ORIGINAL ARTICLE

LEARNED PUBLISHING

(wileyonlinelibrary.com) doi: 10.1002/leap.1610 Received: 21 September 2023 | Accepted: 22 April 2024 | Published online in Wiley Online Library: 17 May 2024

Exploring named-entity recognition techniques for academic books

Pablo Calleja Ibañez [1] and Elea Giménez-Toledo [2*]


Pablo Calleja Ibañez


Elea Giménez-Toledo

[1]Artificial Intelligence Department, Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain

[2]Interdisciplinary Thematic Platform ES CIENCIA, Institute of Language, Literature & Anthropology (ILLA), Spanish National Research Council (CSIC), Madrid, Spain

ORCID:
P. Calleja Ibañez: 0000-0001-8423-8240
E. Giménez-Toledo: 0000-0001-5425-0003

*Corresponding author: Elea Giménez-Toledo, Interdisciplinary Thematic Platform ES CIENCIA, Institute of Language, Literature & Anthropology (ILLA), Spanish National Research Council (CSIC), Madrid, Spain.
E-mail: elea.gimenez@cchs.csic.es

Abstract

Recent advances in the natural language processing (NLP) field have achieved impressive results in various tasks. However, NLP techniques are underrepresented in the analysis of Humanities and Social Science texts and in languages other than English. In particular, academic books are a highly valuable source of information that has not been exploited by these techniques at all. The recognition of named entities (person names, organizations or locations) and their semantic annotation over books could enrich the visibility and discoverability of the information by users. This is an opportunity for academia and the academic publishing industry in which semantic search is a central task and now books can be queried by named entities of interest that are in their content. This work proposes a methodology to apply named-entity recognition to publish the results into an ontological semantic-web format. The work has been performed over a corpus of academic books provided by UNE (*Unión de Editoriales Universitarias Españolas*, Union of Spanish University Presses). Results show an enrichment of the information extracted over the books and of the possibilities of querying them at the individual level but also within the whole set of books, increasing the possibilities for books to be discovered or retrieved beyond metadata.

Keywords: academic books, discoverability, multilingualism, name entity recognition (NER), onomastic index, ontology, semantic-web

INTRODUCTION

Natural language processing (NLP) techniques and artificial intelligence make it possible to enter into large sets of academic texts and perform different types of analysis on them, giving rise to applications that are both surprising and valuable. Uses include the possibility of performing complete reviews of the scientific literature on a topic (Wagner et al., 2022), to the development of automatic indexing systems (Du et al., 2019); systems that allow the analysis of large sets of articles in repositories and check whether they are supported or contested by other scientific articles (Kamshi, 2020); automatic generation of abstracts or reports (Monshi et al., 2020; Roberts & Fisher, 2020); tools for suggesting proper journals according to certain topics; Zero GPT to detect manuscripts that may have been written using Chat GPT; software to identify potential reviewers or to predict the possibility of accepting or rejecting a manuscript by analysing reviewers'

reports (Kousha & Thelwall, 2023) or retrieval and discoverability of scientific information.

The *Helsinki Initiative on multilingualism in science communication* (Helsinki Initiative on Multilingualism in Scholarly Communication, 2019) and the agreement of the *Coalition for advancing research assessment* (CoARA, 2022), which advocates the valuation of the different forms of scientific production regardless of the language in which they are produced, paint a different and enriching picture for the recognition of the different ways of transferring scientific knowledge in different languages. In fact, one of the working groups recently approved by CoARA is focused on multilingualism in scholarly communication. And although these initiatives allow us to initiate a very important cultural change, which implies richness and plurality in the ways of communicating science, there is another challenge ahead. Working with NLP techniques applied to corpora of scientific literature in a given language is crucial for the development of language models in each language. It will also allow scientific content produced in languages other than English to be more visible and discoverable in the digital environment. OPERAS European infrastructure has developed multilingual scientific content discovery system, GoTRIPLE (Achenbach et al., 2022; Di Donato et al., 2021), in which metadata, translation and multilingual terminology are fundamental to their performance and, ultimately, to the projection of scientific content in different languages. This is a significant example of how relevant it is to work with NLP in every language.

Objectives

This paper explores name-entity recognition (NER) techniques in scholarly publishing with the aim of answering these research questions:

- Is it possible to improve scholarly book discoverability using NER techniques? To what extent?
- Is it possible to go beyond information retrieval and discoverability based on metadata?
- What is the value of an ontology in academic content discoverability?

This paper presents the preliminary results of experimental research oriented to the discovery of named entities (NE) within academic books in Spanish.

The project aims to apply artificial intelligence techniques to corpora of scientific literature in Spanish and, specifically, to academic books in the Humanities and Social Sciences, which have rarely been the object of these analyses. It has been possible thanks to the collaboration with UNE (*Unión de Editoriales Universitarias Españolas*/Union of Spanish University Presses). The main objectives are:

- Extract substantive information, in particular that referring to recognized entities.

Key points

- Applying Name Entity Recognition techniques to a corpus of academic books in the Humanities and Social Sciences in Spanish.
- Exploring the benefits for publishing industry & multilingual information retrieval and discoverability tools.
- Development that seeks to make visible outstanding outputs in the SSH (books) and in languages other than English.
- With retrieved information represented in the ontology format, new knowledge could be queried and discovered.
- It serves as a powerful and huge onomastic index.

- Increase the possibilities for browsing large corporate contents.
- Increase the visibility of scientific content in Spanish and, with it, to some extent, contribute to the development of multilingual discovery tools.
- Explore applications of these techniques in the academic publishing sector.

