ORIGINAL ARTICLE

A novel NLP-driven approach for enriching artefact descriptions, provenance, and entities in cultural heritage

Sara Ferro[1] · Riccardo Giovanelli[1] · Madison Leeson[1] · Michela De Bernardin[1] · Arianna Traviglia[1]

Received: 3 February 2025 / Accepted: 13 June 2025

Abstract
Despite the availability of numerous open datasets on cultural heritage, limited research has focussed on structuring and normalising this type of data, particularly through the extraction of entities from unstructured texts. This step is crucial for enriching, analysing, and understanding these complex datasets. This study presents a procedure designed to streamline the creation of domain-specific datasets for training natural language processing models and evaluates their performance across three distinct datasets generated using this procedure. A zero-shot learning model, the Generalist and Lightweight Model for Named Entity Recognition, was assessed alongside pre-trained spaCy models on three datasets created in the framework of the European Union-funded Research Intelligence Technology for Heritage and Market Security project: one containing provenance information on artefacts from North American museums, another detailing stolen cultural goods in Romania, and a third with structured yet unclassified data on WWII-looted Polish art. Further training of spaCy models on these newly defined datasets revealed that fine-tuned models significantly outperform their non-fine-tuned counterparts, with the best results from the Transformer model fine-tuned on provenance data. This success can be largely attributed to the standardised conventions in provenance research. In contrast, the model fine-tuned on descriptive information performed poorly, likely due to extensive descriptions containing non-essential data that increased model uncertainty. This work highlights the potential of automating entity extraction to build knowledge graphs for cultural object databases, enabling advanced analytical approaches such as Network Analysis.

Keywords Natural language processing · Information retrieval · Unstructured data processing

1 Introduction

Unstructured data refers to information that does not have a pre-defined format or organisational model. Unlike structured data, which can be easily organised into standardised keys and values, unstructured data encompasses a wide variety of formats such as text, images, audio, video, and social media posts. This type of data lacks a consistent schema, making it more challenging to clean and analyse, especially when dealing with large volumes that require automated processing. Despite these challenges, unstructured data are a rich source of valuable insights, requiring advanced tools and technologies to manage its complexity and extract meaningful information for data analysis and knowledge enhancement [1].

https://doi.org/10.1007/s00521-025-11449-2

Published online: 13 August 2025

The European Union (EU)-funded project Research Intelligence and Technology for Heritage and Market Security (RITHMS),[1] developed by a consortium of researchers, data analysts, Law Enforcement Agencies (LEAs), and scientific institutions across Europe, aims to build a digital platform to assist LEAs in combating cultural goods trafficking. This platform will facilitate the identification and analysis of relationships between criminal and non-criminal actors involved in the circulation and transfer of ownership of cultural property. At its core, the platform utilises Social Network Analysis (SNA), which applies graph theory to unravel social systems, their properties, and underlying mechanisms. The network structure, represented as a mathematical adjacency matrix, is particularly effective in handling large volumes of entities and their relationships by organising them as nodes and connected edges [2].

To achieve its goals, the RITHMS project leverages a comprehensive Knowledge Graph (KG) database. A KG is a large-scale graph that consolidates data from diverse sources, aimed at organising and conveying knowledge about the real world. In the graph, nodes represent entities such as individuals, organisations, dates, and locations, while edges denote relationships among them. The data are structured using graph-based models, often incorporating ontologies or rules to represent quantified statements. This approach allows for the organisation of specific knowledge domains and the accumulation of new knowledge through inductive or deductive reasoning [3].

Conventional SNA typically focuses on a single type of entity (e.g. ‘person’) and a single type of relationship (e.g. ‘friendship’). However, real-world systems often involve entities of the same type connected by diverse types of relationships, as well as relationships connecting diverse types of entities (e.g. organisations, locations, and dates) [4]. To address this complexity, multiple perspectives can be modelled through multiplex, heterogeneous, or multiplex heterogeneous networks [5], which offer a more effective approach for analysing intricate systems.

In the context of the cultural property trade, a multiplex heterogeneous network structure consists of several interconnected layers, each sharing the same set of nodes but linked by different relationships. Entities within this network may include individuals (e.g. collectors, intermediaries, looters, and restorers), organisations (e.g. auction houses, museums, and companies), dates, locations, events, and cultural goods themselves. Relationships among and within these layers can encompass ownership, transactions, money transfers, restoration processes, partnerships, friendships, legal disputes, loans, exhibitions, historical connections, and more.

Currently, no comprehensive database captures all these nuances. To address this gap, the RITHMS project has developed a set of automated modules to gather and structure data for this complex, interdisciplinary network. These modules draw data from a variety of sources, including existing open-source datasets, news articles, mobile traffic data, satellite imagery analysis reports (in connection with the EU’s Copernicus Earth Observation program), databases of stolen objects, online forums, and websites of galleries and auction houses. While some of the collected data are already well-structured or semi-structured and can be easily normalised through simple rule-based methods, other raw, unstructured data require more sophisticated information extraction processes to be transformed into structured, semantically and numerically representable data.

Natural Language Processing (NLP) tools play a key role in this transformation. For instance, Named Entity Recognition (NER) identifies and categorises the semantic types of entities in text (e.g. person, organisation, location, etc.), while Relation Extraction (RE) infers relationships among them. By leveraging these tools, knowledge extracted from unstructured data can be converted into semantic triples—three-part statements that follow the structure ‘subject’, ‘predicate’, ‘object’—for example, ‘Person A’ ‘is related to’ ‘Person B’ [6]. In graph theory, these triples represent adjacent nodes and their incident edges [7], allowing the effective capture and representation of semantic information.

Early NLP systems relied heavily on manually defined domain-specific features to achieve satisfactory performance. However, with the advent of Deep Learning models and the availability of vast, labelled corpora of texts across the web, the performance and precision of these tools have improved significantly, especially for

[1] https://rithms.eu/.

English-language content [8, 9]. Despite these advancements, when applied to information that differs significantly from the original training dataset in terms of domain, language, and/or morphological structure, the available tools often underperform and require costly and time-consuming domain adaptation [10]. Additionally, challenges such as ambiguity and implicitness of relevant information, especially in event extraction, further complicate the application of NLP models in specialised fields [11].

The present study aims to: a) propose a faster and more efficient workflow for adapting or training NLP models without the need for extensive manual annotation, leveraging existing models like Meta's LLAMA and Explosion's spaCy; b) generate high-quality labelling in the absence of domain-specific training data; c) evaluate the performance of the models on the new datasets; and d) evaluate and compare the quality of the entities extracted in the context of cultural heritage cataloguing and provenance research. The fine-tuned models have been made available on GitHub at the link https://github.com/IIT-CCHT/NER-models-CH-datasets.

2 Materials and methods

2.1 Data collection

The data utilised in the present study were collected from databases of missing, stolen, protected, and unprovenanced cultural goods. Custom-built web scrapers were developed to extract and pre-process information from thirty repositories for ingestion in the RITHMS platform. These repositories included fourteen national and international databases focussing on problematic (e.g. unprovenanced, stolen, missing) or protected objects, fourteen repositories dedicated solely to WWII-looted cultural goods, and two databases containing information on provenance and individuals. Target sources were chosen due to their reliability, scope, and accessibility. While the project focuses primarily on European databases, it also incorporates sources from North and South America due to their relevance in tracking objects at heightened risk of trafficking or illicit import into European markets. Notable examples include the database managed by the Association of Art Museum Directors (AAMD), the Stolen Cultural Assets database managed by Chile's National Cultural Heritage Service, and the Iraq Museum Database developed by the University of Chicago's Oriental Institute.

From the initial thirty data sources, over two million entities—including individuals, objects, dates, locations, and organisations—were gathered and identified. This extensive data collection resulted in an unprecedented, consolidated dataset of stolen, protected, and unprovenanced objects, preliminary analyses on which have yielded promising results [12, 13].

While some properties within each entity were already structured or required minimal deterministic effort to extract, others consisted of unstructured or quasi-structured texts containing rich semantic information. Depending on the source, data needing further processing were broadly subdivided into three categories: provenance information, descriptive details, and unclassified mixed information.

Provenance information concerns texts that relate an object's ownership history, typically following a format like: 'Person Name purchased from Organisation, City, Date; then to Second Person, City, Date; acquired by Museum, Location, through donation on Date.' The second category—descriptive details—concerns free text that contains details of the artefact, for example: 'oil painting on canvas, framed, in good condition, depicting a landscape with artist's signature in bottom-left corner'. The final category, of 'unclassified mixed information', concerns relatively short samples like 'ABC Museum,' 'European art market,' and 'John Smith,' which must be classified as organisation, location, and person, respectively. Each of these three categories is outlined in greater detail in the following section.

https://doi.org/10.1007/s00521-025-11449-2

2.2 Datasets

This paper examines three distinct datasets (cf., Table 1): the AAMD Object Registry, the Obiecte Furate database of the Romanian Police, and the Polish Ministry of Culture's Catalogue of Wartime Losses. They encompass the three types of data requiring additional processing: provenance information, detailed object description, and misclassified or unclassified values, respectively.

