How to Turn Card Catalogs into LLM Fodder

Mary Ann Tan, Shufan Jiang, Harald Sack

How to Turn Card Catalogs into LLM Fodder

Mary Ann Tan, Shufan Jiang, Harald Sack
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Eggenstein-Leopoldshafen, Germany
Karlsruhe Institute of Technology, Karlsruhe, Germany
{ann.tan, shufan.jiang, harald.sack}@fiz-kalrsruhe.de

Abstract

Bibliographical metadata collections describing pre-modern objects suffer from incompleteness and inaccuracies. This hampers the identification of literary works. In addition, titles often contain voluminous descriptive texts that do not adhere to contemporary title conventions. This paper explores several NLP approaches where greater textual length in titles is leveraged to enhance descriptive information.

Keywords: NLP, named entity recognition, question-answering, large language model, digital libraries

1. Introduction

Cultural heritage (CH) institutions have been spending considerable resources digitizing their vast collections resulting in an overwhelming volume of digitized objects and their metadata. A large proportion of these are organized as linked data. Notable examples include the Rijksmuseum (Alani et al., 2018), WarSampo (Hyvönen et al., 2016), and Europeana (Purday, 2009).

Recently, the European Parliament identified the challenges facing cultural heritage institutions in the context of the emergence of Artificial Intelligence (AI) solutions. One of these challenges is uneven metadata quality (Pasikowska-Schnass and Lim, 2023). Metadata consists of a set of information that describes and provides context to resources.

As the German national aggregator to the Europeana, the *Deutsche Digitale Bibliothek* (DDB) collects metadata from other cultural heritage institutions all over Germany. Its metadata collection has been published on the web[1] and has been made accessible through an API[2].

More than a quarter of the DDB's entire holdings is composed of 13.5 million[3] digitized texts from the libraries. Part of the digitization process and the subsequent creation of these metadata involved taking information from both existing physical catalog cards and digital sources. The fact that the age of these objects spans several millennia leads to a high level of uncertainty.

Due to the evolution of cataloging standards [4] and the age of some of the objects, author attribution, creation date, and subject heading classifications are missing (See Section 2, Figure 1). The absence of this information, which facilitates item identification in a contemporary library, makes search and retrieval a laborious process. These challenges also makes content exploration and recommendations unfeasible.

Using Semantic Web Technologies (SWT), the metadata collection of the DDB is currently encoded as linked open data and stored in a knowledge graph (KG) (Tan et al., 2021b).

Section 2 provides a thorough description of the metadata collection. Section 3 provides a review of related literature, while Section 4 and 5 describe in detail the main contributions of this paper:

- How different NLP tasks and models can be leveraged to address the challenges of metadata incompleteness and inaccuracy.
- How the results of the experiments can help librarians improve their metadata.

Finally, section 6 presents the conclusion and future work.

2. The DDB Collection

The metadata collection of the DDB conforms to the information exchange and description standard specified by the Resource Description Framework (RDF). The metadata is represented using an extension of the Europeana Data Model (EDM) [5]. In accordance with the EDM standards, the DCMI Metadata Element Set [6] (Dublin Core or DC) and the DCMI Metadata Terms [7] (DC Terms or DCT) properties are used to describe a resource.

[1] DDB, https://www.deutsche-digitale-bibliothek.de

[2] DDB Rest API, https://labs.deutsche-digitale-bibliothek.de/app/ddbapi/

[3] As of March 2024

[4] The provenance of catalogs used as source of the metadata is not available to the DDB. This assumption is primarily based on the time span covered by the collection.

[5] EDM, https://pro.europeana.eu/page/edm-documentation

[6] Dublin Core, https://www.dublincore.org/specifications/dublin-core/dces/

[7] DC Terms, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

DLnLD: Deep Learning and Linked Data @LREC-COLING-2024, pages 53–65
21 May, 2024.

The DDB dataset is divided into seven (7) sectors, each corresponding to the type of institution from which the metadata originates, namely, archives, libraries, historical preservation, research, media libraries, museums, and the rest. Participating institutions are numbered in the hundreds. This paper is focused on the metadata provided by libraries.

The flexibility afforded by the EDM in the cataloging process and the large number of contributing institutions lead to the uneven quality of the metadata collection, since only `dc:title` is indicated as mandatory.

In the DDB, a single book may be composed of several digitized objects, such as the front cover, *Ex Libris* page, table of contents, a chapter, a section, or a page showing an illustration. Each digitized object is equivalent to a single metadata record, which is then defined as an instance of the class *edm:ProvidedCHO*. To distinguish these digitized objects from each other, the data property *ddb:hierarchyType* is used.

In addition, an object can either be a *primary* or *secondary* object. This is indicated by the object property *dcterms:isPartOf*. The primary object of a book is the cover page, while the other components are the secondary objects.

Due to the heterogeneous, hierarchical and highly-granular nature of the bibliographic collection, the metadata is aligned to another data model that reflects the standards defined by the Functional Requirements for Bibliographic Records (FRBR) (Tillet, 2004). The main classes in FRBR correspond to the four (4) conceptual entities: *frbr:**Work***, *frbr:**Expression***, *frbr:**Manifestation***, and *frbr:**Item*** or "*WEMI*".

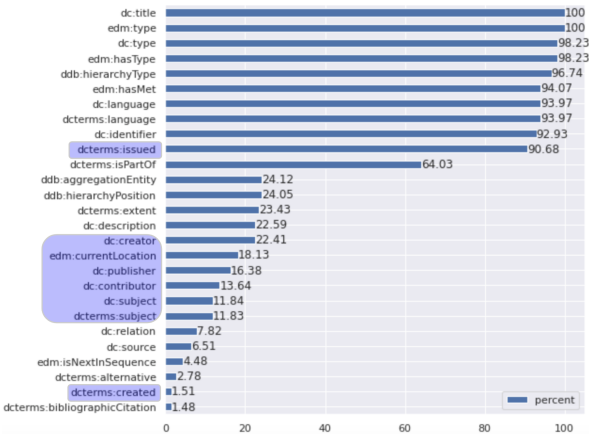

Figure 1: List of properties in the DDB

Tan et al. (2021a) map instances of *edm:ProvidedCHO* to their respective entities in FaBiO or FRBR-Aligned Bibliographic Ontology. Ideally, a primary object such as a cover page can be mapped to its corresponding *Work* entity using properties that distinguish one literary work from another. The title, author, creation date, and subject headings are required to properly identify a literary work. This mapping is necessary since users are more likely to search for higher level representations, *Work* and *Expression* levels, rather than *Manifestation* and *Item* levels.