NER techniques applied on XML of academic books in Humanities and Social Sciences allow the identification of places, institutions or people's names quite precisely. This type of analysis, which requires pointing techniques, software, system training, and so on, generates results similar to the traditional onomastic indexes of monographs or edited volumes, being able to answer questions such as in which part of this book does a certain writer, city or an institution appear? But the possibility of applying the technique to a set of books increases the power of the question: in which books from one or more publishers does a certain person, country or an international organization appear?

If entities identification, extraction, semantic tagging and adaptation to web data is performed on a large corpus of academic books in Spanish, the chances that the information in the books will be visible and retrievable for any user increases significantly (it may be worthwhile to synthesize all the techniques necessary to accomplish this process). This represents a priori a number of important opportunities for academia and for the publishing industry. In this study, we focus on two: information retrieval from the full text of books and discoverability of books from entities treated in the texts.

The exploratory study could be the basis for developments applied to other multilingual corpora, thus contributing to the discoverability of scientific content in different languages and/or with accessibility difficulties. The discoverability of academic books is often based on metadata as key elements and it is not always possible to retrieve them by words contained in the text, due to the lack of availability of the full texts (common in academic books that are not open access). This is why NER

techniques are explored in this paper for reaching academic contents that are not always discoverable. This is particularly important to Social Sciences and Humanities research, where scholarly books are the main channels for disseminating academic knowledge.

RELATED WORK

NER is a task in Information Extraction which finds and labels meaningful pieces of information called NE. NER also plays an essential role in many NLP tasks such as text understanding (Cheng & Erk, 2020; Zhang et al., 2019), information retrieval (Guo et al., 2009; Petkova & Croft, 2007), automatic text summarization (Aone C. et al., 1998), question answering (Mollá et al., 2006), machine translation (Babych & Hartley, 2003), and so on.

The term 'Named Entity' was first used at the 6th Message Understanding Conference (MUC) (Grishman & Sundheim, 1996), as the importance of the semantic identification of organizations, people and geographic locations in the text, as well as numerical expressions such as time and quantities. Since then, there has been an increased interest in NER with various scientific events such as CoNLL2003 (Tjong Kim Sang & De Meulder, 2003), ACE (Doddington et al., 2004) and IREX (Demartini et al., 2009).

NER tasks were traditionally oriented to identify a span text and classify in a group persons, locations, organizations or miscellaneous. However, other domains have defined their own classification groups, such as the biological domain, in which the groups are disease names, proteins, DNA, and so on. Currently, the state of the art is not focused only on those traditional groups; new challenges explore terms or words that belong to a particular group such as profession names (Miranda-Escalada et al., 2021).

NER has been applied in other different domains such as medicine (Song et al., 2020) or legal (Tamper et al., 2020) but there are fewer NER approaches used in Social Sciences and Humanities. However, there are relevant works in which the NER process has been applied in this domain, particularly, in Literature Studies (Rodrigues Alves et al., 2018; Van Dalen-Oskam et al., 2014), in Historical Newspapers (Ehrmann et al., 2020), in Automatic Indexation (Goh, 2017), in social media texts (Nie et al., 2020) or for creating social networks from novels (Dekker et al., 2019). However, there are not works focused on the NER process in academic books to exploit the retrieved information for indexation or for enriching their content for their search.

On the other hand, the semantic web and the ontological field have been focused on the representation of books with the target of making them more accessible and creating relations between them. For instance, works such as Thanapalasingam et al. (2018) and Dabrowski et al. (2009) put emphasis on the ontological representation of books from book publishing houses or other sources, and their properties and metadata. Moreover, there are works (Peroni & Shotton, 2012) that propose ontologies for the representation of bibliographic resources and citations along books. However, these works are not oriented to exploit the content inside the books and the mentions of named entities that appear in them. Those named entities contain highly valuable information about their content, their relations with other named entities in the text and their relations with the authors that are writing them. This work proposes an ontology to capture important metadata of the books and important content, such as the named entities that appear in them, with a NER process to create a scenario in which more complex information can be retrieved from it.

METHODOLOGY

Figure 1 presents the overview of the proposed methodology which is divided into four tasks: PDF conversion, database indexation, information extraction and ontology mapping. The following subsections will detail each of the tasks.

The proposed methodology has been applied in a real use case in collaboration with UNE (Union of Spanish University Presses), which provided a book corpus and a real scenario in which they work. They provided a set of books from two different university presses – PUZ (University of Zaragoza Press) and UGR (University of Granada Press). A total of 780 books of heterogeneous domains were obtained with their metadata.

Document conversion

The first task consists of the conversion of PDF documents into structured data that can be accessed by a computer. For this task, the tool GROBID[1] was used to transform PDF into TEI XML documents. TEI is a text-centric community of practice in the academic field of digital humanities that has defined an XML format that can represent digital documents correctly following the standards of the W3C.

The GROBID tool has been trained for scientific papers, but it can be used correctly in books. The most interesting result that GROBID provides is the correct identification of sections and paragraphs even if they are split over different pages or if there are images with caption labels inside them. In this first task of the methodology, a set of 780 TEI XML documents were created with all their sections and paragraphs detected.

Document segmentation and database indexation

Once books are transformed into a structured format that can be exploited, an extraction of sections and paragraphs is performed. Sections of the documents in the TEI XML format are presented between the labels and paragraphs between the labels <p>. Both are standard labels in HTML format.