At the time of writing, the AAMD Object Registry stores records on over 2300 art objects acquired by AAMD member museums since 2008. These objects lack pre-1970 provenance, referring to documentation of their presence on the international market before the adoption date of the UNESCO Convention aimed at preventing cultural goods trafficking. The original database includes various fields: the current museum, accession number, title, artist or producer, size, creation date, credit line (indicating the previous owner and the channel through which the museum acquired the object, such as by gift or purchase), country and culture of origin, object type, material or technique, provenance, exhibition and publication history, and—specifically for Nazi-era objects—details on the resolution of restitution claims. Each record also provides a direct link to the institution currently holding the object and a reference to the specific AAMD guideline permitting its acquisition despite insufficient provenance.

In this dataset, the unstructured text within the 'provenance' field requires processing via NLP models to extract the entities and relationships associated with the artefact's ownership history. This field provides critical information for the RITHMS project, enabling the identification of individuals and organisations connected to specific artefacts. Provenance information typically follows a chronological structure, detailing events from the earliest to the most recent, separated by semicolons or periods. This format reflects the practices of data providers (such as museums and galleries), who adhere to established conventions in provenance research. Although explicit verbs may be absent in the sentences, the implied action (typically acquisition or transfer of ownership) is clear. The data are classified as 'quasi-structured' as they adhere to a consistent format but contain free-text elements that require specialised extraction and classification of distinct entities.

The second data source is the Obiecte Furate database (ROF), managed by the Romanian Police. At the time of writing, the database holds records on approximately 900 stolen cultural objects based on police reports filed by individuals and institutions. The fields retrieved from each record are object category (e.g. 'Pictures,' 'Decorative Art,' 'Book'), title, reference number assigned by the Romanian Police, date of theft, date of registration in the database, and description.

The 'description' field, which is the focal point of this study, contains valuable details such as the artist, year of creation, place of publication (for books), material, size, weight, and previous location. During the data collection phase, deterministic pre-processing methods were applied to this unstructured description to extract and organise relevant details. However, these methods were limited by pre-defined rules, resulting in the loss of meaningful

Table 1 Datasets specifications

Dataset (language)	Field to process	# Of samples	# Of unique values	Type
AAMD object registry (English)	Provenance	2336	1268	Quasi-Structured
Obiecte furate (Romanian)	Description	890	866	Unstructured
Catalogue of wartime losses (English/Polish)	Author/workshop/School	4214	1026	Quasi-Structured

The content of the fields is considered unique if it varies by one or more characters. The language utilised in the datasets is also specified data not captured within these constraints. This limitation underscores the need for an NLP-based approach to enhance data extraction.

The database is in Romanian, and translation is required using the Google Translator library from the Deep-Translator Python package before proceeding with entity extraction through NLP. Like the AAMD Object Registry, the sentences in this dataset often lack explicit verbs, with actions implied and inferable. Unlike the AAMD dataset, however, the ROF dataset does not adhere to standardised conventions for data reporting. The high variability in data formatting and content values led to its classification as unstructured.

The final dataset analysed in this study originates from the Catalogue of Wartime Losses, a database of art objects maintained by the Division for Looted Art (DLA) within Poland's Ministry of Cultural and National Heritage. This database records items looted from Poland or sold under duress during World War II. It stores information on over 11,000 objects, contributed by both private individuals and museums and includes categories such as 'Archaeology,' 'Ceramics,' 'Glass,' and 'Painting.' Each record can contain details such as title, artist or producer, inventory number, object type and category, date of creation, material and technique, size, weight, signature, heir or legal owner, and related keywords.

The current study focuses on the data provided in the 'Author / School / Workshop' field, which encompasses multiple entity types, including individuals, organisations (e.g. schools or workshops), and locations. Distinguishing among these types is critical for accurate tagging and mapping in the resulting KG. Proper classification facilitates the association of distinct entities, enabling the construction of meaningful relationships in the processed database. Since the data are available in both Polish and English, English was selected as the primary language for processing to ensure consistency, as well as compatibility with the other datasets (AAMD and ROF). Although this field lacks structured sentences, its quasi-structured nature allows NLP models to classify and extract relevant entities effectively.

These three datasets were selected as representative examples of the data relevant to the RITHMS project. The AAMD dataset provides provenance information written in a format that has become a de facto standard, commonly adopted by museums and auction houses. The ROF database was chosen to represent object description databases, as it offers comprehensive details, not only about the artist and materials used but also about the depicted subject matter, which is often omitted in other databases. Finally, the DLA dataset serves as an example where fields are reported too broadly, necessitating further refinement to narrow down the classification.

As highlighted, the datasets used in this research exhibit significant diversity due to the unique structure of each data source. They range from quasi-structured entries with concise textual information to datasets with extensive free-text fields; necessitating fine-tuned NLP approaches for information extraction.

2.3 The named entity recognition models utilised

The absence of pre-existing labelled datasets tailored to cultural heritage object descriptions or provenance information presents a significant obstacle to creating automatic entity extraction models. This issue is further compounded by the substantial time and financial resources required for manual data labelling. To overcome this, a strategy was implemented to first generate annotated data and subsequently evaluate its accuracy through manual validation. This procedure leveraged pre-trained NLP models for named entity extraction. While traditional NLP NER models are effective, they are typically limited to pre-defined entity types. In contrast, Large Language Models (LLMs) offer greater flexibility by extracting any kind of entity through natural language instructions.

To identify the most effective solution, the performance of advanced generative LLMs capable of labelling raw data without supervision was compared. Recent advancements in generative models, equipped with extensive knowledge bases, have demonstrated their capability to efficiently tackle a wide range of NLP tasks. At the time of writing, the leading model available was Llama3 from Meta, which provides a smaller and more efficient model than GPT-4 (a state-of-the-art LLM developed by OpenAI [14]). Moreover, unlike GPT-4, it is freely available. Meta has provided a comparison between the best LLM and Llama3, showcasing that Llama3 performs the best in many common tasks [15]. Llama3 is Meta AI's largest family of LLMs. Two versions of Llama3.0 have been released: a smaller one of 8B parameters and a greater one of 70B parameters. Two different versions of these models have been released: 'Instruction-Tuned Models' designed for assistant-like chat interactions and training, and the general 'Pre-trained Models' which can be fine-tuned for various applications. Given Llama3's better performance and accessibility, the 8B parameter 'Instruction-Tuned Model' was selected to label a subset of the datasets, due to constraints of computational resources which could not load the 70B one.

The chosen model was 'instructed' to get the desired annotations through a series of different prompts designed to extract named entities from the input text. After some attempts, the most effective way to interact with Llama3 was found to be asking for one specific class of entities at a time. As a result, the query.[2] was to return labelled data with one single class of tags per query, and this process was repeated for each class to ensure full labelling across all tags.

While Llama3 was used to create a labelled subset of the datasets, smaller, specialised models for NER were subsequently fine-tuned using the tagged dataset. Several Python libraries have been released to perform NER, including spaCy [16], Huggingface [17], Flair, CoreNLP, and Scikit-learn, to name a few. Among these, spaCy, developed by Explosion, is one of the most widely used libraries for NER due to its high performance and fast, easy-to-use, production-ready APIs. Explosion also provides both open-source implementation and 'open-weight models' of different kinds, based on either Convolutional Neural Networks (CNNs) or Transformers like BERT and ROBERTa [18–20]. SpaCy English models include three CNN-based pipelines of different sizes: 'en_core_web_sm' (12 MB), 'en_core_web_md' (40 MB), and 'en_core_web_lg' (560 MB). Additionally, a model based on the Transformer architecture was released, 'en_core_web_trf' (436 MB), which by default includes the RoBERTa Transformer.

This study examined two of the largest models from spaCy: the CNN-based model ('en_core_web_lg') and the Transformer model ('en_core_web_trf'), as they are known to offer the best performance for named entity recognition [16]. The 'en_core_web_lg' achieves a precision, recall, and F1-score of 0.85, 0.86, and 0.85, respectively, while the equivalent scores for 'en_core_web_trf' are 0.90, 0.90, and 0.90. Despite these promising metrics, the direct application of these pre-trained models to the RITHMS datasets yielded sub-optimal results. This is primarily due to the significant differences between the RITHMS datasets and the data used to train the models, such as OntoNotes 5. The RITHMS datasets feature shorter contexts and distinct sentence structures, like the challenges faced by NER models applied to X posts [21]. In contrast, OntoNotes 5 consists of a vast array of data, including news articles, conversational telephone speech, weblogs, Usenet newsgroups, broadcasts, talk shows, web data, Wikipedia, and various book corpora. The hyperparameters used to train spaCy models were primarily the default settings recommended by Explosion[3]. In addition to spaCy models, the zero-shot model GLiNER (Generalist and Lightweight Model for Named Entity Recognition) was utilised. GLiNER, which leverages a bidirectional type of Transformer, enables parallel output computations, making it an attractive alternative for performing entity labelling without needing training data. This model was assessed to determine its potential for labelling the RITHMS datasets, given its ability to handle entity extraction in a zero-shot, domain-independent manner.