In a contemporary library, these properties are readily available, often written on catalog cards. However, as can be seen in Figure 1, the property corresponding to the author (*dc:creator*) exists only 22% of the time, while creation date (*dcterms:created*) is specified 1.5% of the time. Moreover, the codificaton of card cataloging rules had not been established prior to the French Revolution; it was only in 1791 when the French Cataloging Code was established (Hopkins, 1992), it is highly likely that inaccurate or incomplete information from old card catalogs, created from the times before then, were carried over during the digitization process.

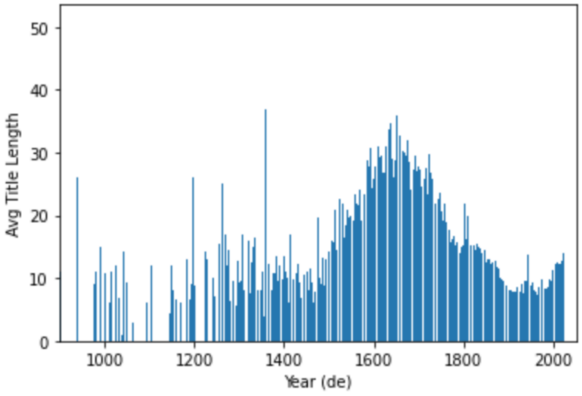

Figure 2: Distribution of title lengths.

A notable example of this phenomenon is the range of values encoded in the data property *dc:title*. As can be seen in Figure 2, the average length of German titles by number of tokens is considerably longer before the French Revolution than after. These titles often contain voluminous descriptive texts that do not adhere to contemporary title conventions.

All example titles from hereon have to be redacted (<...>) due to space constraints. Appendix A lists these examples including full titles, their translations, URLs, and metadata.

As can be seen in Example 1, the title contains the author, creation date and location, subject heading, and a short description of the content.

> *Die Letzte Predigt,* ***Doctoris Martini Lutheri****, heiliger Gedechtnis: So er gethan hat zu* ***Wittemberg*** *...* ***den 17. Januarij, im 1546. Jar*** *:Darinnen wir für falschen Lehrern gewarnet ... werden*

Example 1: Martin Luther's last sermon.

Taking into account the very features that would be considered a disadvantage by present day cata-

loging standards, this paper explores several NLP approaches where greater textual length and more information contained in the titles might be advantageous.

Moreover, the name seen in this example is "*Doctoris Martini Lutheri*", which is the genitive case of Martin Luther's latinized name. The names found in the titles are not normalized: they can be misspelled or in the wrong language, they can contain professional titles ("*Doctoris*"), honorifics, and/or official designations. These naming variations require a more forgiving matching criterion during evaluation.

Another notable example is a dedication written by Lorenz Pscherer for King Gustav II Adolph of Sweden (r. 1611-1632) (Example 2). The metadata attributes authorship (`<dc:creator>`) to "`Horky, Martin *ca. 17. Jh.*`", while the role of "`Pscherer, Lorenz`" is labeled as one of the `<dc:contributor>`'s. Moreover, the dedicatee, King Gustav II Adolph of Sweden, is described as a contributor rather than a subject heading.

> *Ein frölicher Triumph* ***Wagen****/ Von der Göttlichen* `[...]` *Gottfürchtige und gelerte Mann* ***Laurentius Pscherer*** *zu* ***Nürnberg*** *gehabt/ und nu mehr dem* ***7. Septembris Anno 1631****. sich* `[...]`

Example 2: The title containing Lorenz Pscherrer's name.

These inaccuracies add another layer of complexity in the automatic construction of an evaluation dataset.

3. Related Work

The popularity of KGs arose from their ability to encode real-world information using nodes and vertices: the nodes to represent entities or individuals, and the vertices to represent the relationships that exist between these entities.

However, due to the open world assumption, KGs are in practice incomplete, or worse, incorrect. To mitigate the issue of incompleteness, KG completion approaches such as Link Prediction (Rossi et al., 2021) and Entity Alignment (Zeng et al., 2021) became the *de facto* solutions. Both approaches harness the approximation power of KG Embeddings (KGEs) by adding missing information into the KG.

Research into KG construction benefited from advances in Information Extraction (IE), an important branch of NLP. IE provides a scalable solution to KG construction by automatically turning unstructured data, such as texts, into structured or semi-structured data. IE pipelines are often composed of several modules, including, but not limited to, the following: Named Entity Recognition (NER), Entity Linking (EL) and Relation Extraction.

It is possible to enhance the DDB metadata collection by identifying pertinent information from lengthy titles using an IE pipeline composed of fine-grained NER and EL. In this work, we focus on the identification and classification of bibliographic entities.

A recent survey (Ehrmann et al., 2023) summarizes the challenges of NER in historical documents by pointing to the variety of historical document types, topics and domains, noisy input derived from optical character recognition (OCR), handwritten text recognition (HTR), dynamics of language and lack of resources. The use of pre-trained language models in transfer learning leverages knowledge from unlabeled historical corpora. It captures historical language idiosyncrasies during the pre-training phase before adapting the models to a specific NER task in the fine-tuning phrase. The pre-training-fine-tuning paradigm requires task-specific model architecture and storage; it also needs a certain amount of expert annotation. We found the following labeled datasets for German historical and bibliographic named entities:

- AjMC dataset (Romanello et al., 2021; Romanello and Najem-Meyer, 2022) consists of NE-annotated multilingual 19 th century classical commentaries and contains 3,500 mentions of German names, of which 356 are classified as authors.
- CLEF-HIPE 2020 (Ehrmann et al., 2020), a multilingual historical news corpus covering a time span of 200 years, contains 660 mentions of organizations, 58 of which are classified as press agencies.
- NewsEye (Hamdi et al., 2021) consists of annotated multilingual historical newspaper materials published between 1850 and 1950, containing 3,500 German names, of which 30 are classified as article authors.