Then, sections and paragraphs are indexed into a text database. The text database selected is Elasticsearch in which each section and each paragraph are treated as independent

[1]Grobid. https://github.com/kermitt2/grobid

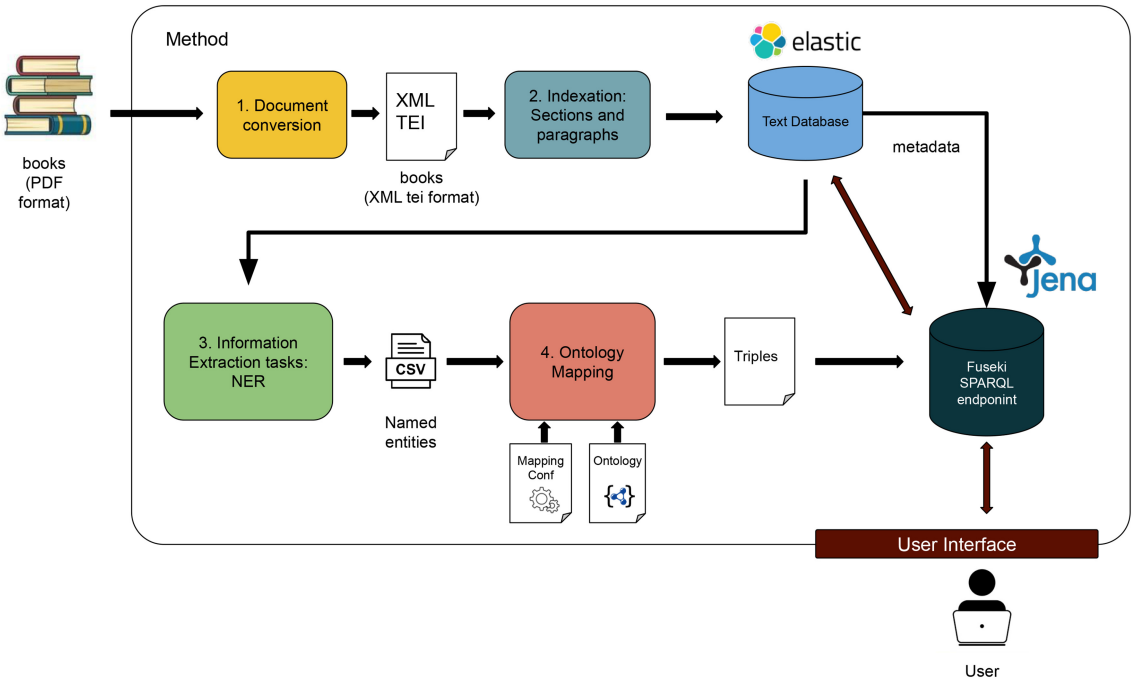

FIGURE 1 Overview of the methodology.

book_id	section	paragraph	text	init	end	type
9788416028672	11	1	Ámsterdam	190	199	location
9788416028672	11	1	Stadionweg	353	363	location
9788416028672	11	1	Amstelveenseweg	365	380	location
9788416028672	11	1	Évora	646	651	location
9788416028672	11	1	Siza	630	634	location
9788416028672	11	1	París	814	819	location
9788416028672	11	2	Rambla	195	201	location
9788416028672	11	2	Parque del Besòs	229	245	location
9788416028672	11	2	Ronda del Litoral	285	302	location

FIGURE 2 Excerpt of the CSV file. It shows nine location named entities of the book 9788416028672, in section 11 and paragraphs 1 and 2.

documents in the database. The format of the identifiers of the documents in the database uses the original book identifier, the section number, and the paragraph number inside the section. Thus, the second paragraph of the first section of a book is stored as *bookId_1_2*. From the 780 TEI XML documents, a set of 62.234 sections and 550.716 paragraphs were detected and stored in an Elasticsearch database.

Information extraction task

Information Extraction is considered an area in which different tasks transform unstructured information (texts) to structured data. From the several tasks that can be performed over texts, this work is focused on one of the most relevant ones, which is NER. The current state of the art is focused on deep learning models based on a fine-tuned version of a language model, which are trained with corpora annotated with named entities (span text and category).

To detect the named entities on the text database, each paragraph is extracted from the database and divided into sentences. Then, each sentence is processed by the NER model. The results obtained are stored in a CSV format in which the information about the book, the section, the paragraph, the span text, the category and the offsets (where the span text starts and ends inside the sentence) is stored. The offsets are local to the paragraphs, so each paragraph starts at position 0. Figure 2 shows an excerpt of the CSV file with named entities of location type (commas have been substituted by tabulations for presentation).

For this task, an already-trained model[2] for NER is used. This model is based on the language model BETO (Cañete et al., 2020) trained with the dataset CoNNL (Tjong Kim Sang, 2002), which contains organization, persons, locations and miscellaneous entities annotated. Thus, this model is trained in the detection of

