First steps towards a platform for the analysis of civil law documentary heritage

Primeros pasos hacia una plataforma para el análisis del patrimonio documental de derecho civil

Hala NEJI (1), Javier NOGUERAS-ISO (1),
Francisco Javier GARCÍA-MARCO (2), Carmen BAYOD LÓPEZ (2)

(1) Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, C/ Mariano Esquillor SN, 50018 Zaragoza, España, {hala.neji,jnog}@unizar.es. (2) Instituto de Patrimonio y Humanidades (IPH), Universidad de Zaragoza, C/ Pedro Cerbuna 12, 50009 Zaragoza, España, {jgarcia,cbayod}@unizar.es

Resumen

El patrimonio documental sobre derecho civil es un importante activo cuyo estudio permite conocer el contexto político, social y cultural de la época a la que se refieren los documentos históricos. Este trabajo presenta el diseño de un prototipo de plataforma para apoyar a los investigadores en el análisis del patrimonio documental sobre derecho civil. Nuestra plataforma consiste en crear una versión en línea de estos materiales, haciéndolos más accesibles. La plataforma proporciona asistentes automáticos para la transcripción, traducción y extracción de elementos de información específicos asociados a conceptos de derecho civil (voces) como citas a fuentes externas y entidades con nombre (lugares, personas y organizaciones) para identificar mejor su contexto. La viabilidad de esta plataforma se ha puesto a prueba con el tratamiento de una obra doctrinal escrita por Miguel del Molino, un conocido experto en derecho civil del siglo XV en el reino de Aragón.

Palabras clave: Derecho civil. Patrimonio documental. Procesamiento de textos. Reconocimiento de entidades nombradas. Aprendizaje automático. Molino, Miguel del.

Abstract

The documentary heritage about civil law is an important asset whose study allows us to learn about the political, social and cultural context of the period referred in historical documents. This paper presents the design of a prototype platform to sup-port researchers in the analysis of civil law docu-mentary heritage. Our platform involves creating an online version of these materials, making them more accessible. The platform provides automatic assistants for the transcription, translation and extraction of specific information items associated to civil law concepts (voices) such as citations to external sources and named entities (locations, persons, and organizations) to identify better their context. The feasibility of this platform has been tested with the processing of a doctrinal work writ-ten by Miguel del Molino, a well-known civil law expert in XV century in the Aragon kingdom.

Keywords: Civil law. Documentary heritage. Text processing. Named entity recognition. Deep learning. Miguel del.

1. Introduction

Civil law regulates the civil or private relations of persons: it deals with the civil status of persons, their family rights and duties, property and other real rights over things, the regime of obligations and contracts, and successions and inheritances. Although in Europe each country has its own civil law code at the state level, this law dates back to past times and has evolved until present times. In many administrative areas this law has its origin in historical kingdoms that regulated the conditions of settlement of their citizens. The books that publish these historical codes of civil law and the doctrinal works that document their interpretation make up an asset of cultural interest whose study allows us to learn about the political, social, and cultural context of this historical period.

Nowadays there are numerous initiatives (e.g. Europeana) that have promoted the digitization of documentary heritage, including that related to civil law. However, simple digitization is not enough. This work is a first step in a research project that aims to promote the study of works of special interest in the field of the history of civil law (specially, civil law authors of XVI, XVII and XVIII centuries) accompanied with the development of the necessary technology to have an online hypertext edition of the works that includes the digital facsimile, a critical edition, a translation and all the complements that are deemed relevant (e.g., legal codes cited and notes or glosses to the text both legal and historical-philological or bibliographical).

The objective of this work is to present a first prototype of the platform that will provide support for the analysis of civil law documentary heritage. Figure 1 presents a use case diagram which outlines various functionalities of the envisioned web platform. Apart from the typical tools for searching, browsing and visualization available for the general public, this platform integrates advanced tools to support transcription, translation, and legal analysis of the books ingested by experts in civil law. With respect to transcription and translation, we aim to customize existent OCR and automated translation algorithms. Related to the legal analysis, we aim to integrate text mining tools able to extract context information from the text defining how civil law concepts should be interpreted. This context consists of citations to regulations and named entities such as locations or authorities (persons and organisations).