[2] The Llama3's 'system' behaviour was set to *"You are a very experienced Natural Language Processing tool. Recognise entities with label 'LABEL/CLASS'* Moreover, we passed an example of the input text to the 'user' role together with the output we would like to have (list of entities recognised of the specific CLASS *[start character index, end character index, CLASS]*) to the 'assistant' role.

[3] https://spacy.io/usage/training#quickstart. We reduced the encoder size of the large model (i.e. *spacy.MaxoutWindowEncoder.v2*) from 256 × 8 to 96 × 4 to lower memory use and approximate the Transformer model size. However, the overall size stayed similar since most space is taken by large static vectors, which improve accuracy, so we kept the rest unchanged. Further details on the models can be found in https://github.com/IIT-CCHT/NER-models-CH-datasets.

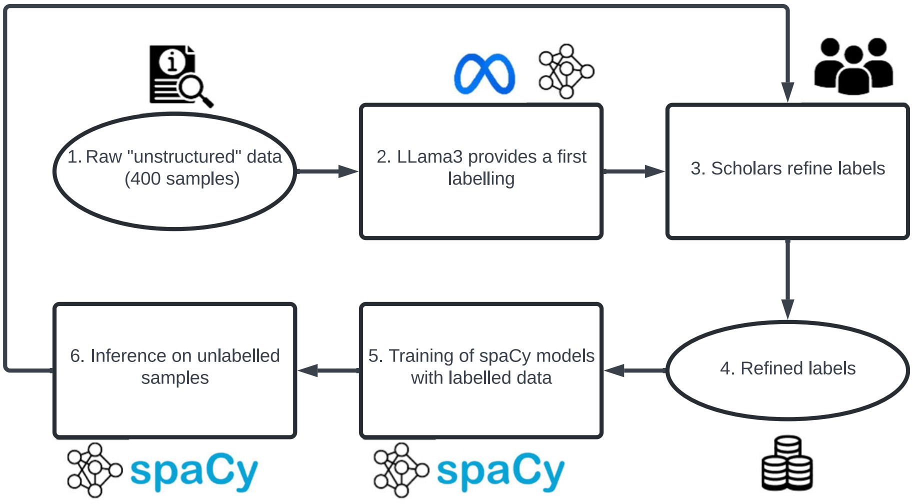

1. Raw "unstructured" data (400 samples)
2. LLama3 provides a first labelling
3. Scholars refine labels
4. Refined labels
5. Training of spaCy models with labelled data
6. Inference on unlabelled samples
spaCy
spaCy

Fig. 1 Pipeline to create the dataset to train Machine Learning models for NER

2.4 Creation of the labelled datasets

The annotation process was an iterative procedure (cf., Fig. 1), performed for each dataset separately.

Llama3 was initially used to label a subset of 400 texts from each dataset. While larger sample are often recommended—sometimes starting at 1, 000 lines—this sample size was deemed sufficient to represent the data and effectively capture the semantic patterns likely to be encountered by the models. Subsequently, experts in computer science and cultural heritage manually refined the classifications to ensure accuracy. The resulting datasets were then used to train spaCy models (cf., Fig. 1), specifically the large English model ('en_core_web_lg') and Transformer-based ('en_core_web_trf') model. Finally, the best-performing fine-tuned model, selected from the large English and Transformer-based models, was used to label the remainder of the datasets. This other part of the dataset, which was annotated using the best model, was then refined by experts in computer science and cultural heritage to provide accurate annotations.

The entity tags used to annotate data were derived from the standard set found in spaCy models, with additional domain-specific tags tailored to the project. To facilitate the understanding of similarities and differences among the datasets Fig. 2 provides a visual representation[4]—a Venn diagram—that illustrates both the shared tags and those unique to each dataset.

In the case of the AAMD database, the labels used are a subset of the classical: ORG, PERSON, GPE, LOC, DATE, EVENT, WORK_OF_ART and CARDINAL. During the annotation and validation phases, a standardised list of conventions was created to ensure consistent tagging. For instance, organisations (ORGs) were extended to include families, private collections, and surname-based groups (e.g. 'the Smiths'). For events (EVENTs), verbs such as 'acquired' and 'purchased' were also classified as event-related. Moreover, OTHER_QUANTITY was introduced to indicate numbers defining percentages and quantities of artefact pieces. Finally, given the frequent occurrence of terms like 'certificate of authenticity' or 'gift agreement' in provenance texts, a DOCUMENT tag was added to capture such references.

In the case of the ROF dataset, the labels used are again a subset of the classical spaCy tags, with additional domain-specific labels. These include ORG, PERSON, GPE, LOC, DATE, EVENT, MONEY, WORK_OF_ART, QUANTITY, and CARDINAL. For PERSONs, individuals depicted in artworks (e.g. portrait subjects or religious figures) were excluded, as they were not relevant to the context of the RITHMS project. Furthermore, additional tags were incorporated to capture specific details, such as: CONDITION (state of conservation), MATERIAL/TECH (material or technique), DESCRIPTION (free descriptive text that does not fall under another

[4] The same kind of representation will be provided for each dataset.

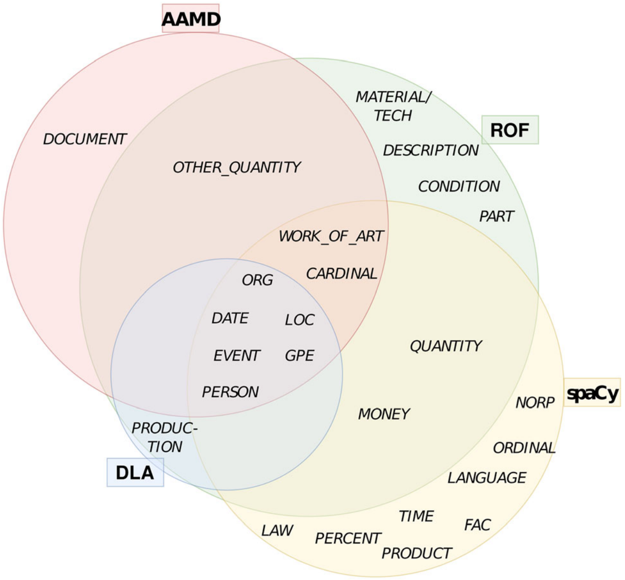


Fig. 2 Entity tags used compared with the spaCy tags

entity type), PRODUCTION (e.g. Roman, Impressionist, Belgrade workshop), and PART (specific parts of an object) and OTHER_QUANTITY (numbers defining percentages and quantities of artefact pieces). One of the most challenging aspects was determining what should be labelled with the DESCRIPTION tag. In some cases, the description includes material-related information—for example, ‘Tempera on wood presents an inscription in *ochre*, in the Greek language’. Although compositional details may appear within the descriptive text, we chose to retain the primary material information under the MATERIAL/TECH tag, while assigning all remaining contextual or interpretive content to the DESCRIPTION tag. Moreover, the content of the descriptions varies: in some cases, the text begins with the title of the opera, which we tag as WORK_OF_ART. In other cases—such as with coins—the title may be absent. In these instances, we instead tag terms like ‘ducat’ or ‘coin’ as WORK_OF_ART, as they function as the primary identifiers within the description. To maintain consistency, we assign a unique WORK_OF_ART tag to each description.

Lastly, the DLA dataset required comparatively fewer tags, and no additional domain-specific labels were added. Since the data from the Catalogue of Wartime Losses was relatively structured and primarily required only the classical spaCy NER labels were applied: PERSON, ORG, GPE, LOC, DATE and EVENT together with PRODUCTION (the only tag that is not part of spaCy’s standard tag set).

To assess the capabilities of the new models, each dataset was divided into three groups, comprising 70% for training, 20% for validation, and 10% for testing. Table 2 reports the resulting distribution of the datasets.

3 Results and discussion

Once the complete labelled datasets were obtained—the so-called “ground truth” (GT)—the qualitative results of the best spaCy models were analysed as they were to assess whether classical pre-processing techniques could yield better results.

When using off-the-shelf spaCy models (also referred to as ‘original spaCy models’), several issues were encountered, which tended to fall into one of two categories. The first common error was data misclassification, for example a PERSON name was classified as an ORG. The second was that certain entities were classified correctly but had missing or extraneous parts, for instance, ‘John’ in ‘Professor John Stone’ was correctly classified as PERSON but the classification performed by the model missed the surname ‘Stone.’ Conversely, it might classify the entire string ‘Professor John Stone’ as PERSON, including the role ‘Professor,’ which should not be labelled as such.