These labeled datasets are relatively small; covering short time spans, a narrow range of topics and limited materials; this necessitated the creation of our own ground truth data.

The Question Answer (QA) task is another possible solution to leverage the potential of existing data in Language Models (LMs) for IE in a low resource setup. Given a passage and a question, the goal is to provide an answer to be extracted from a given passage. Best performing approaches use the SQUAD (Rajpurkar et al., 2016) dataset and its extensions for training and fine-tuning. This dataset contains handcrafted, general questions and answers drawn from excerpts of top Wikipedia pages. Depending on the passage and the type of question, the expected answer may be simple or complex. For the current use case, the title is

used as the passage. The questions are formulated such that the expected answers are simple and explicit (i.e. author names, creation date, etc.)

The recent proliferation of Large Language Models (LLMs) spurred intense research activity due to the generalization, language understanding and generation ability of LLMs (Wei et al., 2022). Several notable studies provided an in-depth analysis of pre-trained language models on how well they can recall factual knowledge using a series of probing questions (Petroni et al., 2019; Poerner et al., 2020). A fact is formulated as a triple consisting of a subject, a relation, and an object. An LLM is said to "know" a fact, if it can fill in the masked relation in a cloze statement, i.e. Dante [MASK] Florence, where [MASK] is the relation *birthplace*. Petroni et al. (2020) concluded that providing relevant context to the LLM improves fact retrieval performance.

Fine-tuning LLMs for the purpose of this study was not possible due to limited access to computational power, neither was it possible to consult experts for manual annotation of the current dataset.

Regarding the application of LLMs for Historical IE, (De Toni et al., 2022) explores the zero-shot abilities of the `T0` model for coarse-grained NER over the CLEF-HIPE 2020 dataset (Ehrmann et al., 2020) with a naive prompt-based approach; it showed the `T0`-like models' potential to probe for language tags and publication dates.

4. Methodology

This section describes dataset construction (Section 4.1), the evaluation procedure and metrics (Section 4.2), and NLP models used for experimentation, and the experimental setup (Section 4.3).

4.1. Dataset

In order to construct our dataset for experimentation and evaluation of the aforementioned approaches, the entire DDB bibliographic metadata collection has been filtered down to a manageable representative sample.

The DDB has objects in more than 200 languages. The scope of this study is limited to digitized textual objects tagged as "`ger`" for German and "`zxx`" for unknown or no language tag. The Python library *langid* (Lui and Baldwin, 2012) is used to confirm that the titles of these objects are indeed in German, since there are objects where the language of the title is not the same as the value indicated in the metadata[8]. There is a considerable number of objects with Latin titles that have been

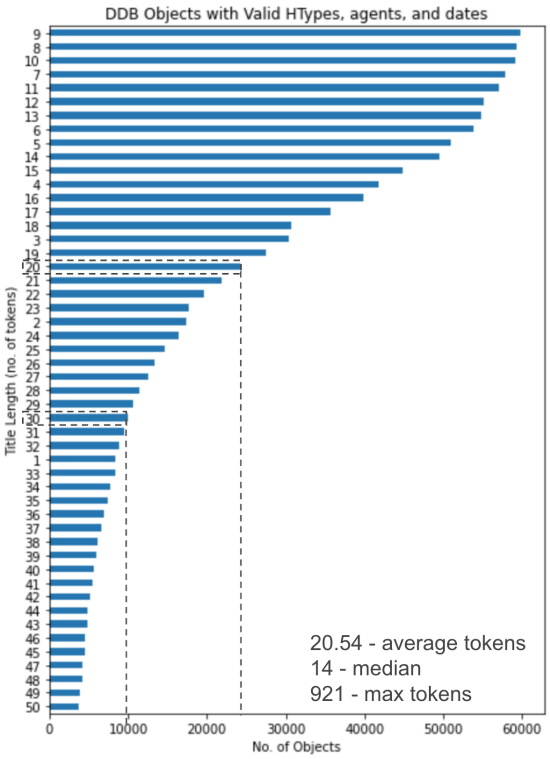

Figure 3: Title token distribution of German titles in the DDB.

tagged as German. In addition, since the collection grew across a long period of time, the German language has had time to evolve. Hence, there are titles written in different versions of the German language from Middle High German (see Example 1) to Standard High German.

Using the *ddb:hierarchyType* mentioned in Section 2, the metadata describing textual objects[9] is further reduced to approximately 30% of its original size. The hierarchy types selected are Monograph, Chapter, Essay, Volume, Manuscript, Letter and Multi-Volume Work, since these are likely to have identifiable titles.

In order to have a ground truth, the representative objects are described with agents (`<dc:creator>`, `<dc:publisher>`, `<dc:contributor>`) and dates (`<dcterms:issued>`, `<dcterms:created>`). Moreover, since the goal is to leverage lengthy titles, the representative objects should have more tokens than the average length of 20.54 (Figure 3). Choosing 30 tokens to be the cut-off still leaves a little over 100k objects after the final pruning criterion. The remaining objects are pruned for the final time to only contain books (`<dc:type> = "Monografie"`), since this object type suggests a physical manifestation in the context of FRBR and can possibly be aligned to their respective higher-level entities in FaBiO.

[8] https://www.ddb.de/item/DNEBFCMMEO52LAQWGT5JHULKPXBU2QYG

[9] `<edm:hasType> == 'TEXT'`

Table 1 provides some statistics about the pruned dataset. Figure 4 in Appendix C shows the distribution of title lengths with respect to age of the objects after pruning.

Characteristic	Value
No. of Objects	108,827
Average Title Length	55.49 Tokens
Median Title Length	47 Tokens
Longest Title Length	364 Tokens

Table 1: Characteristics of the dataset.

4.2. Evaluation Guidelines

The goal of this study is to find out how well a particular approach can retrieve identifying information included in the title, such as dates and agents.

Dates are trivial to compare. The dates stored in the evaluation dataset only include the year element in 'YYYY' format. For metadata values that include month and day, a regular expression is applied to retrieve the year. As in the example in section 2, "1956" is compared against the values of either `<dcterms:created>` or `<dcterms:issued>`, while ignoring "*den 17. Januarij, im...*"

Agents, in the bibliographic domain, refer to the persons responsible, in any capacity, for the creation of the object. Author, editor, and publisher are the roles often attributed to these agents. The properties `<dc:creator>`, `<dc:contributor>`, and `<dc:publisher>` store names of persons in the format of "`last name, first name`" following the German version of the name, without title, honorific, or official designation.

