiations.

Convention	Description	Example
Camel case instantiation	Instances use camel case	roleType1 (ARE8 Role Type)
Individuals instantiation	Individuals are instantiated using their class name followed by an auto-incremented identifier that keeps track of the number of individuals of that class	place4 (E74 Group)
Appelations instantiation	Appellation of terms are instantiated using the term's followed by "name"	eventName1 (E41 Appellation)
DataObject class instantiation	Instances from DataObject ontology that are linked to CIDOC-CRM or the ArchOnto extension have names ending with an underscore	eventName1_ (DOE8 String)
Temporal entity instantiation	Appellations and DataObject entities related to temporal entities use the date format yyyy-mm-dd, which may include full or partial dates	1910-11-30 (E41 Appellation) 1910_ (DOE11 Interval)

(see Section 6.1.2). In response to RQ1, the BiLSTM-CRF variant combining FlairEL and Word2Vec embeddings achieved the best performance on the NER task for OCR-extracted text. Due to the limited availability of annotated data in Portuguese, we were unable to fine-tune a sequence labeling model for relation extraction, preventing a comparison between early sequence labeling architectures and transformer-based models for the RE task.

In response to RQ2, model performance on the NER task was consistently higher on general-domain Portuguese text compared to domain-specific archival text. In contrast, relation extraction performance dropped on general-domain data due to excessive false positives from the classification of numerous dangling named entities without corresponding gold relations. For example, entities like *E54 Dimension* have no associated properties or triples to extract in Re.Gen.Test, leading the RE models to predict relations where none exist.

Analysis from Sections 5.3, 6.1.2, and 6.2.2 allows us to answer RQ3. The results show that while OCR-induced noise hindered entity extraction by failing to recognize certain words, as was illustrated in Figure 6, where the OCR-extracted text does not contain the word "autor" (author), it did not generally degrade performance on entities available for recognition using BiLSTM-CRF-based models, even in the presence of gibberish text. This is evident by directly

```
Prefix erlangen: <http://erlangen-crm.org/200717>
Prefix data_object:
↪ <http://www.episa.inesctec.pt/data_object#>
Prefix ligacao:
↪ <http://www.episa.inesctec.pt/ligacao#>
Prefix record:
↪ <http://www.episa.inesctec.pt/archonto/record#>

# Extraction of person's name
INSERT DATA {
    # Class assertions
    record:person2 a erlangen-crm:E21_Person .
    record:personName2 a
    ↪ erlangen-crm:E41_Appellation .
    record:personName2_ a
    ↪ data_object:DOE17_PersonName .

    # Object property assertions
    record:person2
    ↪ erlangen-crm:P1_is_identified_by
    ↪ record:personName2 .
    record:personName2 ligacao:L2DO_hasValue
    ↪ record:personName2_ .

    # Datatype property assertions
    record:personName2_ data_object:DOP5_name
    ↪ "Jacques Valdour"^^xsd:string .
}
```

Listing 1: SPARQL query format to generate assertions from identified person's name.

comparing the results obtained with the NER.Spec.OCR and NER.Spec.Human and noting that there is not a performance drop in the OCR evaluation dataset. OCR-induced noise affected the ability to identify relations between named entities, with only six from the thirteen documents in the archival dataset having identifiable triples (see Table 8). Relation extraction performance with OCR-extracted text also had a small performance degradation.

Addressing RQ4, entity extraction quality showed minimal influence on downstream ontology-aligned relation extraction, likely due to the small sample size of identified relations between named entities in the domain-specific archival datasets. Even when NER performance varied, the scarcity of entity pairs limited the propagation of upstream errors to the RE phase.

Regarding RQ5 and using the running example, the end-to-end extraction pipeline demonstrated superior performance on clean archival texts mainly due to higher NER quality enabling accurate relation identification. Nevertheless, certain examples revealed relation extraction's ability to compensate for OCR-degraded NER quality, successfully recovering satisfactory ontology representation despite upstream entity recognition errors.