Table 2 Distribution of training, validation, and testing samples from each dataset

Dataset (language)	Field to process	# Of training samples	# Of validation samples	# Of test samples
AAMD object Registry (English)	Provenance	887	253	128
Obiecte furate (Romanian)	Description	606	173	87
Catalogue of Wartime losses	Author / Workshop / School	718	205	103

Pre-processing steps commonly employed to enhance the performance of NLP models were applied to evaluate their effectiveness in improving the entity recognition of the datasets under analysis. These included removing punctuation marks and titles (e.g. 'Mr.') and eliminating stop words (e.g. 'the' and 'and'). However, the removal of punctuation led to an even higher misclassification error. This was the case for both the 'en_core_web_lg' and 'en_core_web_trf' models. For instance, in the provenance 'John Stone, Jersey City, N.J, 1980–2003; Anna and John. Grae, New York, 2003–2010; Yale University Art Gallery, New Haven, Conn.', removing the punctuation marks led to the entire string 'Yale University Art Gallery New Haven' being wrongly classified as an ORG. Conversely, when retaining the punctuation marks, 'New Haven' was correctly classified as a separate entity distinct from 'Yale University Art Gallery,' and the two were rightly classified as GPE and ORG, respectively. A similar error resulted from removing punctuation from the following provenance sample: '(Art Treasures Gallery), Hong Kong; Purchased by Mr. Stone [b. 1942] and Henry Stain [b. 1930], 1997, Englewood, CO; Gifted to the Denver Art Museum, 2018 [1] on loan since 2016.' In this case, the removal of the punctuation led to the entire string 'Mr. Stone b 1942' being wrongly classified as a PERSON. Furthermore, as is clear from the examples provided, certain punctuation marks (such as semicolons and periods) delineate the beginning and end of distinct events, which is a widespread practice when writing the provenance of artefacts. Consequently, the decision was made to retain punctuation, which resolved these issues.

Similarly, the removal of honorifics such as 'Mr.', 'Ms.', and 'Mrs.' led to names being incorrectly classified as organisations (ORG) instead of individuals (PERSON). This issue was particularly pronounced in the large English model compared to the Transformer-based model. For example, the name 'Red' was misclassified as an ORG after honorifics were removed from the following sentence: 'According to a signed and notarized statement by Mr. Red supplied to Sotheby's, New York (May 11, 2012), he acquired the figures in Nayarit in 1946 and brought them to California in 1951.' Importantly, this misclassification did not occur when titles were retained. The same issue was seen in the example, 'Acquired by John Stone, London, at Portobello Road market, London, ca. 1969-1971. Consigned by Mr. Stone to Red, London, April 28, 2010, lot 20. Sold to Rhea Gallery, Zurich. Sold by Rhea Gallery, Zurich, to the Museum' where 'Stone' without title was incorrectly classified as an ORG. Notably, this error was particularly evident when only surnames were provided, while the presence of the first name typically led to the correct classification as a PERSON.

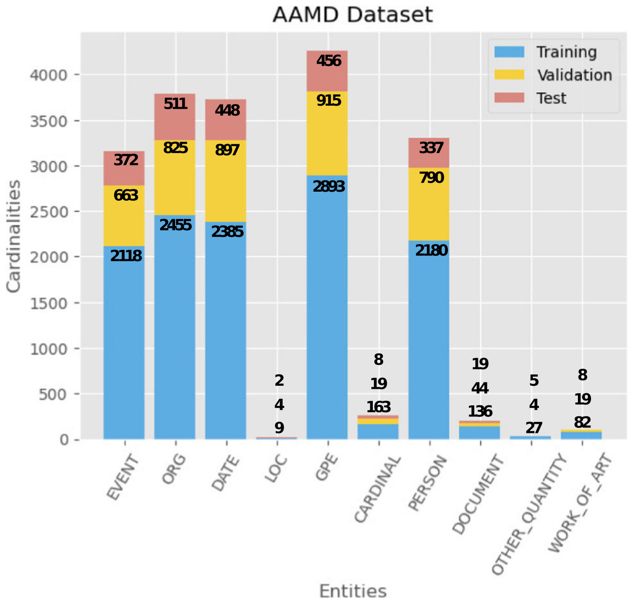
The removal of stop words also proved detrimental to the performance of NLP models, as demonstrated in the following example: '(Arte Primitivo), New York; purchased by Theodore and John T. Mohar, 1973; by descent to the Mohar family; consigned to (Arte Primitivo), New York; purchased by Gordon DeBiasi, Birmingham, MI, 2006; gifted to the Denver Art Museum, 2014.' Unsurprisingly, the removal of the word 'and' resulted in the name 'Theodore' being classified incorrectly. Also consider this example: 'Ex coll. Robert Stone, United States, said to have been purchased from Albert Rayan (1928-1973), New York, New York. Said to have been sold by Stone to a London dealer. Ex coll. Robert Stone, said to have been purchased from an English dealer, New York, New York, November 1999'. In this case, removing stop words led to 'Stone London' being misrecognised as a single entity.

These examples highlight the useful role that punctuation, honorifics, and stop words play in maintaining the semantic integrity of text and ensuring accurate classification. While the removal of these elements is considered a best practice to enhance model efficiency, our results demonstrate that it can result in significant inaccuracies and

Table 3 Results of the spaCy English and the GLiNER models on the AAMD dataset. Bold indicates the best result for each metric, while italics denotes the second-best value

Model	Entity	Precision	Recall	F1-score
Original spaCy large English model	DATE	71.74%	73.66%	73.66%
	PERSON	67.51%	71.51%	71.51%
	GPE	**92.54%**	**92.54%**	**92.54%**
	ORG	69.04%	37.96%	37.96%
	EVENT	0.00%	0.00%	0.00%
	CARDINAL	27.59%	47.06%	47.06%
	LOC	0.00%	0.00%	0.00%
	WORK_OF_ART	0.00%	0.00%	0.00%
Original spaCy Transformer-based English model	DATE	76.94%	77.46%	77.20%
	PERSON	84.53%	66.47%	74.42%
	GPE	**92.54%**	**92.54%**	**92.54%**
	ORG	70.67%	51.86%	59.82%
	EVENT	0.00%	0.00%	0.00%
	CARDINAL	37.25%	55.88%	44.71%
	LOC	0.00%	0.00%	0.00%
	WORK_OF_ART	0.00%	0.00%	0.00%
GLiNER large English model (Transformer-based)	DATE	74.63%	55.80%	63.86%
	PERSON	68.47%	**63.80%**	**66.05%**
	GPE	0.00%	0.00%	0.00%
	ORG	**100.00%**	0.20%	0.39%
	EVENT	23.64%	3.49%	6.09%
	CARDINAL	18.18%	5.88%	8.89%
	LOC	0.00%	0.00%	0.00%
	WORK_OF_ART	0.00%	0.00%	0.00%
	DOCUMENT	33.33%	31.58%	32.43%
	OTHER_QUANTITY	0.00%	0.00%	0.00%
Fine-tuned spaCy large English model	DATE	93.27%	92.86%	93.06%
	PERSON	90.70%	92.58%	91.63%
	GPE	**97.14%**	**96.71%**	**96.92%**
	ORG	94.56%	91.78%	93.15%
	EVENT	91.20%	91.94%	91.57%
	CARDINAL	96.97%	94.12%	95.52%
	LOC	0.00%	0.00%	0.00%
	WORK_OF_ART	50.00%	50.00%	50.00%
	DOCUMENT	80.00%	63.16%	70.59%
	OTHER_QUANTITY	60.00%	60.00%	60.00%
Fine-tuned spaCy Transformer-based English model	DATE	**95.69%**	93.16%	94.41%
	PERSON	88.34%	88.86%	88.60%
	GPE	92.98%	**97.03%**	**94.96%**
	ORG	94.30%	90.78%	92.51%
	EVENT	92.62%	91.77%	92.19%
	CARDINAL	86.05%	92.50%	89.16%
	LOC	25.00%	50.00%	33.33%
	WORK_OF_ART	72.73%	47.06%	57.14%
	DOCUMENT	25.00%	50.00%	33.33%
	OTHER_QUANTITY	72.73%	47.06%	57.14%

Fig. 3 Entities' cardinalities in the AAMD dataset



loss of meaning. Rather, retaining these characters and strings was proven to improve the overall performance of NLP models and prevent erroneous entity recognition and classification.

3.1 Comparative evaluation of the results for each fine-tuned model against the original model

This section presents and compares the results of the original models with those of the fine-tuned models. Table 3 presents the precision, recall, and F1 scores for both the original and fine-tuned models on the AAMD dataset, with bolded and italicised figures indicating the best and second-best results, respectively. Furthermore, Fig. 3 presents the cardinality of each entity label or class. Analysing these cardinalities helps in interpreting the results, as data-driven models, as is the case with neural network-based models, are highly dependent on both the quality and quantity of the training data. As we will demonstrate empirically, model performance is strongly influenced by the number of available samples for each class.

Among the original three, the spaCy Transformer-based model demonstrated equal or superior performance across all entity classifications, but slightly worse on the classification of CARDINAL entities. However, the zero-shot learning model, GLiNER, was also capable of effectively classifying entities, particularly for domain-specific tags that were added to the "classical labels" (the ones used by the pre-trained spaCy models), such as DOCUMENT. Regardless, the spaCy Transformer-based model performed both better and more consistently.