Exact name matching is non-trivial, as mentioned in Section 2 and illustrated in Example 1. To facilitate approximate name matching, an extension of the Python package *sqlite-spellfix* [10] is used.

Spellfix is implemented as a virtual table that stores all the vocabulary terms and uses Levenshtein distance (Levenshtein, 1966) to compute edit distance in order to gauge the lexical similarity between the vocabulary terms and the search string.

The agents' names found in the ground truth are collected and stored as vocabulary terms. The reference table of the *Spellfix* virtual table is composed of two (2) columns: the person's name normalized in the format of "`firstname lastname`"; and the object ID, a 32-character unique identifier of an object, associated with the agent. To illustrate, if a model is able to extract "*Martini Lutheri*" from the text, this string is used to lookup the most similar names found in the *Spellfix* virtual table and the corresponding object ID linked to these names as defined in the primary table. If the ID of the object currently being evaluated is found in the list of the object IDs resulting from the *Spellfix* lookup results, then it is considered a match.

It is important to note that some objects are annotated with agents that cannot be found in the title. Authors are more likely mentioned in the title, like "Schiller's Robbers"[11], while editors and publishers rarely are. However, the latter roles can still be associated with the properties *dc:contributor* and *dc:publisher*. This will not affect the evaluation results, since we test the results on the list of all agents and role-specific agents ("Ground Truth" column in Tables 6 and 8). In addition, a more forgiving *Precision@n* metric is used, where *n* varies depending on the number of agents associated with an object. If there are 2 agents associated with an object, and only a single name gets a match, *Precision@n* will be equal to 1 for this specific object. Appendix B shows some of the matches related to the LLM experiments.

The applied metric deviates from the customary precision, recall, and F_1 score combination for IE, due to the nature of the ground truth, and the variety of name formats found in the text. On the other hand, this metric is similar to the *Top1Acc* measure for the extractive QA task meant for closed-domain evaluation, where 1 point is attributed if the predicted answer has a single word overlap with the labeled answer. As for the task involving LLM, the model is instructed to only provide names without justifications. Therefore, the same metric is used during evaluation.

4.3. NLP Tasks and Models

This subsection describes how the use case of the DDB is recast into the three chosen NLP tasks: (1) **NER**, (2) Extractive **QA** and (3) Open Generative QA using an **LLM**. Moving forward, Task 2 will simply be referred to as **QA** while task 3 as **LLM**

NER. Since the goal is to extract the people, dates, and possibly, subject headings, from a lengthy title, it is appropriate to adopt an IE pipeline. The current state-of-the-art, general-purpose, open source, and off-the-shelf model is the FLAIR English NER Large Model (FLERT[12]) (Schweter and Akbik, 2020). Despite being classified as an English model, its pre-trained Language Model (PLM) is based on XLM-R (Conneau et al., 2020). Choosing FLERT is motivated by its multilingual representations capability and its ability to identify 18

[10] sqlite-spellfix, https://pypi.org/project/sqlite-spellfix

[11] Schiller's Räuber, https://www.ddb.de/item/FXHCBDNJAAHI7PSMOYBMKZS5I47NX36J

[12] FLERT, https://huggingface.co/flair/ner-english-ontonotes-large

different entity types, including dates (DATE) and works (WORK_OF_ART).

Further classification of PERSON entities according to specific bibliographic roles, whether author, editor, or publisher, calls for a fine-grained NER approach. Such a requirement necessitates an expert-annotated dataset that can be used for fine-tuning (Radford and Narasimhan, 2018; Peters et al., 2019) to produce a domain-specific, fine-grained NER model as in LegalNER (Leitner et al., 2019; Akbik et al., 2018). This currently exceeds the scope of this study and is being considered for future work.

For this task, the model is expected to find entity mentions and to classify them given a title. In example 2, the FLERT model recognizes 3 highlighted entities: “Laurentius Pscherer" as PERSON, “Nürnberg" as GPE and “7. Septembris Anno 1631" as DATE. For this specific use case, only the PERSON and DATE entities are scrutinized.

Following the evaluation procedure and metric described in section 4.2, FLERT’s prediction results in a single matching point for each person and each date.

QA. Existing “Extractive" QA models can be retrofitted for the purpose of the DDB. Using the title as the passage, below is the list of simple questions posed to the models:

1. Who is the author? (“*Wer ist der Autor?*")
2. Who wrote the text? (“*Wer hat den Text geschrieben?*")
3. Who is the publisher? (“*Wer ist der Herausgeber?*")

The best German QA models available are fine-tuned using the German equivalent of the Wikipedia articles used in SQUAD, aptly named GermanQUAD (Möller et al., 2021). Using GELECTRA (Chan et al., 2020) as the PLM, these models are fine-tuned with the goal of *extracting* relevant parts of the passage with dense representation to be most similar to the corresponding dense representation of the question. Since the goal is to retrieve the names of the persons from the passage whose specific role is indicated in the question, and the German QA models adopted are not trained on unanswerable QA pairs, it is necessary to ensure that only titles with names are included in the test by using those identified by the NER model to have PERSON entities.

To find out how well the German QA models compare to one of the top 3[13] English QA models, experiments also used the roberta-large-squad2 model[14] published by Deepset. This model is fine-tuned using Squad 2.0 (Rajpurkar et al., 2018). Squad 2.0 is an extension of SQUAD that includes an additional set of 50,000 handcrafted, adversarial questions that have no answers but are very similar to existing answerable questions.

In order to do so, the German titles are translated into English using the DE-EN machine translation model submitted by Facebook’s FAIR for the WMT19News Translation Task[15], which boasts a SacreBLEU score of 40.8 (Ng et al., 2019). The translations of the titles used in the examples are listed in Appendix A

Table 2 shows an example of the answers of the different QA models: both the GELECTRA-based models are provided with the original title in German as context, while the RoBERTa-based (Liu et al., 2019) model is fed with the English machine-translated title. Despite being given a translated text produced by a moderately performing machine translation model, the confidence score of the English QA model is still considerably higher than the German QA models. Nevertheless, these scores are not taken into account since only the answers matter during evaluation.