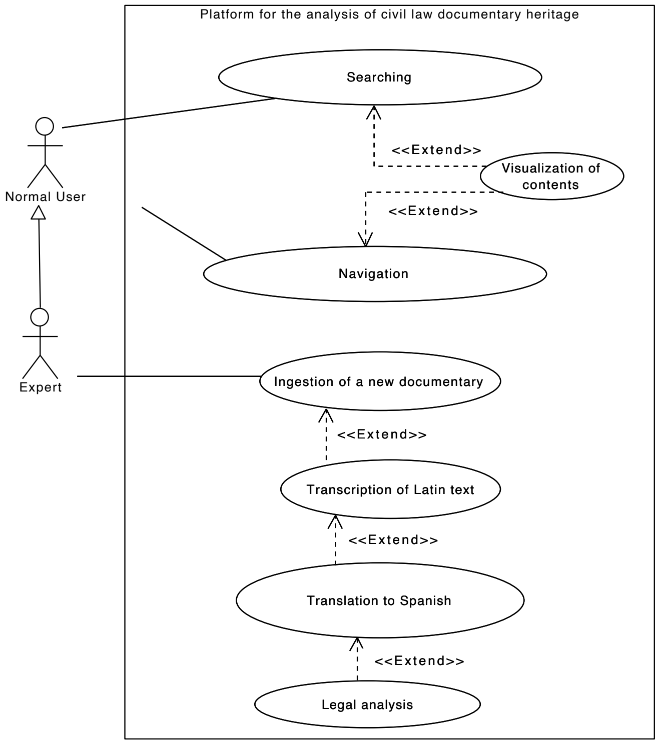

Platform for the analysis of civil law documentary heritage
Searching
<<Extend>>
Visualization of contents
<<Extend>>
Normal User
Navigation
Expert
Ingestion of a new documentary
<<Extend>>
Transcription of Latin text
<<Extend>>
Translation to Spanish
<<Extend>>
Legal analysis

Figure 1. Use case diagram

The rest of this work is organized as follows. Section 2 reviews the state-of-the-art in documentary heritage analysis. Section 3 describes the design of the platform. Then, section 4 presents some experimental results of the platform for the analysis of a work of Miguel del Molino, a well-known civil law expert in XV century in the Aragon kingdom (currently an autonomous community in Spain). Finally, section 5 provides some concluding remarks.

2. State of the art

This platform presents important challenges regarding transcription, translation and named entity recognition.

With respect to the *transcription* task, it must be noted that the images from the digitization of historical documents present various levels of degradation due to the use of handmade fonts, ink stains on the paper or the noise generated during the digitization process (interference from double-sided printed text, deviations, blurring in double-page scans, etc.) (Gupta et al., 2015). All of this makes optical character recognition (OCR) difficult and motivates research into the pre-processing of images to eliminate noise (Neji et al., 2024), or the training of character recognition models specialized in Gothic and round letters using machine learning based on neural networks (Lacasta et al., 2022; Kodym et al., 2021).

It is also worth noting the existence of the Pero-OCR tool, which is semi-supervised machine learning method for automatic handwritten and printed text transcription (Kišš et al., 2023). It employs a SoftCTC (Connectionist Temporal Classification) loss function that allows to manage complex transcription scenarios.

In addition, automated *translation* from Latin to Spanish also represents an important challenge. Although there are currently numerous online translators (Google Translator, Yandex, DeepL, Translateking, imTranslator or Translateking, among others) and some configurable open source tools (for example, NVIDIA NeMo or OpenNMT), the translation of medieval Latin is not very advanced due to the scarcity of parallel Latin and Spanish corpora in different domains (Tiedemann et al., 2012).

There are incipient works that exploit deep neural networks based on transformer-type architectures (Transformer) for this type of problems. For instance, Martínez García and García Tejedor (2020) developed an advanced Neural Machine Translation system for Latin-Spanish, aiming to make historical texts more accessible. Using Transformer-based models trained on Bible and Saint Augustine corpora, the study explores domain adaptation challenges and emphasizes the significance of sufficient data for accurate translations, especially for low-resourced languages like Latin, but for the moment each translation work in this context requires the preparation of a personalized training corpus. Fischer et al. (2022) also crafted a dedicated Latin-German Neural Machine Translation (NMT) system for 16th-century letter translation. Their meticulous data collection and NMT model development led to superior translations for short to medium sentences, outperforming GoogleTranslate. While centered on Swiss reformer Heinrich Bullinger's correspondence, their work offers broader utility for translating texts from the era.