Notably, the fine-tuned models achieve better results than the original, though neither consistently outperformed the other across all categories. For example, the Transformer-based model achieved a higher F1-score for classifying DATE, GPE, EVENT, LOC, and WORK_OF_ART entities. Conversely, the large model better-classified PERSON, ORG, CARDINAL, DOCUMENT, and OTHER_QUANTITY entities. The entities that were classified most effectively were DATE, PERSON, GPE, ORG, EVENT, and CARDINAL. LOC, WORK_OF_ART, OTHER_QUANTITY, and DOCUMENT were poorly classified. This is probably due to the small sample of these entity types in the chosen datasets. Indeed, the training dataset included only 9 instances of LOC entities, 82 of WORK_OF_ART, 27 of OTHER_QUANTITY and 136 of DOCUMENT, which are significantly less represented than the other classes (cf., Fig. 3).

Further examination of the errors revealed that the misclassification of both LOC and WORK_OF_ART was primarily due to the span of information the model considered. For example, in the GT, the entity 'tomb of Nesbanedjed' was classified as LOC. In practice, however, the model classified 'Nesbanedjed' alone as PERSON, which is still technically correct. Overall, the model struggles with WORK_OF_ART classifications. For instance, the GT label for ‘held base of Funerary Couch’ is WORK_OF_ART, but the model later classified ‘held’ as an EVENT, informed by the tagging convention which annotated informative verbs as events. In this case, however, the text is a description of the base of the artefact instead of an actual EVENT. To improve the results, we suggest limiting verb annotations to those referring to real EVENTs (like ‘to purchase’ and ‘to acquire’), where possible. For example, we believe it is unnecessary to tag verbs like ‘hold’ if it is referring to part of a WORK_OF_ART and are not related to a ’provenance EVENT’ of legal or physical ownership. There are also discrepancies in the case of OTHER_QUANTITY tags, which likely stemmed from differences in the labelling of the data. Most of the discrepancies arose from the presence of an article preceding the entity. For instance, if the GT was ‘a group of ten,’ the resulting tagged part by the models was ‘group of ten,’ which correctly encompassed the information of interest, but omitted the article. In this case, we suggest removing unnecessary information from what is labelled, as it could be the case with the articles.

Interestingly, when classifying DOCUMENT, the large model outperformed the Transformer model. However, the test set contained only 19 samples of this entity type, meaning that even a small number of correct classifications can significantly affect the percentage error.

Just as LOC and WORK_OF_ART were occasionally confused, similar errors were observed in the classification of DOCUMENTs, where misclassification resulted from the model’s inability to interpret an object title as such. Nonetheless, the labelling of these parts was still largely correct. For example, in the case of ‘Charles Stone Ltd. Antiquities Catalogue 173’ (which should be tagged as DOCUMENT because it is the title of an auction catalogue), the model classified ‘Charles Stone Ltd.’ as an organisation (ORG). This classification is correct per se, but it would be ideal for the model to label the whole string as DOCUMENT. We believe this type of error can be addressed through post-processing, by merging the ORG and DOCUMENT entities when an organisation name precedes or follows terms like ‘catalogue’, ‘archive’, or ‘document’. In such cases, the organisation should be treated as part of a single, unified DOCUMENT entity.

Table 4 presents the original and fine-tuned models’ precision, recall, and F1 scores for the ROF dataset. Comparing the original models, the Transformer-based model performed equally well or better in almost all cases except for CARDINAL and QUANTITY. Further, GLiNER performed worse than the original spaCy models and was unable to classify custom domain-specific labels like PRODUCTION, DESCRIPTION, and CONDITION. As with the processing of the AAMD dataset, the fine-tuned models showed improved results over the original models.

The fine-tuned spaCy models outperformed the original models, particularly when classifying CARDINAL and OTHER_QUANTITY entities; this is relatively unsurprising, considering the original spaCy models were not trained on the OTHER_QUANTITY label, hindering their ability to classify it. Conversely, the CARDINALITY label represents a class referring to numerical information, which is inherently more identifiable and less ambiguous. Moreover, the categories that differentiate numerical classes, such as MONEY, QUANTITY, OTHER_QUANTITY, and CARDINAL, are defined in a way that ensures clear distinctions among them. MONEY refers to information on monetary values (in this case the metrics are zero because there are no instances of MONEY in the test set, cf. Fig. 4). QUANTITY refers to size and distance measures as well as weights (aligned with the definition provided by spaCy). OTHER_QUANTITY includes percentages (e.g. the percentage of silver in a metal artefact) and portions of items, as in the case ‘one third preserved’, referring to a broken artefact. CARDINAL is defined as all the numbers that do not fall into other categories (as per the common spaCy definition). It should be noted that we have not used all the spaCy labels to annotate the datasets, but only those which were relevant to the datasets under study (e.g. ORDINAL was not found to be relevant), cf., Fig. 4.

Overall, the Transformer-based model outperformed the large model for all tags except CARDINAL, QUANTITY, and OTHER_QUANTITY. In the case of CARDINAL and OTHER_QUANTITY, the F1 scores between the large and Transformer-based models are quite close, at 82.67% versus 79.49% (CARDINAL) and 63.64% versus 66.67% (OTHER_QUANTITY). A greater difference was seen in the F1-score for the two models’ classification of QUANTITY: 63.59% and 56.62%. As stated, the Transformer-based model achieves

Table 4 Results of the spaCy English and the GLiNER models on the ROF dataset. Bold indicates the best result for each metric, while italics denotes the second-best value

Model	Entity	Precision	Recall	F1-score
Original spaCy large English model	DATE	33.33%	22.41%	26.80%
	PERSON	31.25%	34.09%	32.61%
	GPE	**43.18%**	39.58%	**41.30%**
	ORG	10.43%	40.00%	16.55%
	EVENT	0.00%	0.00%	0.00%
	CARDINAL	22.45%	**64.71%**	33.33%
	LOC	0.00%	0.00%	0.00%
	WORK_OF_ART	0.00%	0.00%	0.00%
	QUANTITY	7.81%	4.59%	5.78%
	MONEY	0.00%	0.00%	0.00%
Original spaCy Transformer-based English model	DATE	38.10%	27.59%	32.00%
	PERSON	30.51%	40.91%	34.95%
	GPE	68.75%	68.75%	68.75%
	ORG	10.00%	3.33%	5.00%
	EVENT	0.00%	0.00%	0.00%
	CARDINAL	23.88%	47.06%	31.68%
	LOC	**100.00%**	**70.59%**	**82.76%**
	WORK_OF_ART	0.00%	0.00%	0.00%
	QUANTITY	25.58%	20.18%	22.56%
	MONEY	0.00%	0.00%	0.00%
GLiNER large English model (Transformer-based)	DATE	**47.62%**	34.48%	**40.00%**
	PERSON	27.05%	**75.00%**	39.76%
	GPE	0.00%	0.00%	0.00%
	ORG	0.00%	0.00%	0.00%
	EVENT	16.67%	1.85%	3.33%
	CARDINAL	0.00%	0.00%	0.00%
	LOC	1.47%	33.33%	2.82%
	WORK_OF_ART	14.65%	22.55%	17.76%
	QUANTITY	14.29%	9.17%	11.17%
	MONEY	0.00%	0.00%	0.00%
	PART	0.00%	0.00%	0.00%
	MATERIAL/TECH	0.00%	0.00%	0.00%
	OTHER_QUANTITY	0.00%	0.00%	0.00%
	PRODUCTION	0.00%	0.00%	0.00%
	DESCRIPTION	0.00%	0.00%	0.00%
	CONDITION	0.00%	0.00%	0.00%
Fine-tuned spaCy large English model	DATE	38.98%	39.66%	39.32%
	PERSON	59.62%	70.45%	64.58%
	GPE	74.07%	41.67%	53.33%
	ORG	38.24%	43.33%	40.62%
	EVENT	64.65%	59.26%	61.84%
	CARDINAL	75.61%	**91.18%**	**82.67%**
	LOC	0.00%	0.00%	0.00%
	WORK_OF_ART	41.18%	34.31%	37.43%
	QUANTITY	63.89%	63.30%	63.59%
	MONEY	0.00%	0.00%	0.00%
	PART	31.37%	27.12%	29.09%
	MATERIAL/TECH	48.95%	39.55%	43.75%

Table 4 (continued)