MODEL	ANSWER	SCORE
gelectra-base-germanquad	Doctoris Martini Lutheri	0.0539
gelectra-large-germanquad	Doctoris Martini Lutheri	0.0115
roberta-large-squad2	Doctoris Martini Lutheri	0.9425

Table 2: Answers of different QA models when asked about the author of Example 1.

LLM. With the optimal mix of instructions, LLMs trained as conversational agents are known to generate impressively coherent and sometimes factual texts. The prompts used for the experiments are patterned after the guidelines provided by Bsharat et al. (2024). Specifically, the following principles are incorporated into the prompts:

- P16: Assign a role to the large language models.
- P8: Use line breaks to separate instructions.
- P25: Clearly state the requirements.

The series of instructions used to test the chosen LLM is provided below. Lines 3-5 are explicitly specified to suit the evaluation procedure described in Section 4.2. Line 6 varies depending on the question that needs to be asked (author, publisher, etc.) Line 7 contains the full title. Given the title as the context, this task is categorized as a *Open* Generative QA task.

1. `You are a librarian doing cataloging work.`

[13] As of March 2024, https://paperswithcode.com/sota/question-answering-on-squad-v2

[14] https://huggingface.co/deepset/roberta-large-squad2

[15] https://huggingface.co/facebook/wmt19-de-en

2. Respond with "I don't know" when uncertain.
3. Enumerate your answers with numbers.
4. Only answer with the name of the persons.
5. Do not provide justifications.
6. Who is/are the publisher/s of this text?
7. "The Last Sermon, Doctoris Martini Lutheri, Sacred..."

Mistral-7B-Instruct-v0.2[16] is an open source LLM developed by Mistral (Jiang et al., 2023). The Mistral-7B PLM is fine-tuned on an instruction dataset developed by HuggingFace. This dataset contains "high-quality, diverse, human-written instructions with demonstrations"[17]. Since the model is trained and fine-tuned with English datasets, the machine-translated titles are used for the succeeding experiments. The choice of Mistral-7B-Instruct-v0.2 is motivated by its availability (open source) and published performance besting state-of-the-art open source LLMs at the time of this writing.

```
1. Martin Luther
2. Wittenberg: Druck von
Paulus Berckmann
Or:
1. Paulus Berckmann (printed
by)
```

Example 3: An Example Response from Mistral-7B-Instruct-v0.2.

Example 3 shows the response of Mistral-7B-Instruct-v0.2 given the aforementioned series of instructions concerning the publisher and provided with the machine-translated title in Example 1. This tests the model's ability to respond with "I don't know". However, the model ignored the instruction, and instead "hallucinated" at least 2 names. The publisher was never in the title and the value of <dc:publisher> in the metadata of Example 1 is "Bergen".

5. Experiments and Results

The goals of the experiments are to find out the following:

- To what extent can coarse-grained, general purpose NER models be used in filling missing metadata descriptions?
- How can a NER model be leveraged to further refine the evaluation dataset for QA and LLM tasks?
- How well can QA models identify the different agent roles?
- How does an LLM-based chat model compare to a QA model in identifying different agent roles?

The configurations and parameters used for the succeeding experiments are kept according to the published default settings.

5.1. NER

Using the dataset constructed in Section 4.1, all 108,827 records are processed with FLERT. Since FLERT is a general-purpose NER model, it is not able to distinguish between the different agent roles. For this task, the ground truth is composed of the values of the 3 agent-related properties: <dc:creator>, <dc:contributor>, and <dc:publisher>.

	PERSON	DATE
Exact Match	9.61%	27.68%
Approx. Match	8.42%	N/A

Table 3: FLERT's Precision@n results.

The disappointing results of Table 3 can be partially explained. Although only lengthy titles are considered, it is possible that despite the large number of tokens, a title might not contain any entity mentions of PERSON or DATE. Table 4 shows the number or proportion of objects from the dataset where PERSON, DATE or both entity types are detected by FLERT.

PERSON	DATE	Both
59.07%	49.12%	32.93%

Table 4: Proportion of objects with PERSON and DATE entity mentions detected by FLERT.

This specific case is shown in Example 4. The NER model correctly identifies "*Das Hohe Lied des Königes Salomons*" (The Song of Songs of King Solomon) as a WORK_OF_ART. A 4-tag model such as Flair's German NER (Large)[18] predicts "Salomon" as a PERSON, which is only partly correct.

Das Hohe Lied des Königes Salomons
: Wie es/ Zu der aus Gott wieder-geboren-
[...] *... ausgefärtiget hat*

Example 4: The Song of Songs, a lengthy title without PERSON or DATE entities.

Although the results of FLERT cannot differentiate between an author and a publisher, this step can already identify possible objects for metadata enhancement, just by identifying the very existence

[16] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[17] https://huggingface.co/datasets/HuggingFaceH4/instruction-dataset

[18] https://huggingface.co/flair/ner-german-large

of the entity mentions. Moreover, by further pruning objects without PERSON entities, the dataset can be further improved for the succeeding tasks, particularly for the extractive QA task where the answer is expected to be present in the context.

5.2. QA

For this task, the evaluation dataset is reduced to only those records that yielded agent matches according to the previous NER model. This step is necessary to ensure that the QA models are provided only with questions where the answers exist in the passage. This cuts down the original dataset by 84% to 17,084 titles.

Ground Truth with all Agents	gelectra-base-german quad	gelectra-large-german quad	roberta-large-squad2
Context	Title (DE)	Title (DE)	Title (EN)
"Who is the author?"	62.94%	**66.23%**	63.07%
"Who wrote" "the text?"	58.12%	60.83%	58.36%

Table 5: QA results against ground truth containing all names.

Table 5 shows that the result of the best performing model, gelectra-large-germanquad, is consistent with its published Exact Match (EM) results of 68.6%. The differences between Middle High German and Standard High German do not seem to matter as much. The results also show that asking direct questions yielded some improvement (i.e. by providing specific roles).

To check whether the models understand the difference between author and publisher, the list of names in the ground truth is made more specific according to the question, such that only values described under dc:creator are included in the reference list when asked about the author, and respectively, when asked about the publisher.