9 Conclusions and Future Work

This paper presented an end-to-end ontology-aligned information extraction pipeline for Portuguese archival texts, demonstrated with an example of the expected and actual output on an archival document. BiLSTM-CRF models for NER achieved solid performance (F1=40.86% on OCR-extracted text), while OCR noise minimally disrupted entity recognition with negligible downstream RE impact due to sparse entity pairng. GLiREL model for RE yielded poor results (F1=7.41% on OCR-extracted text), limiting its usability to generate reliable ontology-guided triples. Future designs could strictly enforce head and tail entity type constraints in ontology-guided RE pipelines, filtering invalid pairs pre-classification based on ontology schema.

As future work, it would be interesting to evaluate the ontology-aligned information extraction pipeline on benchmarks of archival annotated triples in various Latin languages beyond Portuguese. We could also compare the performance of the models explored in this paper against models trained on domain-specific archival data.

Acknowledgements

This work is financed by National Funds through FCT - Foundation for Science and Technology I.P., within the scope of the EPISA project - DSAIPA/DS/0023/2018. We would like to thank the archival professionals at DGLAB for their guidance on ontology-based archival descriptions.

References

[1] Hyvönen E. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*, volume 3. Morgan & Claypool Publishers, 2012. DOI:10.2200/S00452ED1V01Y201210WBE003.

[2] Candela G, Cuper M, Holownia O et al. A Systematic Review of Wikidata in GLAM Institutions: a Labs Approach. In Antonacopoulos A, Hinze A, Piwowarski B et al. (eds.) *Linking Theory and Practice of Digital Libraries*. Cham: Springer Nature Switzerland. ISBN 978-3-031-72440-4, pp. 34–50. DOI:10.1007/978-3-031-72440-4_4.

[3] Koch I, Ribeiro C and Lopes CT. ArchOnto, a CIDOC-CRM-Based Linked Data Model for the Portuguese Archives. In *Digital Libraries for Open Knowledge: 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020, Lyon, France, August 25–27, 2020, Proceedings*. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-54955-8, p. 133–146. DOI:10.1007/978-3-030-54956-5_10.

[4] Hawkins A. Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web. *Archival Science* 2022; 22. DOI:10.1007/s10502-021-09381-0.

[5] Candela G, Dobreva M, Alkemade H et al. A Use Case Lens on Digital Cultural Heritage, 2025. URL https://arxiv.org/abs/2509.08710. 2509.08710.

[6] CRM C. Definition of the CIDOC conceptual reference model. Technical report, ICOM/CIDOC CRM Special Interest Group, 2010. URL https://cidoc-crm.org/sites/default/files/cidoc_crm_version_5.0.2.pdf.

[7] Bruseker G, Carboni N and Guillem A. *Cultural Heritage Data Management: The Role of Formal Ontology and CIDOC CRM*. Cham: Springer International Publishing. ISBN 978-3-319-65370-9, 2017. pp. 93–131. DOI:10.1007/978-3-319-65370-9_6.

[8] Ferreira-Lopes P and González-Gracia E. Performing a systematic literature review on the implementation of the cidoc crm in cultural heritage. *J Comput Cult Herit* 2025; 18(4). DOI:10.1145/3771098.

[9] Clavaud F and Wildi T. ICA Records in Contexts-Ontology (RiC-O): a Semantic Framework for Describing Archival Resources. In *Linked Archives 2021: Proceedings of Linked Archives International Workshop 2021 co-located with 25th International Conference on Theory and Practice of Digital Libraries (TPDL 2021)*. CEUR Workshop Proceedings (CEUR-WS.org) (ISSN 1613-0073) : Free Open-Access Proceedings for Computer Science Workshops, 2021. pp. p. 79–92. URL https://enc.hal.science/hal-03965776.

[10] Clavaud F. Records in Contexts (RiC) aux Archives nationales de France : enjeux, réalisations, perspectives. In *Demi-journée d'information "RIC (Records in Contexts). Quels changements et quelles perspectives?"*. Lausanne (CH), Switzerland: Association vaudoise des archivistes (AVA). URL https://enc.hal.science/hal-03957469.