Model	Entity	Precision	Recall	F1-score
	OTHER_QUANTITY	**100.00%**	50.00%	66.67%
	PRODUCTION	55.56%	21.74%	31.25%
	DESCRIPTION	20.93%	18.65%	19.73%
	CONDITION	39.29%	25.58%	30.99%
Fine-tuned spaCy Transformer-based English model	DATE	54.24%	55.17%	54.70%
	PERSON	71.15%	84.09%	77.08%
	GPE	66.67%	45.83%	54.32%
	ORG	46.67%	46.67%	46.67%
	EVENT	59.66%	65.74%	62.56%
	CARDINAL	70.45%	**91.18%**	**79.49%**
	LOC	14.29%	33.33%	20.00%
	WORK_OF_ART	53.76%	49.02%	51.28%
	QUANTITY	56.36%	56.88%	56.62%
	MONEY	0.00%	0.00%	0.00%
	PART	42.16%	36.44%	39.09%
	MATERIAL/TECH	56.02%	52.54%	54.23%
	OTHER_QUANTITY	**87.50%**	50.00%	63.64%
	PRODUCTION	41.18%	30.43%	35.00%
	DESCRIPTION	28.64%	30.57%	29.57%
	CONDITION	38.71%	27.91%	32.43%

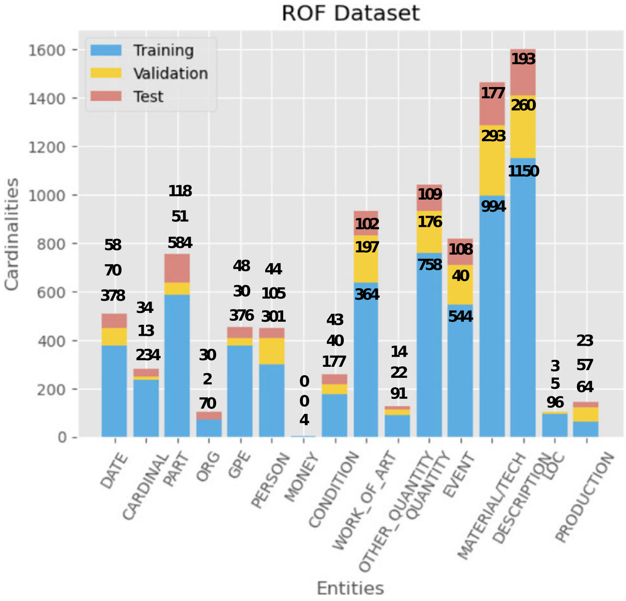


Fig. 4 Entities' cardinalities in the ROF dataset

higher scores for most tags (13 out of 16) compared to the large model. This improvement is likely due to the superior ability of Transformer-based models to remember and process longer patterns, in contrast to the CNN-based models. One poorly performing category is LOC, understandably, as there are very few samples (3) in the test dataset. Another problematic label for both models was DESCRIPTION, but for a different reason. The DESCRIPTION category is the most challenging as it contains free text that could be classified under other tags, but doing so would be unconstructive and even detrimental. For example, consider the object description 'icon depicting Mary, mother of Jesus, and the Christ child on a gold background.' In this case, Mary and Jesus would

Table 5 Results of the spaCy English and the GLiNER models on the DLA dataset. Bold indicates the best result for each metric, while italics denotes the second-best value

Model	Entity	Precision	Recall	F1-score
Original spaCy large English model	DATE	**95.12%**	**92.86%**	**93.98%**
	PERSON	85.71%	64.86%	73.85%
	GPE	71.79%	80.00%	75.68%
	ORG	0.00%	0.00%	0.00%
	EVENT	0.00%	0.00%	0.00%
	LOC	66.67%	12.50%	21.05%
Original spaCy Transformer-based English model	DATE	**100.00%**	88.10%	**93.67%**
	PERSON	76.92%	67.57%	71.94%
	GPE	77.50%	**88.57%**	82.67%
	ORG	0.00%	0.00%	0.00%
	EVENT	0.00%	0.00%	0.00%
	LOC	87.50%	43.75%	58.33%
GLiNER large English model (Transformer-based)	DATE	**90.70%**	**92.86%**	**91.76%**
	PERSON	61.36%	72.97%	66.67%
	GPE	0.00%	0.00%	0.00%
	ORG	0.00%	0.00%	0.00%
	EVENT	0.00%	0.00%	0.00%
	LOC	23.33%	87.50%	36.84%
	PRODUCTION	0.00%	0.00%	0.00%
Fine-tuned spaCy large English model	DATE	**100.00%**	**100.00%**	**100.00%**
	PERSON	78.95%	81.08%	80.00%
	GPE	73.17%	85.71%	78.95%
	ORG	0.00%	0.00%	0.00%
	EVENT	87.50%	**100.00%**	*93.33%*
	LOC	66.67%	12.50%	21.05%
	PRODUCTION	57.14%	66.67%	61.54%
Fine-tuned spaCy large Transformer-based English model	DATE	**100.00%**	**100.00%**	**100.00%**
	PERSON	*88.89%*	*86.49%*	*87.67%*
	GPE	66.67%	85.71%	75.00%
	ORG	**100.00%**	**100.00%**	**100.00%**
	EVENT	**100.00%**	**100.00%**	**100.00%**
	LOC	**100.00%**	25.00%	40.00%
	PRODUCTION	80.00%	66.67%	72.73%

both be tagged PERSON, but to include them in the resulting KG would have a detrimental impact on the analysis of gathered data as they are not real actors in the circulation of cultural goods. Consequently, this information (which often follows the key word "depicting", "depicts", "representing" or 'represents") should be tagged DESCRIPTION, to be added as a property to the artefact data object. To reduce the number of misclassifications of DESCRIPTION, we believe that it would be beneficial to adopt a more deterministic classification approach, relying on 'trigger verbs' like those just mentioned.

Lastly, Table 5 presents the original and fine-tuned models' precision, recall, and F1-score from processing the DLA dataset. Overall, the original spaCy models performed similarly while the performance of the GLiNER model was worse. Notably, the latter was unable to classify custom labels like PRODUCTION. However, as with the case of the AAMD dataset, the fine-tuned models demonstrated improved performance compared to the original models. Of these, the Transformer-based model outperformed the larger model for all tags except GPE. While both models can classify all the tags, their performance in classifying LOC entities was significantly lower, though this kind of entity is not the least represented in the dataset (cf., Fig. 5). Upon analysing the errors in the

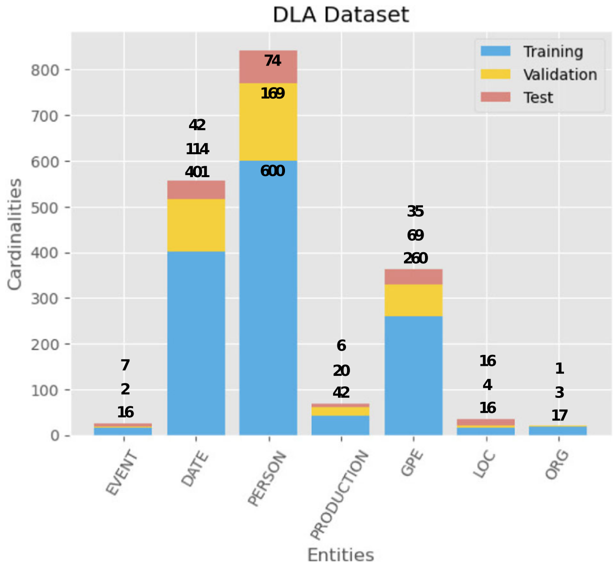


Fig. 5 Entities' cardinalities in the DLA dataset

classification of LOC entities, it was observed that most misclassifications occurred with LOC being classified as GPE. This type of error is not a major concern for us, as our primary goal is to distinguish between places, organisations, and people, regardless of whether the place is precisely defined spatially. Our suggestion would be to use only one unique label to tag both non-political locations, or geographical features, and GPEs.

3.2 Evaluation of model robustness

This section examines whether a model trained on one dataset can classify the same types of entities in another dataset. As shown in Table 6, the models perform poorly on other datasets, reinforcing that their efficacy derives from having been fine-tuned on a dataset with a specific structure, scope, and set of labels. This outcome was anticipated, given the significant differences in the information contained within each dataset and the way that information is presented. Notably, the classes that are classified most accurately are GPE and DATE, which are two of the easiest to distinguish among entity types.

3.3 Determination of the best model

Table 7 presents the F1 scores of the fine-tuned models across the three datasets. The results are promising, with high F1 scores achieved for the AAMD Object Registry and the Catalogue of Wartime Losses, despite the models being fine-tuned with only a few hundred labelled samples.

The poorest results are observed with the labelling of the ROF dataset, which contains the highest number of tags (16 different labels). In contrast, the AAMD dataset comprises 10 entities, while the DLA dataset includes only 7 tags. While the training sample sizes are similar across all datasets that as the number of entity types increases, so does the requirement for additional training samples.