	Ground Truth	gelectra-large-germanquad	roberta-large-squad2
"Who is the author?"	<dc: creator>	**32.19%**	31.16%
"Who is" "the publisher?"	<dc: publisher>	0.85%	0.78%

Table 6: QA results against ground truth containing all names.

The results in Table 6 are inconclusive, because publishers are rarely mentioned in the title. However, looking closely at the title in Example 5, the two names mentioned have two distinct roles:

- <dc:creator>: "Ignatz"[19]
- <dc:contributor>:"Johann Jacob Ferber"[20]

> *Des Hrn.* ***Ignatz****, Edl. von Born, Ritters, K.K. Berg-Raths,* [...] *Gesellschaft zu Padua Mitglieds [et]c. Briefe über* [...] *und Nieder-Hungarn, an den Herausgeber derselben,* ***Johann Jacob Ferber****, Mitglied der Königl.* [...] *zu Florenz, geschrieben*

Example 5: The agent roles of Ignaz von Born and Johann Jacob Ferber.

Table 7 shows the responses of gelectra-large-germanquad and roberta-large-squad2, incorrect answers are highlighted.

Question	gelectra-large-german quad	roberta-large-squad2
...Author<->Autor?	**Johann Jacob Ferber**	**Johann Jacob Ferber**
...Editor<->Redakteur?	Johann Jacob Ferber	Johann Jacob Ferber
...Herausgeber?	Johann Jacob Ferber	-
...Verfasser?	**Johann Jacob Ferber**	-
...Publisher<->Verleger?	**Johann Jacob Ferber**	**Mr. Ignatz, Edl. von Born**

Table 7: QA models not being able to tell the different agent roles.

Depending on the historical context of the object, the translation of the German term *Herausgeber* can either be editor or publisher. *Redakteur* is almost always the direct translation of editor, while *Verfasser* means author, and *Verleger* means publisher. Despite providing the passage in the language native to the respective QA models, these models have difficulty distinguishing agent roles. This limitation could be due to the fact that the titles are fragmented texts and the roles being asked do not explicitly appear with the names mentioned.

5.3. LLM

Using the instructions described in Section 4.3, Table 8 shows that the LLM is less precise when asked about the author, but performs better compared to the QA model in all other experiments conducted.

Question: "Who is the ...?"	**Ground Truth**	**LLM** mistral-7b-instruct-v0.2	**QA** gelectra-large-germanquad
Author	all agents	51.60%	**66.23%**
	<dc: creator>	**37.60%**	32.19%
Publisher	<dc: publisher>	**2.70%**	0.85%

Table 8: LLM vs QA results.

When inspecting the responses closely, Mistral7BInstructv0.2 occasionally makes up names (See Example 3), is not following instructions with regard to formatting (Appendix B #5) and still provides justifications (Appendix B #6), despite being told not to do so.

[19] https://d-nb.info/gnd/118661868

[20] https://d-nb.info/gnd/118686690

5.4. Discussions

Revising tens of millions of metadata records is a daunting task. With the help of these NLP models, it is possible to identify candidate objects for refinement. Concretely, an object lacking in descriptive information, but with a lengthy title from which an NER model may be able to extract pertinent entities, can automatically signal further attention from librarians. Even when the extracted entities are not entirely accurate, the results can be used as suggestions in a post-ingestion editing workflow. The level of post-processing required for each of the objects can also be automatically determined. For instance, those objects whose titles do not yield any results when fed into an NER model will require more work than others.

Since there is currently no gold standard dataset, both QA models and LLMs are meant to gauge their efficacy in determining fine-grained agents. In this setting, objects identified by these models as having authors, but without matching values against `<dc:creator>` indicate the need for manual intervention. In this scenario, the models' results can be leveraged for possible refinements. For example, an extracted agent may already be indicated as a `<dc:contributor>`, in which case, the metadata can be made more accurate by defining this agent as a `<dc:creator>`. Another possibility would be to fill out missing `<dc:subject>`.

The disparity of the adapted models in terms of their published performance and the results shown in this paper can be attributed to several factors.

Primarily, the titles used for the experiments contain fragmented texts in older versions of the German language where spelling and naming conventions changed over time. In contrast, these off-the-shelf models were pre-trained and fine-tuned on contemporary, general purpose texts. This limitation calls for future work in adapting models trained on texts whose age and domain overlap with the DDB dataset. In addition, the absence of a gold standard evaluation dataset limits the validity of the results. This limitation will be the first to be addressed in the next iteration of our work.

Although the experiments conducted with `gelectra-base-germanquad` and `gelectra-large-germanquad` lacked adversarial questions, this limitation was partly mitigated by comparing their results with `roberta-large-squad2`, an English QA model trained on the Squad 2.0 dataset, which includes unanswerable questions. Nonetheless, this calls for further experiments that include titles without entity mentions.

The last limitation concerns the unpredictability of the LLMs and the difficulty of formulating the most optimal prompts. This affects the reproducibility of the experiments conducted in section 5.3.

6. Conclusion

The challenge of incomplete and inaccurate bibliographical metadata collection, the linked data source of DDB-KG, can be addressed using a combination of NLP tasks. The results show that NER, QA and LLMs can, to some extent, be used to extract some bibliographic properties from lengthy titles of historical objects. A domain-specific dataset is currently being prepared for a fine-grained NER model capable of determining literary work title, agent roles, dates, and subject headings.

The experiments make use of an evaluation dataset where the agent roles encoded in the metadata are not entirely accurate. While domain experts are necessary in the preparation of a more precise dataset for future DDB-KG enhancement initiatives leveraging AI models, domain experts can also benefit from the rapid approximation capabilities of AI models. In particular, the list of objects with `PERSON` entities that are not matching any answers provided by either QA models or LLMs may be used as an initial list of objects to undergo expert scrutiny for a possible revision.

Further experiments are planned to compare NLP models that are relevant to the DDB dataset. It is worthwhile to test the efficacy of models trained on 19th-20th historical German text (Ehrmann et al., 2023). Moreover, once the aforementioned gold standard dataset is available, further experiments will be conducted using state-of-the-art commercial LLMs.

It is the ultimate goal to develop a collaborative tool for metadata providers where inputs from both domain experts and AI models can be combined to provide better results in search, retrieval and exploration of cultural heritage.