[11] Pandolfo L, Pulina L and Zielinski M. Arkivo: an ontology for describing archival resources. In *CILC*. pp. 112–116. URL https://www.academia.edu/download/101443823/paper12.pdf.

[12] Wang Z, Song Z, Yu G et al. An ontology for chinese government archives knowledge representation and reasoning. *IEEE Access* 2021; 9: 130199–130211. DOI:10.1109/ACCESS.2021.3112001.

[13] Vsesviatska O, Tietz T, Hoppe F et al. Ardo: an ontology to describe the dynamics of multimedia archival records. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. SAC '21, New York, NY, USA: Association for Computing Machinery. ISBN 9781450381048, p. 1855–1863. DOI:10.1145/3412841.3442057.

[14] Koch I, Teixeira Lopes C and Ribeiro C. Moving from isad(g) to a cidoc crm-based linked data model in the portuguese archives. *J Comput Cult Herit* 2023; 16(4). DOI:10.1145/3605910.

[15] Bountouri L, Damigos M, Drakiou M et al. The Semantic Mapping of RiC-CM to CIDOC-CRM. In Goh DH, Chen SJ and Tuarob S (eds.) *Leveraging Generative Intelligence in Digital Libraries: Towards Human-Machine Collaboration*. Singapore: Springer Nature Singapore. ISBN 978-981-99-8088-8, pp. 90–99. DOI:10.1007/978-981-99-8088-8_8.

[16] Oliveira C, Löw M and Henrique Bragato Barros T. Knowledge Organization Possibilities for Archives: Comparative Semantic Analysis Between CIDOC-CRM and RiC-CM. *Knowledge Organizationh* 2024; 51: 362–270. DOI:10.5771/0943-7444-2024-5-362.

[17] Giagnolini L, Koch I, Tomasi F et al. Comparative insights into semantic archival modelling: evaluating ric-o and archonto representation capabilities. *Journal of Documentation* 2025; 81(4): 1003–1031. DOI:10.1108/JD-12-2024-0310.

[18] Ji H. *Information Extraction*. Boston, MA: Springer US. ISBN 978-0-387-39940-9, 2009. pp. 1476–1481. DOI:10.1007/978-0-387-39940-9_204.

[19] Zheng S, Hao Y, Lu D et al. Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing* 2017; 257: 59–66. DOI:https://doi.org/10.1016/j.neucom.2016.12.075. Machine Learning and Signal Processing for Big Multimedia Analysis.

[20] Chiticariu L, Li Y and Reiss FR. Rule-based information extraction is dead! long live rule-based information extraction systems! In Yarowsky D, Baldwin T, Korhonen A et al. (eds.) *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 827–832. URL https://aclanthology.org/D13-1079/.

[21] Morwal S, Jahan N and Chopra D. Named entity recognition using hidden markov model (hmm). *International Journal on Natural Language Computing (IJNLC) Vol* 2012; 1. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3758852.

[22] Peng F and McCallum A. Information extraction from research papers using conditional random fields. *Inf Process Manage* 2006; 42(4): 963–979. DOI:10.1016/j.ipm.2005.09.002.

[23] Bundschus, Markus and Dejori, Mathaeus and Stetter, Martin and Tresp, Volker and Kröger, Peer. Extraction of semantic biomedical relations from text using conditional random fields. *BMC bioinformatics* 2008; 9: 207. DOI:10.1186/1471-2105-9-207.

[24] Panchendrarajan R and Amaresan A. Bidirectional lstm-crf for named entity recognition. In *Proceedings of the 32nd Pacific Asia conference on language, information and computation*. URL https://aclanthology.org/Y18-1061.pdf.

[25] Souza F, Nogueira RF and de Alencar Lotufo R. Portuguese Named Entity Recognition using BERT-CRF. *CoRR* 2019; DOI:10.48550/arXiv.1909.10649. 1909.10649.