Thirdly, to facilitate the *legal analysis* of the works, it is relevant to adapt text mining techniques that allow the recognition of named entities such as place names, person names and organisation names to contextualize the voices included in the doctrinal works of civil law. Identifying people, places, and other historical entities is an essential task in automatic understanding of historical documents (Aljalbout and Falquet, 2017). Named entity recognition and classification (NER for short) is very often the first step of entity linking, which can support the cross-linking of multilingual and heterogeneous heritage collections based on authority files and knowledge bases and can greatly support the search and exploration of historical documents. Nowadays there are several approaches making profit of machine learning methods for NER (Li et al. 2020). For instance, Erdmann et al. (2016) presented a CRF-based model with handcrafted features for Latin historical texts and motivated the choice of Part-of-Speech (POS) tagger by the fact that this NLP tool leverages the highly informative morphological complexity of Latin. Hubkova at al. (2019) also proposed a BiLSTM-based model by applied a character-based CNN to encode the different spellings of words. Nonetheless, it must be noted also the difficulties for applying NER to historical documents and the consequent degradation of performance metrics (Rodriguez et al., 2012; van Strien et al., 2020). In such a context it is essential to count on annotated corpus and benchmarks. For instance, Hubkova et al. (2019) curated and annotated a corpus from scanned Czech historical newspapers. In the same line, Hamdi et al. (2019) delineated a German gold standard for NER within the domain of historical biodiversity literature.

3. Design of the platform

Figure 2 presents the architectural design of our proposed web platform. The architecture contains three main components: database, content management, and search/visualization.

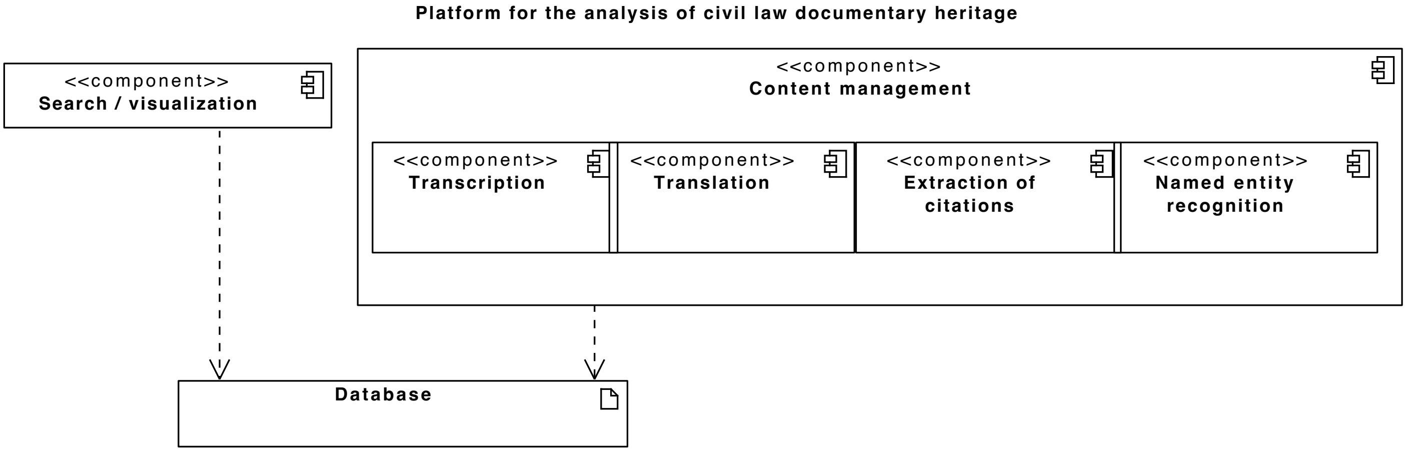
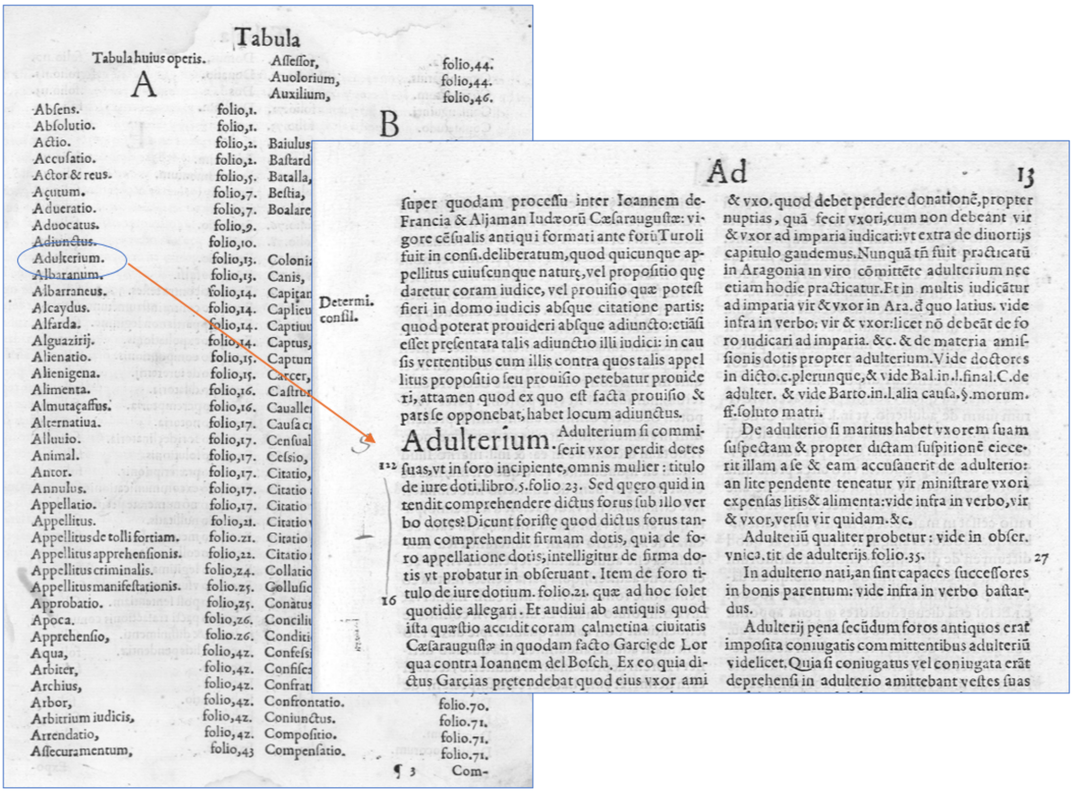