[2]https://huggingface.co/mrm8488/bert-spanish-cased-finetuned-ner

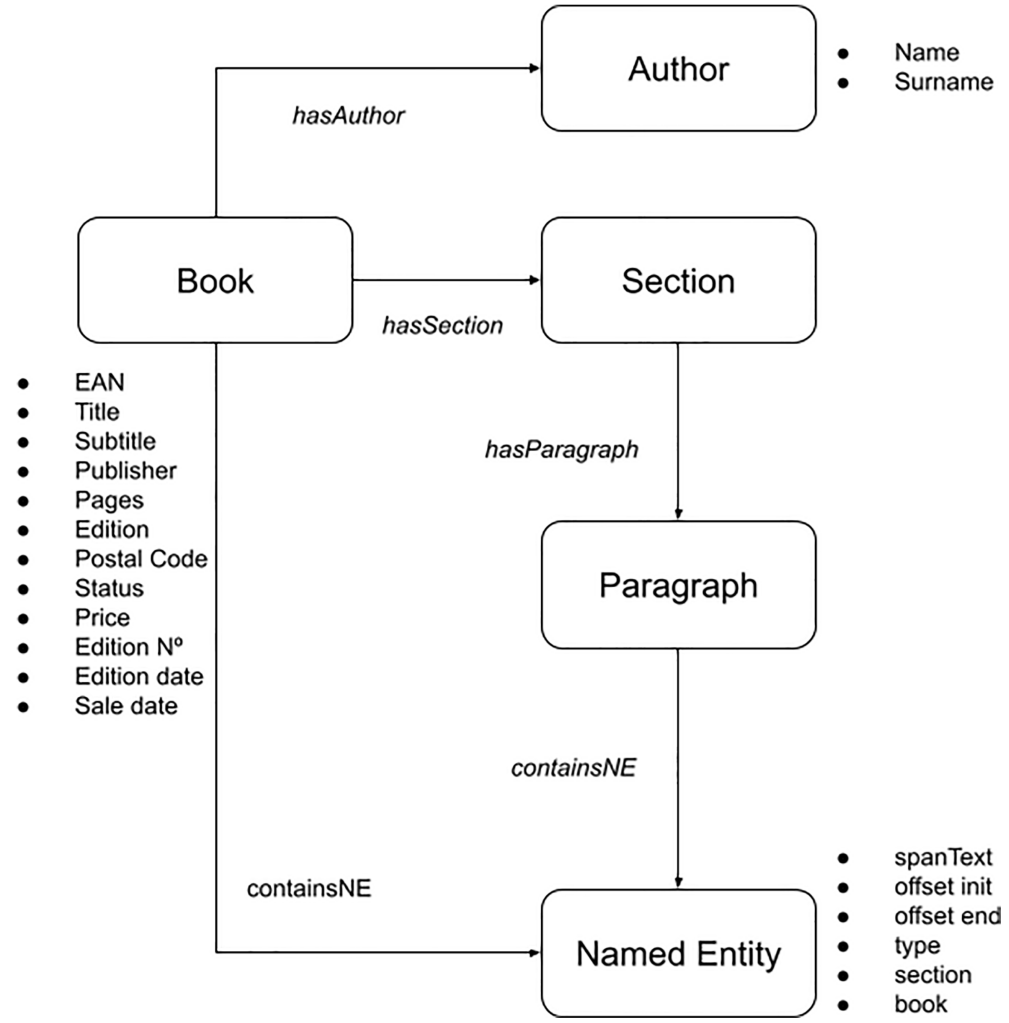

FIGURE 3 Ontology for the representation of books, authors, section, paragraphs and named entities.

named entities of the same type. A total of 1,342,385 persons, 441,462 organizations, 755,758 locations and 857,258 miscellaneous named entities were detected and stored in a CSV file.

Ontology mapping and storage

Once the named entities are detected and stored in a CSV file, the results have to be mapped to populate an ontology that represents the information about books, sections, paragraphs and named entities. This ontology has been defined previously for the representation of UNE main information and it has been developed in collaboration with UNE experts. The resulting triples obtained in the mapping process are then stored in an Apache Jena Fuseki.

This ontology mapping process is made by scripts described in a language named RDF Mapping Language (RML) (Dimou et al., 2014). However, to ease the creation of the scripts, the YARRRML language was used (Heyvaert et al., 2018). YARRRML is a human readable text-based representation for declarative Linked Data generation rules and eases the creation of RML from sources such as CSV with simple language that indicates the information stored in each column and how a correct mapping is created.

Ontology definition

An ontology has been defined to represent the information extracted from the books. An overview of the ontology is represented in Fig. 3. The main concept is *Book* which contains attributes that represent all the main metadata provided by UNE, such as EAN, Title, Status, Price or Publisher. *Author* has been defined as another class in which attributes of name and surname are represented. Future work will represent others such as nationality, age, date of birth or date of death. *Book* and *Author* are related by the relation *hasAuthor*, which can represent that a book has more than one author.

The classes *Section* and *Paragraph* are oriented to represent the sections and paragraphs in which a book is divided. The

```
prefixes:
  une: http://une.linkeddata.es/scheme/
mappings:
  book:
    sources:
      - [ejemplodataset.tsv~csv]
    s: une:book/$(book_id)
    po:
      - [une:hasSection, une:section/$(book_id)/$(section)~iri]
      - [une:hasEntity, une:entity/$(book_id)/$(section)/$(paragraph)/$(init)_$(end)~iri]
      - [a, une:Book]
  entity:
    sources:
     - [ejemplodataset.tsv~csv]
    subjects: une:entity/$(book_id)/$(section)/$(paragraph)/$(init)_$(end)
    po:
      - [une:type, une:type/$(type)~iri]
      - [une:init, $(init),xsd:integer]
      - [une:end, $(end),xsd:integer]
      - [une:spanText, $(text),xsd:string]
```

FIGURE 4 An excerpt of the YARRRML script for the generation of RML rules that maps the content in the CSV file into triples of the ontology.

relations are *hasSection* and *hasParagraph*. Then, the class *Named Entity* is defined and is related by *containsNE* to *Paragraph*, to show the named entities that appear in it, and to *Book*, to represent all the named entities that a book mentions. The class *Named Entity* contains the attributes of Span Text, the init offset, the end offset, the type of named entity, the section and the book in which it appears.

Mapping process

As mentioned before, a script in RML language is created to translate the information stored in the CSV file (the named entities, their features and their relations with the paragraphs, sections and books) to RDF triples in which the linked data information is stored. As RML is not an easy language description for humans, a YARRRML script is used to create the RML mappings. YARRRML is a human readable text-based representation for declarative Linked Data generation rules and it eases the creation of triples (subject, predicate, object) from CSVs.