3.4 Comparative evaluation of the best model, observations, and challenges

Figure 6 presents three different provenance samples from an open dataset of objects in the collection of the Metropolitan Museum of Art ('the Met') that have been linked to known traffickers. We utilised the best model

Table 6 F1 scores of the best models tested on the other datasets tested evaluated on datasets other than those they were fine-tuned on

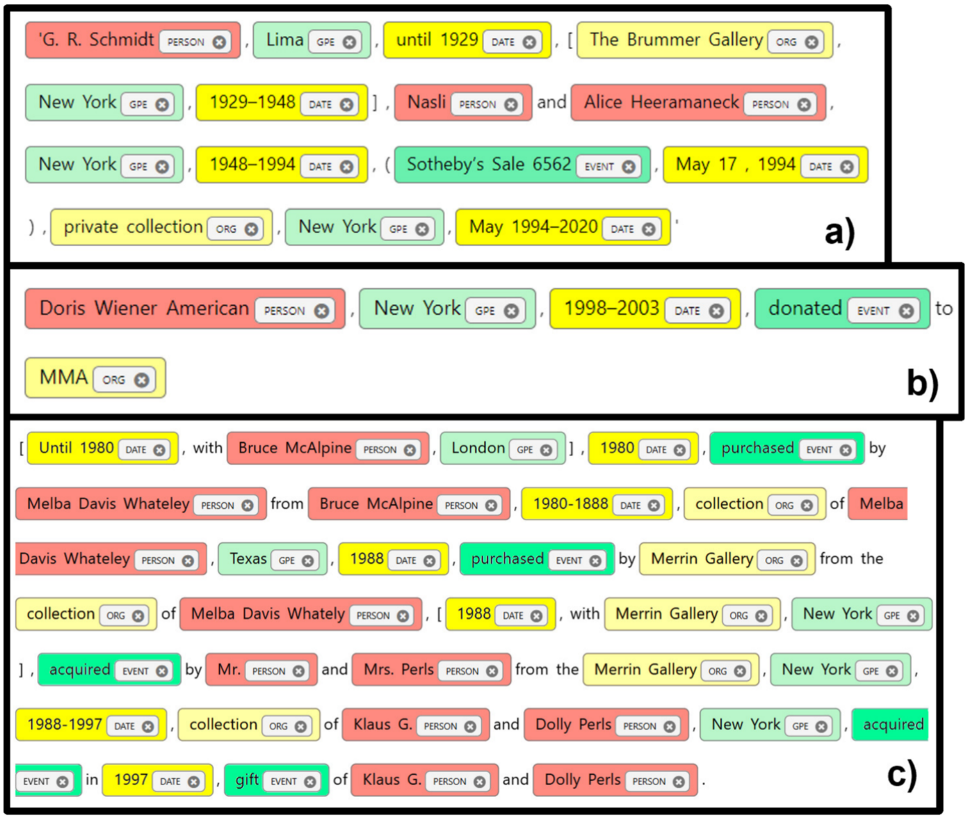
Model	Entity	F1-score
Transformer fine-tuned on DLA used on AAMD	ORG	37.82%
	DATE	38.04%
	LOC	0.00%
	PRODUCTION	0.00%
	EVENT	8.75%
	GPE	**53.61%**
	PERSON	*50.95%*
Transformer fine-tuned on AAMD used on DLA	ORG	11.76%
	DATE	**94.25%**
	LOC	30.00%
	PRODUCTION	0.00%
	EVENT	22.22%
	GPE	78.38%
	PERSON	*79.45%*
Transformer fine-tuned on AAMD used on ROF	ORG	19.39%
	DATE	36.04%
	LOC	0.00%
	PRODUCTION	0.00%
	EVENT	17.85%
	GPE	**37.50%**
	PERSON	*35.38%*
	CARDINAL	27.10%
	WORK_OF_ART	7.80%
	OTHER_QUANTITY	2.13%
	ORG	19.39%
	DATE	36.04%
Transformer fine-tuned on ROF used on AAMD	ORG	50.78%
	DATE	54.66%
	LOC	0.00%
	PRODUCTION	0.00%
	EVENT	*68.33%*
	GPE	47.32%
	PERSON	67.94%
	CARDINAL	**74.70%**
	WORK_OF_ART	12.77%
	OTHER_QUANTITY	0.00%
	ORG	50.78%
	DATE	54.66%
Transformer fine-tuned on DLA used on ROF	ORG	*34.21%*
	DATE	15.38%
	LOC	0.00%
	PRODUCTION	0.82%
	EVENT	17.96%
	GPE	**45.36%**
	PERSON	14.69%
Transformer fine-tuned on ROF used on DLA	ORG	0.00%
	DATE	**96.47%**
	LOC	0.00%

Table 6 (continued)

Model	Entity	F1-score
	PRODUCTION	0.00%
	EVENT	0.00%
	GPE	*76.19%*
	PERSON	62.07%

Table 7 Cumulative F1-scores of the fine-tuned models

	F1-score of the fine-tuned large English model	F1-score of the fine-tuned transformer-based English model
AAMD	92.93%	**94.23%**
ROF	81.01%	**85.15%**
DLA	46.08%	**50.23%**



'G. R. Schmidt PERSON , Lima GPE , until 1929 DATE , [The Brummer Gallery ORG , New York GPE , 1929–1948 DATE] , Nasli PERSON and Alice Heeramaneck PERSON , New York GPE , 1948–1994 DATE , (Sotheby's Sale 6562 EVENT , May 17 , 1994 DATE) , private collection ORG , New York GPE , May 1994–2020 DATE '

Doris Wiener American PERSON , New York GPE , 1998–2003 DATE , donated EVENT to MMA ORG

[Until 1980 DATE , with Bruce McAlpine PERSON , London GPE] , 1980 DATE , purchased EVENT by Melba Davis Whateley PERSON from Bruce McAlpine PERSON , 1980-1888 DATE , collection ORG of Melba Davis Whateley PERSON , Texas GPE , 1988 DATE , purchased EVENT by Merrin Gallery ORG from the collection ORG of Melba Davis Whately PERSON , [1988 DATE , with Merrin Gallery ORG , New York GPE] , acquired EVENT by Mr. PERSON and Mrs. Perls PERSON from the Merrin Gallery ORG , New York GPE , 1988-1997 DATE , collection ORG of Klaus G. PERSON and Dolly Perls PERSON , New York GPE , acquired EVENT in 1997 DATE , gift EVENT of Klaus G. PERSON and Dolly Perls PERSON .

Fig. 6 NER detections on three samples from the Met database using the best NER model fine-tuned on the AAMD provenance dataset

fine-tuned on the AAMD dataset, which contains similar provenance information, to evaluate its performance on this external dataset. Our findings indicate that the model performed quite accurately. In sample (a), there is only one error: 'Sotheby's' in the text 'Sotheby's Sale 6562' should be classified as an ORG rather than as part of the EVENT. However, this type of error can be corrected through a post-processing step. Sample (b) features 'American,' which is incorrectly classified as part of a PERSON name but should be categorised as a GPE. Finally, the third example (c) is classified correctly.

In the context of the RITHMS project, the most crucial data for extracting named entities pertains to provenance information. As introduced above, this type of data provides insights into the ownership history of artefacts, which is valuable for understanding the connections between specific organisations and individuals associated with each artefact [22]. Notably, this type of dataset (represented by the AAMD) yields the highest performance, attaining a high F1-score of 94.23%.

The most significant challenge was extracting named entities from the ROF dataset (cf., Table 4). This difficulty arises because the DESCRIPTION often contains information that can easily be classified into other categories. Therefore, a potential solution is to consider removing the description of what is represented in the artefact as a pre-processing step, as it can introduce uncertainty in the labelling process for the models.

3.5 Opportunities for cultural heritage research

The methodology outlined in the present study contributes to the broader development of tools for automating provenance analysis and detecting risk in cultural heritage datasets. There are also other, relevant and in some ways similar, projects in the field of cultural heritage that aim to improve the accessibility and interoperability of heritage datasets. For instance, ArCo, the Italian Cultural Heritage Knowledge Graph, has structured extensive cultural heritage data into Resource Description Framework (RDF) triples, facilitating semantic interoperability across tools [23].