[26] Shi P and Lin J. Simple BERT Models for Relation Extraction and Semantic Role Labeling, 2019. URL https://arxiv.org/abs/1904.05255. 1904.05255.

[27] Markus E and Adrian U. *Span-Based Joint Entity and Relation Extraction with Transformer Pre-Training*. IOS Press, 2020. DOI:10.3233/faia200321.

[28] Wang S, Sun X, Li X et al. GPT-NER: Named entity recognition via large language models. In Chiruzzo L, Ritter A and Wang L (eds.) *Findings of the Association for Computational Linguistics: NAACL 2025*. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-195-7, pp. 4257–4275. DOI:10.18653/v1/2025.findings-naacl.239.

[29] Hu Y, Chen Q, Du J et al. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association* 2024; 31(9): 1812–1820. DOI:10.1093/jamia/ocad259.

[30] Zhou W, Zhang S, Gu Y et al. UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition, 2024. URL https://arxiv.org/abs/2308.03279. 2308.03279.

[31] Orlando R, Huguet Cabot PL, Barba E et al. Retrieve, read and link: Fast and accurate entity linking and relation extraction on an academic budget. In *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics.

[32] Zaratiana U, Tomeh N, Holat P et al. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In Duh K, Gomez H and Bethard S (eds.) *Proceedings*

of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Mexico City, Mexico: Association for Computational Linguistics, pp. 5364–5376. DOI:10.18653/v1/2024.naacl-long.300. URL https://aclanthology.org/2024.naacl-long.300.

[33] Boylan J, Hokamp C and Ghalandari DG. Glirel – generalist model for zero-shot relation extraction, 2025. URL https://arxiv.org/abs/2501.03172. 2501.03172.

[34] Ehrmann M, Hamdi A, Pontes EL et al. Named entity recognition and classification in historical documents: A survey. *ACM Comput Surv* 2023; 56(2). DOI:10.1145/3604931.

[35] Vieira R, Olival F, Cameron H et al. Enriching the 1758 Portuguese Parish Memories (Alentejo) with Named Entities. *Journal of Open Humanities Data* 2021; 7. DOI:10.5334/johd.43.

[36] DigitArq. Memórias Paroquiais, 2021. URL https://web.archive.org/web/20211019034700/https://digitarq.arquivos.pt/details?id=4238720.

[37] Zilio L, Finatto MJ and Vieira R. Named Entity Recognition Applied to Portuguese Texts from the XVIII Century. In Trojahn C, Finatto MJ, de Paiva V et al. (eds.) *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Virtual Event, Fortaleza, Brazil, 21st March, 2022*, *CEUR Workshop Proceedings*, volume 3128. CEUR-WS.org, pp. 1–10. URL https://ceur-ws.org/Vol-3128/paper10.pdf.

[38] Semmedo JC. *Observaçoens Medicas Doutrinaes de Cem Casos gravissimos*. na officina de Antonio Pedrozo Galram, 1707.

[39] Lisboa JL, dos Reis Miranda TCP and Olival F (eds.) *Gazetas Manuscritas da Biblioteca Pública de Évora*, volume 1 (1729-1731). Évora: Publicações do Cidehu, 2002. DOI:10.4000/books.cidehus.3083.

[40] Paixão de Sousa MC. O Corpus Tycho Brahe: contribuições para as humanidades digitais no Brasil. *Filologia e Linguística Portuguesa* 2014; 16(esp.): 53–93. DOI:10.11606/issn.2176-9419.v16ispep53-93.

[41] Cunha LFdC and Ramalho JC. NER in Archival Finding Aids: Extended. *Machine Learning and Knowledge Extraction* 2022; 4(1): 42–65. DOI:10.3390/make4010003.

[42] ADB. Arquivo Distrital de Braga . URL http://pesquisa.adb.uminho.pt/.