7. Acknowledgements

We would like to express our gratitude to Frank Pöhlmann who dedicated his time and editing experience to provide valuable feedback and constructive criticism on this paper.

8. Ethics Statement

The authors of this paper are affiliated with FIZ Karlsruhe and Karlsruhe Institute of Technology. Funding was provided solely by FIZ Karlsruhe. The work done on this paper has been conducted ethically. No part of this paper was suggested, generated, improved, or corrected using generative AI models.

9. Bibliographical References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Harith Alani, Chris Dijkshoorn, Lizzy Jongma, Lora Aroyo, Jacco van Ossenbruggen, Guus Schreiber, Wesley ter Weele, and Jan Wielemaker. 2018. The rijksmuseum collection as linked data. *Semant. Web*, 9(2):221–230.

Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. 2024. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Francesco De Toni, Christopher Akiki, Javier De La Rosa, Clémentine Fourrier, Enrique Manjavacas, Stefan Schweter, and Daniel Van Strien. 2022. Entities, Dates, and Languages: Zero-Shot on Historical Texts with T0. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 75–83, virtual+Dublin. Association for Computational Linguistics.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, 56(2).

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Extended overview of clef hipe 2020: named entity processing on historical newspapers. In *CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum*, volume 2696. CEUR-WS.

Ahmed Hamdi, Elvys Linhares Pontes, Emanuela Boros, Thi Tuyet Hai Nguyen, Günter Hackl, Jose G. Moreno, and Antoine Doucet. 2021. A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2328–2334, Virtual Event, Canada. ACM.

Carla Hayden. 2017. *The Card Catalog: Books, Cards and Literary Treasures*, 1st edition. The Library of Congress, Chronicle Books, San Francisco, CA.

Judith Hopkins. 1992. The 1791 french cataloging code and the origins of the card catalog. *Libraries & Culture*, 27(4):378–404.

Eero Hyvönen, Erkki Heino, Petri Leskinen, Esko Ikkala, Mikko Koho, Minna Tamper, Jouni Tuominen, and Eetu Mäkelä. 2016. Warsampo data service and semantic portal for publishing linked open data about the second world war history. In *The Semantic Web. Latest Advances and New Domains*, pages 758–773, Cham. Springer International Publishing.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained Named Entity Recognition in Legal Documents. In *Semantic Systems. The Power of AI and Knowledge Graphs. Proceedings of the 15th International Conference (SEMANTiCS 2019)*, pages 272–287.

Vladimir I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Timo Möller, Julian Risch, and Malte Pietsch. 2021. Germanquad and germandpr: Improving non-english question answering and passage retrieval.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Magdalena Pasikowska-Schnass and Young-Shin Lim. 2023. Artificial intelligence in the context of cultural heritage and museums: Complex challenges and new opportunities. Technical Report PE 747.120, European Parliamentary Research Service, Brussels.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy. Association for Computational Linguistics.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-BERT: Efficient-yet-effective entity embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 803–818, Online. Association for Computational Linguistics.

Jon Purday. 2009. Think culture: Europeana.eu from concept to construction. *Bibliothek Forschung und Praxis*, 33(2):170–180.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Matteo Romanello and Sven Najem-Meyer. 2022. Guidelines for the annotation of named entities in the domain of classics, march 2022. *DOI: https://doi. org/10.5281/zenodo*, 6368101.

Matteo Romanello, Sven Najem-Meyer, and Bruce Robertson. 2021. Optical character recognition of 19th century classical commentaries: the current state of affairs. In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, HIP '21, page 1–6, New York, NY, USA. Association for Computing Machinery.

Andrea Rossi, Denilson Barbosa, Donatella Firmani, Antonio Matinata, and Paolo Merialdo. 2021. Knowledge graph embedding for link prediction: A comparative analysis. *ACM Trans. Knowl. Discov. Data*, 15(2).

Stefan Schweter and Alan Akbik. 2020. Flert: Document-level features for named entity recognition.

Mary Ann Tan, Tabea Tietz, Oleksandra Bruns, Jonas Oppenlaender, Danilo Dessì, and Harald Sack. 2021a. DDB-EDM to FaBiO: The Case of the German Digital Library. *The Semantic Web – ISWC 2021*.

Mary Ann Tan, Tabea Tietz, Oleksandra Bruns, Jonas Oppenlaender, Danilo Dessì, and Harald Sack. 2021b. DDB-KG: The German Bibliographic Heritage in a Knowledge Graph. In *6th International Workshop on Computational History at JCDL – Histoinformatics*, volume 2981. CEUR-WS.org.

Barbara Tillet. 2004. What is FRBR?: A Conceptual Model for the Bibliographic Universe.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Kaisheng Zeng, Chengjiang Li, Lei Hou, Juanzi Li, and Ling Feng. 2021. A comprehensive survey of entity alignment for knowledge graphs. *AI Open*, 2:1–13.

A. Appendix A

Titles and their Details

A.1. Example 1: Martin Luther's Last Sermon.

- **Title:** *Die Letzte Predigt, Doctoris Martini Lutheri, heiliger Gedechtnis: So er gethan hat zu Wittemberg ... den 17. Januarij, im 1546. Jar :Darinnen wir für falschen Lehrern gewarnet ... werden*
- **Google Translate:** `The last sermon, Doctoris Martini Lutheri, holy memory: What he did in Wittemberg... January 17th, 1546: In which we are warned... for [sic] false teachers`
- **WMT19:** `The Last Sermon, Doctoris Martini Lutheri, Sacred Memory: If he did at Wittemberg... January 17, 1546. Yar: In which we are warned for false teachers...`
- **URL:** https://ddb.de/item/6563H62JUWEVSVTH3T7TJWCPK2NOMLK7
- **Metadata:** https://ddb.de/item/xml/6563H62JUWEVSVTH3T7TJWCPK2NOMLK7

A.2. Example 2: The title containing Lorenz Pscherrer's name.