Figure 2. Architecture of our platform

Firstly, the *database* provides a semantic repository (RDF triplestore) for representing the structural contents of a civil law book. The type of books that we aim to analyse in this platform consist of a set of civil law concepts (voices) with its associated juridical interpretation. For instance, the left side of Figure 3 shows an example of the table of contents in the work *Repertorium Fororum et Observantiarum Regni Aragonum: una pluribus cum determinationibus consilii iustitiae Aragonum practicis atquae cautelis eisdem fideliter annexis* written by Miguel del Molino in an edition published in 1585. On the right side of the figure, the initial page containing the *Adulterium* concept is displayed. Considering this, Figure 4 shows the main classes and properties proposed for the RDF triplestore. Whenever possible, we have reused terms from well-known vocabularies such as the DCMI Metadata Terms (terms with *dct* prefix), the DCMI Type vocabulary (terms with *dctype* prefix), the FOAF vocabulary (terms with *foaf* prefix), the Simple Knowledge Organization System (SKOS, terms with *skos* prefix) and the Core Location Vocabulary (terms with *locn* prefix).

In addition, we have defined new classes and properties in our *Civil* vocabulary to define better the peculiarities of a book of civil law (*civil:Book*) as an extension of a printed text resource (*dctype:Text*). It must be noted also that we have defined a particular class for representing the civil law concepts (*civil:Concept*) discussed in these books. This civil law concept is an extension of the typical concept (*skos:Concept*) in a knowledge organisation system. Apart from a preferred label, it is also annotated with properties containing links to digitized page images (*dct:source*), the original transcribed text associated to the concept (*civil:transcription*), its translation into a modern language (*civil:translation*), and any type of reference (*dct:references*) to bibliographic resources, locations, persons or organizations.



Tabula

Tabula huius operis.