An excerpt of the YARRRML script is presented in Fig. 4. The script starts defining the prefixes (*prefixes*) that will be used in the URIs created for each resource. In this case, a new prefix is created for this project named UNE (http://une.linkeddata.es/scheme). Then, the mapping rules are defined for each class of the ontology. In the presented excerpt of Fig. 4, the mapping rules for book and entity are presented.

Following the first named entity example of Fig. 2 (book_id = 9788416028672, section = 11, paragraphs = 1, text = Amsterdam, init = 190, end = 199, type = location), the YARRRML works as follows. First, for the definition of the class *book*, the source is defined, which is the CSV in which the information is stored. Then, the subject (*s*) is defined. In this case, the rule for the creation of the book is defined by the prefix plus the word book and the value of the column in the CSV file: *une:book/9788416028672*. The script defines the value of the name of the column using the dollar symbol such as *$(COLUMN_NAME)*.

Then, the script defines the predicates and objects (*po*) of the subject *book*. First, the section and its relation with the book is created. The relation is created as defined in the ontology with *une:hasSection* and the section is defined with URI *une:book/9788416028672/1* which is the section 1. The class book is also related with the named entity and the next rule creates this relation *une:hasEntity* and the named entity URI *une:entity/9788416028672/1/1/190_199* for the example of Amsterdam named entity. Finally, the last rule defines the relation of the subject with its class, the *une:book/9788416028672* is *a* (instance of) *une:Book*.

Once the YARRRML scripts have been executed and the RML mapping files have been created, the tool Morph-KGC (Arenas-Guerrero et al., 2022) is used as the engine to generate the triples. For storage and publication, a public Apache Jena Fuseki datastore[3] is used, which provides a SPARQL endpoint to perform queries over it. The Jena Fuseki of the project is presented publicly.[4]

RESULTS

Once the ontology has been populated and published in the public Jena Fuseki datastore, users can start making queries and obtaining results for their benefit. In this project, the queries and their results are being evaluated by UNE experts in their own use cases. UNE provided a set of queries of interest that can now

[3] https://jena.apache.org/index.html

[4] https://fuseki.pcalleja.linkeddata.es/dataset.html

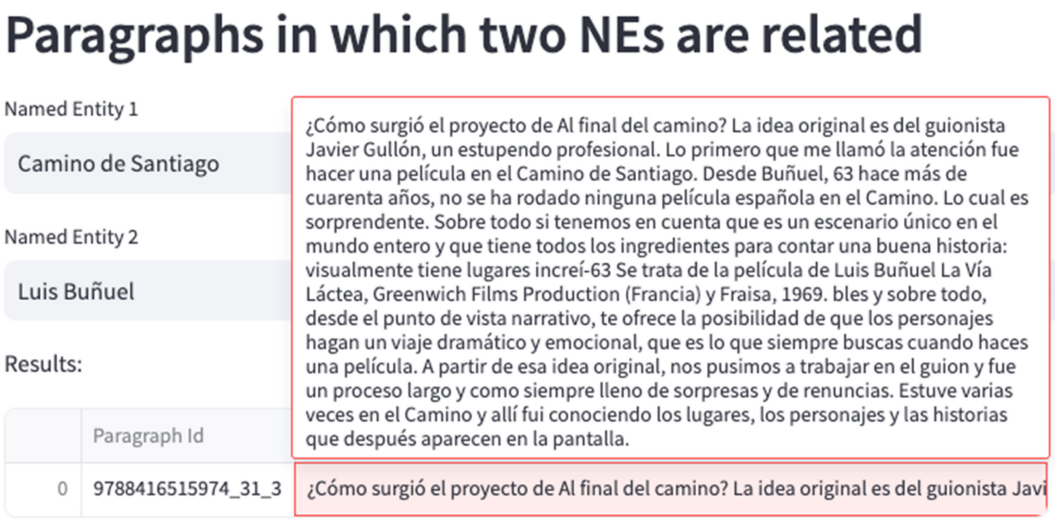

FIGURE 5 A web interface in which the user includes two named entities and the ontology finds the result and the database shows the text. In this example, the retrieved paragraph contains the named entities 'Luis Buñuel' and 'Camino de Santiago'.