Leveraging more computational methods, the Heritage Connector project has employed Machine Learning to link museum catalogues with external datasets such as Wikidata, improving accessibility and the consolidation of data on related records. The project's methodology, discussed in [24] shares many similarities with our own approach, particularly in the use of NLP and NER techniques for entity extraction and KG construction within the cultural heritage domain. However, while the Heritage Connector project aims to construct a KG using neural networks, the methodology described in the present study goes beyond KG construction to integrate SNA, providing insights into the relationships and patterns among the entities in the source datasets. Further, our fine-tuned models boast much higher F1 scores with a quite small number of labelled samples (few hundred), suggesting a strong balance between precision and recall, leading to more accurate and reliable results. This is particularly valuable for the context in which our models were developed, where both the inclusion of relevant entities and the exclusion of irrelevant ones are critically important. Prior research using the present methodology (also conducted within the scope of RITHMS project) has shown that NLP-based entity extraction is a valuable tool for identifying problematic provenance records and highlighting patterns that may indicate illicit activity [25]. In a more in-depth analysis of the AAMD dataset, a number of records initially caused challenges due to vague, inconsistent, or unusually structured provenance entries. These included instances where ownership histories contained bracketed or ambiguous text or listed generic entities such as "private collection" rather than verifiable individuals or institutions. One example involved an earthenware figure from Ecuador, whose provenance linked it to two individuals who have been the focus of other research into trafficking activities. Manual investigation revealed that both individuals had prior involvement in the trade of looted artefacts from Latin America, and that the object's provenance was not only vague but potentially designed to obscure its illicit origins. After processing the available provenance entries and extracting entities and relationships with the fine-tuned NLP model, the resulting KG (constructed of the extracted nodes and edges) enabled the reconstruction of a sprawling social network of over 72,000 entities and more than 110,000 relationships. Centrality analyses (a statistical method within SNA) conducted on this system revealed central and influential actors within the art market, including auction houses like Sotheby's New York as well as individuals and organisations who have been the target of previous investigations into trafficking networks such as Nicolas Koutoulakis and the Merrin Gallery. More significantly, the NLP-KG-SNA methodology helped identify lesser-known figures whose proximity to problematic actors warrants further scrutiny. The insights returned from this research demonstrate the potential for computational approaches not only to organise and better structure provenance data but also to flag hidden actors and systemic patterns within networks of cultural property exchange. As such, the NLP models and methods developed in the present study offer a scalable and innovative opportunity for provenance researchers, cultural heritage professionals, and LEAs engaged in the fight against art crime and trafficking.

4 Conclusions

Models to perform named entity recognition have been improved over the past years, reaching near-human-level performance. However, performance is highest when models are trained with a large quantity of labelled data and can leverage 'long' information, which potentially provides informative context and helps the recognition process. While large, labelled datasets have been created for projects in other disciplines, it is rare to have such materials specifically pertaining to cultural heritage, despite the availability of open data in this field. Moreover, in the case of using data from open datasets available online—as was the case for the RITHMS project - informative texts are often composed of just a few lines. Therefore, models cannot utilise long information and must rely on minimal data to capture the complete context from the input texts of the datasets.

In this work, three distinct datasets and models have been specifically designed to create structured data, addressing this challenge, and facilitating the development of graph databases which leverage similar domain-specific materials.

Among the datasets, we argue that the most important for the RITHMS project is the AAMD, with provenance information. The best-performing models have enabled extraction of entities like people and organisations, which are crucial actors in reconstructing the circulation of artefacts. Moreover, our convention that expands the definition of EVENT entities to include verbs as well as named activities significantly enhances the value of these unstructured texts, in which actions such as 'acquisition' are more often implied than explicitly stated. This is a valuable step forward in further defining, reconstructing, and visualising relations among entities.

Future efforts aim to optimise the NER methodology—for example, by removing redundant labels through the merging of related categories, such as LOC and GPE. Another improvement could involve adding a module prior to the NER stage that identifies the most informative parts of the text. This pre-processing step can be implemented using either deterministic algorithms or topic modelling techniques. Its goal wold to determine whether the extracted information reflects the artefacts themselves, rather than what is represented on the artefacts. Moreover, we would develop optimised relation extraction algorithms for the three diverse types of datasets to permit complete KG creation, which is the basis for performing social network analysis. This research will continue to leverage open data and novel technological methods for contributing to the study of the trade of cultural heritage objects.

Acknowledgements Our thanks to the AAMD, Romanian Police, and Polish Ministry of Culture and National Heritage for managing the databases on which this research is based and for making these resources available for public use. This research was conducted under the auspices of RITHMS project (G.A. 101073932), funded by the European Union. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Commission; neither the European Union nor the granting authority can be held responsible for them.

Author Contributions Conceptualisation contributed by SF, RG, ML, MD, and AT; methodology contributed by SF, RG, and ML; software and data collection contributed to SF and ML; formal analysis contributed by SF and RG; writing—original draft preparation contributed by SF, RG, and ML; writing—review and editing contributed by MD and AT; visualisation contributed by SF; supervision and project administration contributed by MD and AT; funding acquisition contributed by AT All authors have read and agreed to the published version of the manuscript.

Funding This research was funded by the European Union, Horizon Europe, under the auspices of the RITHMS project (G.A. 101073932) and carried out at the Centre for Cultural Heritage Technology of the Italian Institute of Technology.

Data Availability The datasets presented in this article are not readily available because they are part of the ongoing RITHMS project. These data were derived from the following publicly available resources: The AAMD Object Registry: https://aamd.org/object-registry/; Obiecte Furate database: https://politiaromana.ro/ro/obiecte-furate; DLA Catalogue of Wartime Losses: http://www.dzielautracone.gov.pl/en/product-war-losses.

Declarations

Conflict of interest The authors declare no Conflict of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Code availability The code is not publicly available, but the trained Natural Language Processing models can be accessed at: https://github.com/IIT-CCHT/NER-models-CH-datasets.

Use of AI tools During the preparation of this work, the authors used ChatGPT to review grammar and enhance the overall reading experience. After using this service, the authors reviewed and edited the content as needed.

References

1. Mishra S, Misra A (2017) Structured and unstructured big data analytics. In: Paper presented at the 2017 international conference on current trends in computer, electrical, electronics and communication (CTCEEC), pp 740–746
2. Barabási A, Pósfai M (2016) Network science. Cambridge University Press, Cambridge
3. Hogan A et al (2021) Knowledge graphs. ACM Comput Surv 54:1–37
4. Brauer F (2017) Mathematical epidemiology: past, present, and future. Infect Disease Modell 2:113–127
5. Pio-Lopez L, Valdeolivas A, Tichit L, Remy É, Baudot A (2021) Multiverse: a multiplex and multiplex-heterogeneous network embedding approach. Sci Rep 11:8794
6. Sowa JF (2014) Principles of semantic networks: explorations in the representation of knowledge. Morgan Kaufmann, San Mateo
7. Manola F, Miller E (2014) Rdf prime. https://www.w3.org/TR/rdf11-primer/
8. Chinchor N, Robinson P (1997) Muc-7 named entity task definition
9. Li J, Sun A, Han J, Li C (2020) A survey on deep learning for named entity recognition. IEEE Trans Knowl Data Eng 34:50–70
10. Sporleder C (2010) Natural language processing for cultural heritage domains. Lang Linguist Compass 4:750–768
11. Piskorski J, Yangarber R (2013) Information extraction: past, present and future. In: Multi-source, multilingual information extraction and summarization, pp 23–49
12. Leeson M, Giovanelli R, De Bernardin M, Traviglia A (2024) War, art, and sanctions: social network analysis on the nacp's databases of sanctioned russian individuals and art collectors. Int J Digital Humanit 1–27
13. Leeson M, Giovanelli R, Ferro S, De Bernardin M, Traviglia A (2025) Overcoming data siloes in cultural heritage crime research: a consolidated osint-derived dataset on art, antiquities, and the trade in cultural goods. Arch Sci
14. Achiam J et al (2023) Gpt-4 technical report
15. Meta (2024) Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/
16. Montani I, Honnibal M, Boyd A, Van Landeghem S, Peters H (2023) explosion/spacy: v3.7.2: fixes for apis and requirements (v3.7.2). https://doi.org/10.5281/zenodo.10009823
17. Face H Hugging face homepage. https://huggingface.co/
18. LeCun Y et al (1989) Backpropagation applied to handwritten zip code recognition. Neural Comput 1:541–551
19. Waswani A et al. (2017) Attention is all you need. In: Paper presented at the 31st conference on neural information processing systems (NIPS 2017), Long Beach
20. Liu Y et al (2019) Roberta: a robustly optimized bert pretraining approach
21. Ushio A, Neves L, Silva V, Barbieri F, Camacho-Collados J (2022) Named entity recognition in twitter: a dataset and analysis on short-term temporal shifts
22. Giovanelli R, Traviglia A (2024) Aikogam: an ai-driven knowledge graph of the antiquities market: toward automatised methods to identify illicit trafficking networks. J Comput Appl Archaeol 7
23. Carriero VA et al (2019) Arco: the Italian cultural heritage knowledge graph. CoRR**abs/1905.02840**
24. Dutia K, Stack J (2021) Heritage connector: a machine learning framework for building linked open data from museum collections. Appl AI Lett 2:e23
25. Leeson M, Giovanelli R, De Bernardin M, Ferro S, Traviglia A (2025) RITHMS digital platform: social network analysis for intelligence-led policing of cultural heritage crime. Springer, Berlin

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Sara Ferro[1] · **Riccardo Giovanelli**[1] · **Madison Leeson**[1] · **Michela De Bernardin**[1] · **Arianna Traviglia**[1]

✉ Arianna Traviglia
arianna.traviglia@iit.it

Sara Ferro
sara.ferro@iit.it

Riccardo Giovanelli
riccardo.giovanelli@iit.it

Madison Leeson
madison.leeson@iit.it

Michela De Bernardin
michela.debernardin@iit.it

[1] Centre for Cultural Heritage Technology (CCHT), Fondazione Istituto Italiano di Tecnologia, Via Adriano Olivetti 1, 31056 Treviso, Italy