		Aſſeſſor,	folio,44.
		Auolorium,	folio,44.
A		Auxilium,	folio,46.
		B	
Abſens.	folio,1.		
Abſolutio.	folio,1.		
Actio.	folio,2.	Baiulus	
Accuſatio.	folio,2.	Baſtard	
Actor & reus.	folio,5.	Batalla,	
Açutum.	folio,7.	Beſtia,	
Adueratio.	folio,7.	Boalare	
Aduocatus.	folio,9.		
Adiunctus.	folio,10.		
Adulterium.	folio,13.	Coloni	
Albaranum.	folio,13.	Canis,	
Albarraneus.	folio,14.	Capiţan	
Alcaydus.	folio,14.	Caplieu	
Alfarda.	folio,14.	Captiuu	
Alguazirij.	folio,14.	Captus,	
Alienatio.	folio,15.	Captum	
Alienigena.	folio,15.	Carcer,	
Alimenta.	folio,16.	Caſtru	
Almutaçaffus.	folio,16.	Cauallei	
Alternatiua.	folio,17.	Cauſa c	
Alluuio.	folio,17.	Cenſual	
Animal.	folio,17.	Celſio,	
Antor.	folio,17.	Citatio,	
Annulus.	folio,17.	Citatio	
Appellatio.	folio,17.	Citatio	
Appellitus.	folio,21.	Citatio	
Appellitus de tolli fortiam.	folio.21.	Citatio	
Appellitus apprehenſionis.	folio,22.	Citatio	
Appellitus criminalis.	folio,24.	Collatio	
Appellitus manifeſtationis.	folio.25.	Gollufic	
Approbatio.	folio,25.	Conàtus	
Apoca.	folio,26.	Concili	
Apprehenſio,	folio.26.	Conditi	
Aqua,	folio,42.	Confeſsi	
Arbiter,	folio,42.	Confiſca	
Archius,	folio,42.	Confrat	
Arbor,	folio,42.	Confrontatio.	folio.70.
Arbitrium iudicis,	folio,42.	Coniunctus.	folio.71.
Arrendatio,	folio,42.	Compoſitio.	folio.71.
Aſſecuramentum,	folio,43	Compenſatio.	folio.71.

¶ 3 Com-

Ad 13

ſuper quodam proceſſu inter Ioannem de-
Francia & Aljaman Iudæorũ Cæſarauguſtæ: vi-
gore cẽſualis antiqui formati ante forũ Turoli
fuit in conſi.deliberatum, quod quicunque ap-
pellitus cuiuſcunque naturę, vel propoſitio quę
daretur coram iudice, vel prouiſio quæ poteſt
Determi. conſil. fieri in domo iudicis abſque citatione partis:
quod poterat prouideri abſque adiuncto: etiãſi
eſſet pręſentata talis adiunctio illi iudici: in cau
ſis vertentibus cum illis contra quos talis appel
litus propoſitio ſeu prouiſio petebatur prouide
ri, attamen quod ex quo eſt facta prouiſio &
pars ſe opponebat, habet locum adiunctus.
Adulterium. Adulterium ſi commiſerit vxor perdit dotes
122 ſuas, vt in foro incipiente, omnis mulier : titulo
de iure doti.libro.5.folio 23. Sed quęro quid in
tendit comprehendere dictus forus ſub illo ver
bo dotes? Dicunt foriſtę quod dictus forus tan-
tum comprehendit firmam dotis, quia de fo-
ro appellatione dotis, intelligitur de firma do-
tis vt probatur in obſeruant. Item de foro ti-
16 tulo de iure dotium. folio.21. quæ ad hoc ſolet
quotidie allegari. Et audiui ab antiquis quod
iſta quæſtio accidit coram çalmetina ciuitatis
Cæſarauguſtæ in quodam facto Garcię de Lor
qua contra Ioannem del Boſch. Ex eo quia di-
ctus Garcias prętendebat quod eius vxor ami
& vxo. quod debet perdere donationẽ, propter
nuptias, quã fecit vxori, cum non debeant vir
& vxor ad imparia iudicari: vt extra de diuortijs
capitulo gaudemus. Nunquã tñ fuit practicatũ
in Aragonia in viro cõmittẽte adulterium nęc
etiam hodie practicatur. Et in multis iudicãtur
ad imparia vir & vxor in Ara. đ quo latius. vide
infra in verbo: vir & vxor: licet nõ debeãt de fo
ro iudicari ad imparia. &c. & de materia amiſ-
ſionis dotis propter adulterium. Vide doctores
in dicto.c.plerunque, & vide Bal.in.l.final.C.de
adulter. & vide Barto.in.l.alia cauſa.§.morum.
ff.ſoluto matri.
De adulterio ſi maritus habet vxorem ſuam
ſuſpectam & propter dictam ſuſpitionẽ ciece-
rit illam a ſe & eam accuſauerit de adulterio:
an lite pendente teneatur vir miniſtrare vxori
expenſas litis & alimenta: vide infra in verbo, vir
& vxor, verſu vir quidam. &c.
Adulteriũ qualiter probetur : vide in obſer.
vnica.tit de adulterijs folio.35. 27
In adulterio nati, an ſint capaces ſucceſſores
in bonis parentum: vide infra in verbo baſtar-
dus.
Adulterij pęna ſecũdum foros antiquos erat
impoſita coniugatis committentibus adulteriũ
videlicet. Quia ſi coniugatus vel coniugata erãt
deprehenſi in adulterio amittebant veſtes ſuas