- **Title:** *Ein frölicher Triumph Wagen/ Von der Göttlichen Offenbarung/ so durch den Engel Gottes/ der Gottfürchtige und gelerte Mann* ***Laurentius Pscherer*** *zu* ***Nürnberg*** *gehabt/ und nu mehr dem* ***7. Septembris Anno 1631****. sich glücklichen angefangen*
- **Google Translate:** `A happy triumph chariot/ From the Divine Revelation/ so through the angel of God/ the God-fearing and learned man Laurentius Pscherer had at Nuremberg/ and now the 7th of September 1631. began to be happy`
- **WMT19:** `A devout triumph chariot / From the Divine Revelation / so had by the Angel of God / the God-fearing and learned man Laurentius Pscherer of Nuremberg / and now more the 7th of September in 1631.`
- **URL:** https://ddb.de/item/WECO4OXGK3FXONM57VUUDZDOHACE4VCK
- **Metadata:** https://www.ddb.de/item/xml/WECO4OXGK3FXONM57VUUDZDOHACE4VCK

A.3. Example 4: The Song of Songs, a lengthy title without PERSON or DATE entities.

- **Title:** *Das Hohe Lied des Königes Salomons : Wie es/ Zu der aus Gott wieder-geboren- und/ durch die Betrachtung himmlischer Dinge/ in Gott verliebten Seelen Geist-feuriger Liebesüb- und Külung/ nach der Ordnung des Textes/ schrifftmässig erkläret gesungen; und/ mit anmutigen Kupffer- und Sinnen-Bildern ... ausgefärtiget hat*
- **Google Translate:** `The Song of Songs of King Solomon: As it is sung/ To the souls who are reborn from God and/ through the contemplation of heavenly things/ in love with God, spirit-fiery love and cultivation/ according to the order of the text/ sung in scriptural terms; and/ with graceful copper and sensual images...`
- **WMT19:** `The Song of Solomon: As it / To the souls born again of God and / through the contemplation of heavenly things / fallen in love with God, the spirit of fiery exercise of love and cooling / as explained in writing according to the order of the text; and / with graceful copper and sensual images...`
- **URL:** https://www.ddb.de/item/6PQAFR3SSP6F5OZPKSIYCRTSFWXP5CAO
- **Metadata:** https://www.ddb.de/item/xml/6PQAFR3SSP6F5OZPKSIYCRTSFWXP5CAO

A.4. Example 5: The agent roles of Ignaz von Born and Johann Jacob Ferber.

- **Title:** *Des Hrn. Ignatz, Edl. von Born, Ritters, K.K. Berg-Raths, der Königl. Akademie der Wissenschaften zu Stockholm, der Großherzogl. zu Siena, u. der Georg. gelehrt. Gesellschaft zu Padua Mitglieds [et]c. Briefe über Mineralogische Gegenstände, auf seiner Reise durch das Temeswarer Bannat, Siebenbürgen, Ober- und Nieder-Hungarn, an den Herausgeber derselben, Johann Jacob Ferber, Mitglied der Königl. Grßherzogl. Akademie der*

Wissenschaften zu Siena, und der Ackerbau-Gesellschaft zu Vicenza und zu Florenz, geschrieben

- **Google Translate:** Of Mr. Ignatz, Edl. by Born, Ritters, K.K. Berg-Raths, the king. Academy of Sciences in Stockholm, the Grand Duke. to Siena, and the George. taught. Society of Padua members [et]c. Letters about mineralogical objects, on his journey through the Timisoara Bannat, Transylvania, Upper and Lower Hungary, to the editor of the same, Johann Jacob Ferber, member of the Royal. Grand Duke Academy of Sciences in Siena, and the Agricultural Society in Vicenza and Florence
- **WMT19:** Mr. Ignatz, Edl. von Born, Ritter, K.K. Berg-Raths, the Royal Academy of Sciences of Stockholm, the Grand Duchy of Siena, and Georg. erudite. Society of Padua member [et] c. Letters on mineralogical objects, on his journey through the Timisoara Bannat, Transylvania, Upper and Lower Hungary, to the editor of the same, Johann Jacob Ferber, member of the Royal Grand Ducal Academy of Sciences of Siena, and the Agricultural Society of Vicenza and Florence, written
- **URL:** https://www.ddb.de/item/H72GA6ENNNPAPYL5EYR4O2WHCJIOCJQP
- **Metadata:** https://www.ddb.de/item/xml/H72GA6ENNNPAPYL5EYR4O2WHCJIOCJQP

B. Appendix B

Sample approximate name matches from section 5.3

1. <ddb:2WQOAYCEL4TN4W5FB2ODYCUEX4KTSKC5> <dc:creator>: **Franz Xaver Brean**
   LLM Response:
   1. Augustine Baarfüssern
   2. Franciscus Xaverius Brean

2. <ddb:X4DLZTZUIPLDGGJURXV5EAFHJIW7YX2K> <dc:creator>: **Hendrik van Deventer**
   LLM Response:
   1. Henrici à Deventer
   2. Henrici à Deventer, Med. Doct.

3. <ddb:UT7GE4Y6URQ64QGDW7WA4HIMTUXI2OPN> <dc:creator>: **Georg Augustin Holle**
   LLM Response:
   1. Huber, Petrum
   2. Holler

4. <ddb:NJXCFR6YQKNOSYURGZ4TAZNFQ2KG7F4V> <dc:creator>: **Edward Wortley Montagu**
   LLM Response:
   1. E. W. Montagu

5. <ddb:ZJMLI4LH5R73SXZEH544TXLQ42FSP3CA> <dc:creator>: **Felix Loë**
   LLM Response:
   I. Deputy Baron Felix von Loë
   (Authors of the stenographic reports) Prussian House of Representatives reporters.)

6. <ddb:JOVL2DKWUICA6464DMGLMHHFI6VRZAL7> <dc:creator>: **Johann Hinrich Lehmann**
   LLM Response:
   1. Johann Hinrich Lehmann
   2. Nicol
   (Note: The given text seems to be written in old German, it might be a part of an obituary or a will, mentioning Johann Hinrich Lehmann and Nicol as authors)

C. Appendix C

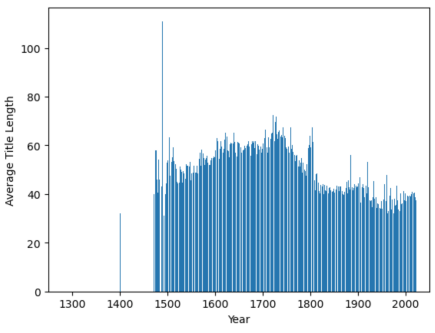

Figure 4: Distribution of title lengths after pruning.