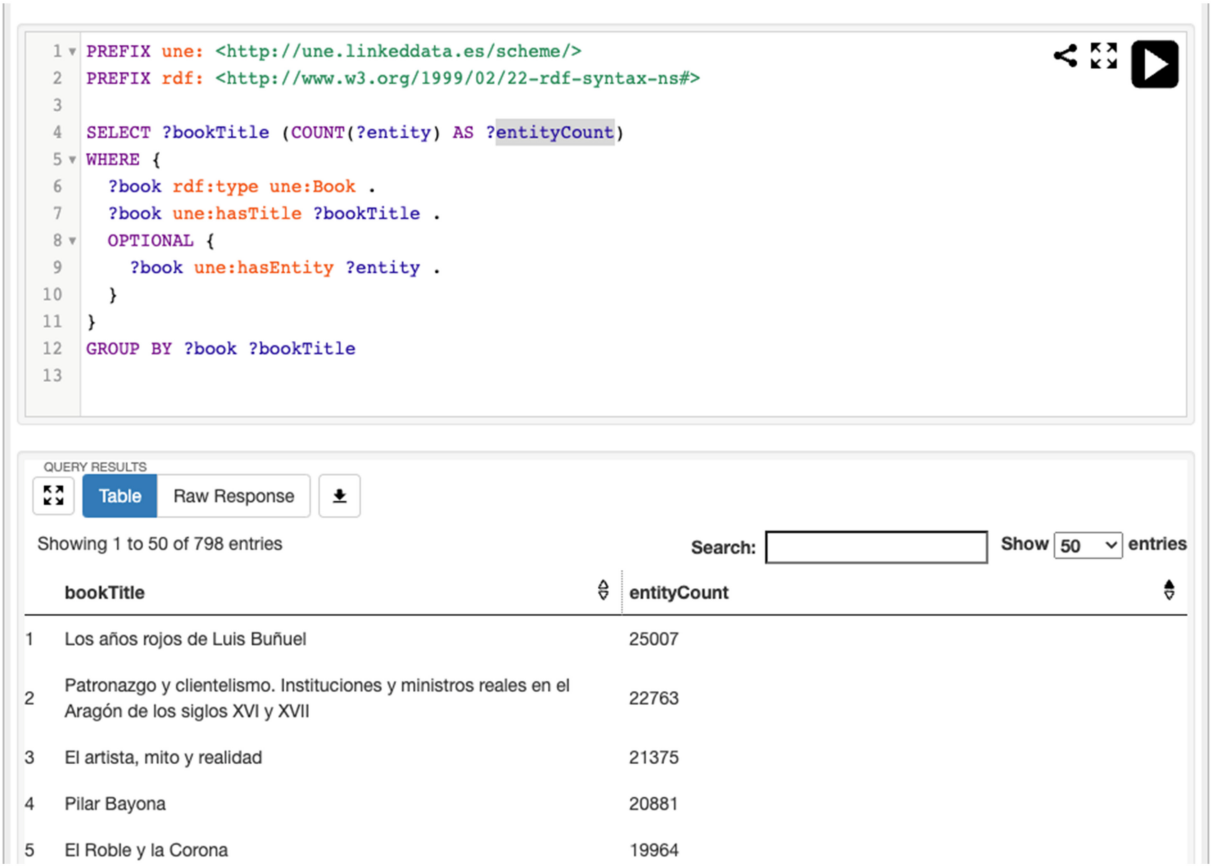

FIGURE 6 SPARQL endpoint showing the number of named entities recognized in each book and the title of the book.

be performed. In particular, queries that could be performed over the books included:

- Get books in which a particular named entity appears
- Get paragraphs in which two particular named entities are mentioned. Users could search if there is a particular relation between them.
- Get the authors that know about a particular named entity. If there is an interest about a specific person or location a query can search those authors that have talked about it in their books.
- Get the main topics of an author relating to the locations, persons or other named entities that the author talks about in their books.

These queries can be verbalized through SPARQL language queries and the results can be jointly retrieved with the Elasticsearch database in which the sections and paragraphs are stored. For instance, query 1, get books in which a particular named entity (e.g., '*Camino de Santiago*') appear and the price of the book could be expressed in SPARQL as:

```
PREFIX une: http://une.linkeddata.es/scheme/%3e
SELECT ?texttitle ?price
WHERE {
?book une:hasTitle ?texttitle .
?book une:price ?price .
?book ?r ?e .
?e une:spanText 'Camino de Santiago' .
} GROUP BY ?texttitle ?price LIMIT 1000
```

On the other hand, web interfaces such as the one presented in Fig. 5 can show the result of query 2, paragraphs in which two named entities occur and the text of the paragraph.

The SPARQL endpoint allows users to generate different queries on demand, explore the data and get quantitative results about the books. For instance, Fig. 6 shows the number of total entities recognized in each book, and Fig. 7 shows the most repeated entities in all the books.

The application of these techniques on this type of corpus allows the retrieval of information beyond the metadata, as well as improving the discoverability of academic books in Spanish (in this case), thus answering the main research questions of this work. On the one hand, users searching for information about an

```
PREFIX une: <http://une.linkeddata.es/scheme/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT  ?entityLabel (COUNT(?entityLabel) AS ?entityCount)
WHERE {
  ?book rdf:type une:Book .
  ?book une:hasEntity ?entity .
  ?entity une:spanText ?entityLabel .
}
GROUP BY  ?entityLabel
ORDER BY DESC(?entityCount)
LIMIT 20
```

QUERY RESULTS

Table | Raw Response

Showing 1 to 20 of 20 entries

Search: Show 50 entries

	entityLabel	entityCount
1	España	41972
2	Estado	23541
3	Madrid	22254
4	Granada	13690
5	Francia	13372
6	Dios	13018
7	Zaragoza	12377
8	Europa	12346

FIGURE 7 SPARQL endpoint showing the most repeated named entities in all the books and the number of occurrences.

'entity' will be able to retrieve results that are normally opaque, since the content of academic books is not usually accessible, except in the case of open access books, works released from copyright or under prior negotiations (as in the case of Google Books). In other words, the technique makes it possible to delve into the content of a book beyond its metadata, which are the usual elements of information retrieval.

By working with a corpus in languages other than English (in this case, Spanish), different entities will be identified and the retrieval of scientific content in Spanish or multilingual languages will be made possible, if connections can be established with vocabularies in different languages, something that has been developed within Go TRIPLE. Developments such as the one presented in this paper could be integrated into search engines such as Isidore or others.

On the other hand, entity searches in a set of books can become a kind of powerful index that will allow the discovery of titles that deal with certain places, people or institutions of interest to the user. Thus, the technique would make it possible to discover book titles that are not normally very visible on scientific content discovery platforms. Overall, this represents an opportunity for publishers to bring their content closer to potential readers, to sell more or increase their access numbers, in the case of OA books and, of course, to offer content in the form of books and in Spanish, helping to diversify the results and content of discovery platforms.

CONCLUSIONS AND FUTURE WORK

This exploratory study has made it possible to apply artificial intelligence techniques, in particular NER, to a corpus of 780 academic books in Spanish. It has been possible to identify 441,462 organizations, 1,342,385 persons, 755,758 locations and 857,258 miscellaneous entities.