[43] Santos J, Cameron HF, Olival F et al. Named entity recognition specialised for portuguese 18th-century history research. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*. Santiago de Compostela, Galicia/Spain: Association for Computational Lingustics, pp. 117–126. URL https://aclanthology.org/2024.propor-1.12/.

[44] Santos D, Seco N, Cardoso N et al. HAREM: An Advanced NER Evaluation Contest for Portuguese. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA), pp. 1986–1991. URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/59_pdf.pdf.

[45] Nothman J, Ringland N, Radford W et al. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence* 2013; 194: 151–175. DOI:10.1016/j.artint.2012.03.006.

[46] Efremova J, García AM, Zhang J et al. Towards population reconstruction : extraction of family relationships from historical documents. In *First International Workshop on Population Informatics for Big Data*. pp. 1–9. URL https://dmm.anu.edu.au/popinfo2015/papers/2-efremova2015popinfo.pdf.

[47] Chantaraj P, Rungrattanaubol J and Na-udom A. Historical Relation Extraction from Buddhist Temple Documents of the Lanna Kingdom. *Journal of Computer Science* 2019; 15(9): 1320–1330. DOI:10.3844/jcssp.2019.1320.1330.

[48] Christou D and Tsoumakas G. Extracting Semantic Relationships in Greek Literary Texts. *Sustainability* 2021; 13(16). DOI:10.3390/su13169391.

[49] Opitz J, Raclé C, Boros E et al. CLEF HIPE-2026: Evaluating Accurate and Efficient Person-Place Relation Extraction from Multilingual Historical Texts. *arXiv preprint arXiv:260217663* 2026; .

[50] Santos D, Mamede N and Baptista J. Extraction of family relations between entities. In *INForum*. Citeseer, pp. 9–10.

[51] Collovini S, Machado G and Vieira R. A Sequence Model Approach to Relation Extraction in Portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 1908–1912. URL https://aclanthology.org/L16-1301.

[52] Freitas C, Mota C, Santos D et al. Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese. In Chair) NCC, Choukri K, Maegaard B et al. (eds.) *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). ISBN 2-9517408-6-7, pp. 3630–3637. URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/412_Paper.pdf.

[53] Collovini S, Neto JFS, Consoli BS et al. IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)*. pp. 390–410. URL https://ceur-ws.org/Vol-2421/NER_Portuguese_overview.pdf.

[54] Akbik A, Blythe D and Vollgraf R. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1638–1649. URL https://aclanthology.org/C18-1139.

[55] Akbik A, Bergmann T, Blythe D et al. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 54–59. DOI:10.18653/v1/N19-4010.

[56] Santos J, Consoli B, dos Santos C et al. Assessing the Impact of Contextual Embeddings for Portuguese Named Entity Recognition. In *Proceedings of the 8th Brazilian Conference on Intelligent Systems*. pp. 437–442. DOI:10.1109/BRACIS.2019.00083.

[57] Lief E. *Deep Contextualized Word Embeddings from Character Language Models for Neural Sequence Labeling*. Master's thesis, Univerzita Karlova, Matematicko-fyzikální fakulta, 2019. URL http://hdl.handle.net/20.500.11956/105144.

[58] Segura-Bedmar I, Martínez P and Herrero-Zazo M. SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA: Association for Computational Linguistics, pp. 341–350. URL https://aclanthology.org/S13-2056.

[59] Tjong Kim Sang EF and De Meulder F. 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*. pp. 142–147. URL https://aclanthology.org/W03-0419.

[60] Tjong Kim Sang EF and De Meulder F. 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*. pp. 142–147. URL https://aclanthology.org/W03-0419.

[61] Dias M and Lopes CT. Consensual ArchOnto representation of 13 Portuguese Historical Archival Records based on their Digital Representations. Data set, 2023. DOI:10.25747/EADP-M943.

[62] Dias M and Lopes CT. Optimization of Image Processing Algorithms for Character Recognition in Cultural Typewritten Documents. *J Comput Cult Herit* 2023; 16(4). DOI:10.1145/3606705.