Answering the research question of *What is the value of an ontology in academic content discoverability?*, the created system shows the utility of using an ontology for representing the information extracted from books as well as the potential for retrieval and discoverability of scientific information in large corpora of scientific literature (in this case, in Spanish and from academic books). Such a system, systematized and extended, performs the functions of a traditional onomastic index, but with much more efficiency and power, since it allows the retrieval of millions of entities within a particular book or a large set of books. This makes it possible to search and browse the content of scholarly books, usually not very discoverable unless they are open access or full-text and machine-readable after negotiation of rights. The required queries by UNE validates its usability, our second research question.

For the first research question, *Is it possible to improve scholarly book discoverability using NER techniques? To what extent?*, this work has demonstrated the potential of a NER model based on Language Models for identifying names in books to populate the proposed ontology.

It should be noted that the existence and richness of any Language Model is crucial for creating and populating these ontologies and, consequently, to be able to make the contents of the books discoverable by means of the entities. This is to say, the more robust the Language Model in each language is, the more useful the ontology will be for discovering contents.

Also, the implementation of NER techniques for this study highlights the relevance of having multilingual full-text corpora available as well as the quality of Language Models to train NER models. Both are critical elements for proposing these techniques as powerful tools for discovering contents beyond metadata and diving into the richness of a set of academic books.

As future lines of work, there are several processes that can be improved. For instance, information extraction can be extended to other NER models for other types of named entities or other information tasks can be included such as co-reference resolution or entity linking. Co-reference resolution aims to identify those named entities that have different span texts mentioning the same concept. As an example of this, variants of persons such as 'Luis Buñuel', 'Buñuel' and 'Luis' in the same section has a high probabilities of being the same concept in the three cases. Moreover, co-reference resolution could identify pronouns (he/she) that also mention the same concept. Entity linking aims to identify the concept of the named entity in another knowledge graph and explore its relations and properties. For instance, the location 'Lorca' is represented in Wikidata as Q475717, which is a municipality of Murcia, Spain. This entity should be linked to another graph where Lorca refers to the renowned Spanish writer Federico García Lorca.

This experiment seeks to make visible outstanding outputs in the Social Sciences and the Humanities (books) and in different languages, in line with all the international initiatives for boosting multilingualism in scholarly communication (GoTRIPLE, Helsinki Initiative, COARA principles, etc.). It could be a powerful tool for academic book discovery useful for (a) researchers or potential readers, reaching contents usually hidden or difficult to find; and (b) academic book publishers, increasing visibility of their titles. Moreover, the combination of entities and metadata (i.e., authors or institutions) could also provide a solution for expert discovery, useful for journalist and communication departments as well as for academic publishers, looking for new authors or reviewers.

AUTHOR CONTRIBUTIONS

Pablo Calleja Ibañez and Elea Giménez-Toledo: Conceived the project. **Pablo Calleja Ibañez:** Developed the NER procedure. **Elea Giménez-Toledo:** Analysed the possibilities of the techniques within publishing industry. Both authors analysed the results and wrote the article.

ACKNOWLEDGEMENTS

Financed by the European Union. NextGeneration EU. Margarita Salas Program (UP2021-035). Supported by CSIC Interdisciplinary Thematic Platform (PTI) Spanish as language of science (Español como lengua de comunicación científica) (PTI-ES CIENCIA).

Authors want to thank UNE for providing access to the corpus of books for experimental development.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available on request from Unión de Editoriales Universitarias Españolas (UNE). Restrictions apply to the availability of these data, which were used under licence for this study. Data are available from the authors with the permission of UNE.

REFERENCES

Achenbach, K., Błaszczyńska, M., De Paoli, S., Di Donato, F., Dumouchel, S., Forbes, P., Kraker, P., & Vignoli, M. (2022). Defining discovery: Is google scholar a discovery platform? An essay on the need for a new approach to scholarly discovery. *Open Research Europe*, *2*(28). https://doi.org/10.12688/openreseurope.14318.1

Aone, C., Okurowski, M. E., & Gorlinsky, J. (1998). Trainable, Scalable Sumamization Using Robust NLP and Machine Leearning. *36th Annual Meeting of the Association for Computational Linguistics* (Vol. *1*, pp. 62-66).

Arenas-Guerrero, J., Chaves-Fraga, D., Toledo, J., Pérez, M. S., & Corcho, O. (2022). Morph-KGC: Scalable knowledge graph materialization with mapping partitions. *Semantic Web*.

Babych, B., & Hartley, A. (2003). Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT Workshop on MT and Other Language Technology Tools, Improving MT Through Other Language Technology Tools, Resource and Tools for Building MT at EACL 2003*. Association for Computational Linguistics.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. PML4DC at ICLR 2020.

Cheng, P., & Erk, K. (2020). Attending to entities for better text understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. *34*, pp. 7554–7561). AAAI Press. https://doi.org/10.1609/aaai.v34i05.6254

Coalition for the Advacement of Research Assessment CoARA. (2022). *The Agreement on Reforming Research Assessment*. https://coara.eu/agreement/the-agreement-full-text/

Dabrowski, M., Synak, M., & Kruk, S. R. (2009). Bibliographic ontology. In *Semantic digital libraries* (pp. 103–122). Springer.

Dekker, N., Kuhn, T., & van Erp, M. (2019). Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science*, *5*, e189. https://doi.org/10.7717/peerj-cs.189

Demartini, G., Iofciu, T., & Vries, A. P. (2009). Overview of the inex 2009 entity ranking track. In *International Workshop of the Initiative for the Evaluation of XML Retrieval* (pp. 254–264). Springer.

Di Donato, F., Dumouchel, S., Monachini, M., & Pohle, S. (2021). The discovery platform gotriple: An eosc service for social sciences and humanities research. AIUCD 2021-DH per la società: e-guaglianza, partecipazione, diritti e valori nell'era digitale.

Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., & Van de Walle, R. (2014). RML: A generic language for integrated RDF mappings of heterogeneous data. In *Proceedings of the 7th Workshop on Linked Data on the Web*. Springer.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M., & Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC* (Vol. *2*, pp. 837–840). European Language Resources Association.

Du, J., Cunningham, R. M., Xiang, Y., Li, F., Jia, Y., Boom, J. A., Myneni, S., Bian, J., Luo, C., Chen, Y., & Tao, C. (2019). Leveraging deep learning to understand health beliefs about the human papillomavirus vaccine from social media. *npj Digital Medicine*, *2*(1), 27. https://doi.org/10.1038/s41746-019-0102-4

Ehrmann, M., Romanello, M., Flückiger, A., & Clematide, S. (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* (Vol. *2696*). CEUR-WS.

Goh, R. (2017). Using named entity recognition for automatic indexing.

Grishman, R., & Sundheim, B. M. (1996). Message understanding conference-6: A brief history. In *The 16th International Conference on Computational Linguistics*. Association of Computer Machinery.

Guo, J., Xu, G., Cheng, X., & Li, H. (2009). Named entity recognition in query. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 267–274). Association of Computer Machinery.

Helsinki Initiative on Multilingualism in Scholarly Communication. (2019). Helsinki: Federation of Finnish Learned Societies, Committee for Public Information, Finnish Association for Scholarly Publishing, Universities Norway & European Network for Research Evaluation in the Social Sciences and the Humanities. https://doi.org/10.6084/m9.figshare.7887059

Heyvaert, P., De Meester, B., Dimou, A., & Verborgh, R. (2018). Declarative rules for linked data generation at your fingertips! In *Proceedings of the 15th ESWC: Posters and Demos*. Springer.

Kamshi, R. (2020). Coronavirus in context: Scite.ai tracks positive and negative citations for COVID-19 literature. *Nature*. https://doi.org/10.1038/d41586-020-01324-6

Kousha, K., & Thelwall, M. (2023). *Artificial intelligence to support publishing and peer review: A summary and review*. Learned Publishing.

Miranda-Escalada, A., Farré-Maduell, E., Lima-López, S., Gascó, L., Briva-Iglesias, V., Agüero-Torales, M., & Krallinger, M. (2021). The profner shared task on automatic recognition of occupation mentions in social media: Systems, evaluation, guidelines, embeddings and corpora. In *Proceedings of the sixth social Media Mining for Health (SMM4H) Workshop and Shared Task* (pp. 13–20). Association for Computational Linguistics.

Mollá, D., Van Zaanen, M., & Smith, D. (2006). Named entity recognition for question answering. In *Proceedings of the Australasian Language Technology Workshop 2006* (pp. 51–58). Australian Language Technology Association.

Monshi, M. M. A., Poon, J., & Chung, V. (2020). Deep learning in generating radiology reports: A survey. *Artificial Intelligence in Medicine*, *106*, 101878. https://doi.org/10.1016/j.artmed.2020.101878

Nie, Y., Tian, Y., Wan, X., Song, Y., & Dai, B. (2020). Named entity recognition for social media texts with semantic augmentation.

Peroni, S., & Shotton, D. (2012). Fabio and cito: Ontologies for describing bibliographic resources and citations. *Journal of Web Semantics*, *17*, 33–43. https://doi.org/10.1016/j.websem.2012.08.001

Petkova, D., & Croft, W. B. (2007). Proximity-based document representation for named entity retrieval. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management* (pp. 731–740). Association for Computing Machinery.

Roberts, J., & Fisher, D. (2020). Preview: The artificially intelligent conference reviewer. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 665–668). IEEE.

Rodrigues Alves, D., Colavizza, G., & Kaplan, F. (2018). Deep reference mining from scholarly literature in the arts and humanities. *Frontiers in Research Metrics and Analytics*, *3*, 21.

Song, B., Bao, Z., Wang, Y., Zhang, W., & Sun, C. (2020). Incorporating lexicon for named entity recognition of traditional Chinese medicine books. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020* (pp. 481–489). Springer.

Tamper, M., Oksanen, A., Tuominen, J., Hietanen, A., & Hyvönen, E. (2020). Automatic annotation service appi: Named entity linking in legal domain. In A. Harth, V. Presutti, R. Troncy, M. Acosta, A. Polleres, J. D. Fernández, J. Xavier Parreira, O. Hartig, K. Hose, & M. Cochez (Eds.), *The semantic web: ESWC 2020 satellite events* (pp. 208–213). Springer International Publishing. https://doi.org/10.1007/978-3-030-62327-2

Thanapalasingam, T., Osborne, F., Birukou, A., & Motta, E. (2018). Ontology-based recommendation of editorial products. In D. Vrandečić, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee, & E. Simperl (Eds.), *The semantic web—ISWC 2018* (pp. 341–358). Springer International Publishing.

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. Association for Computational Linguistics.

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL* (Vol. *2003*, pp. 142–147). Association for Computational Linguistics.

Van Dalen-Oskam, K., de Does, J., Marx, M., Sijaranamual, I., Depuydt, K., Verheij, B., & Geirnaert, V. (2014). Named entity recognition and resolution for literary studies. *Computational Linguistics in the Netherlands Journal*, *4*, 121–136.

Wagner, G., Lukyanenko, R., & Paré, G. (2022). Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology*, *37*(2), 209–226. https://doi.org/10.1177/02683962211048201

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the ACL* (pp. 1441–1451). Asscociation for Computational Linguistics.