Figure 3. An example of pages from Miguel del Molino's book

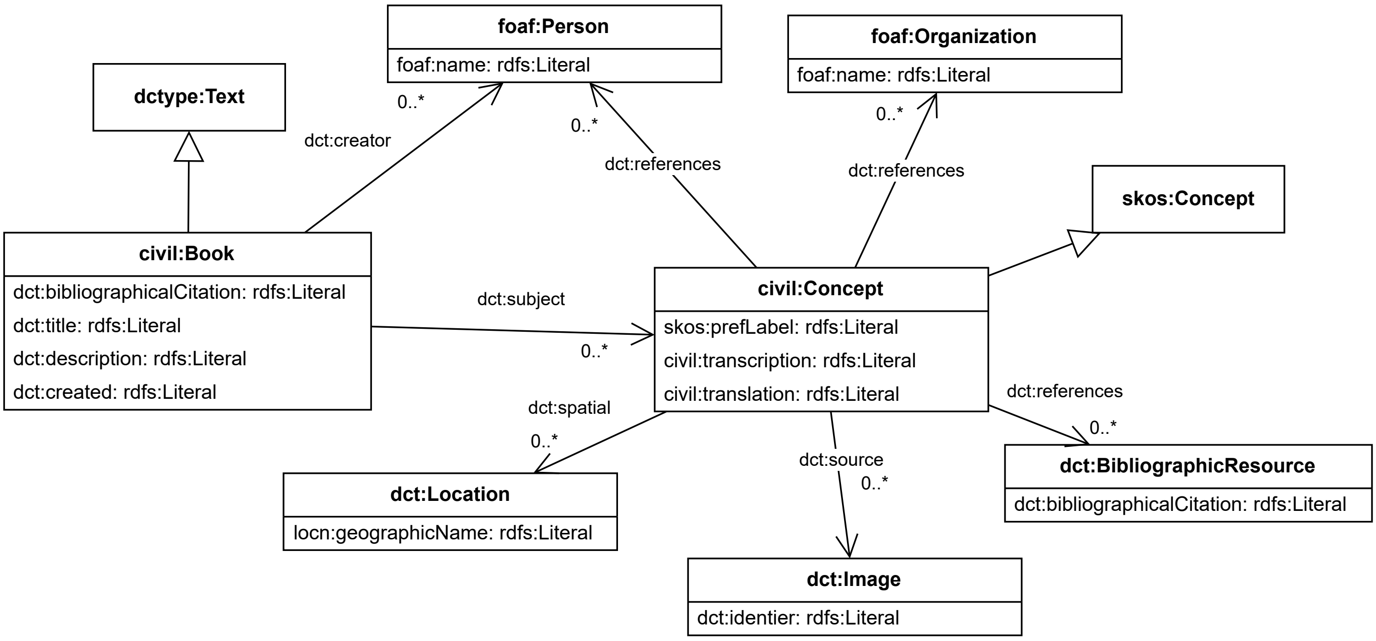

foaf:Person
foaf:name: rdfs:Literal
foaf:Organization
foaf:name: rdfs:Literal
dctype:Text
0..*
dct:creator
0..*
dct:references
0..*
dct:references
skos:Concept
civil:Book
dct:bibliographicalCitation: rdfs:Literal
dct:title: rdfs:Literal
dct:description: rdfs:Literal
dct:created: rdfs:Literal
dct:subject
0..*
civil:Concept
skos:prefLabel: rdfs:Literal
civil:transcription: rdfs:Literal
civil:translation: rdfs:Literal
dct:spatial
0..*
dct:references
0..*
dct:source
0..*
dct:Location
locn:geographicName: rdfs:Literal
dct:BibliographicResource
dct:bibliographicalCitation: rdfs:Literal
dct:Image
dct:identier: rdfs:Literal

Figure 4. Conceptual model (UML class diagram) of the semantic repository

Secondly, the *content management* component cordinates a set of automated tasks to facilitate the processing and analysis of documents. Figure 5 shows an activity diagram representing the processing workflow applied to a new book ingested in the platform and the inputs/outputs that are generated.

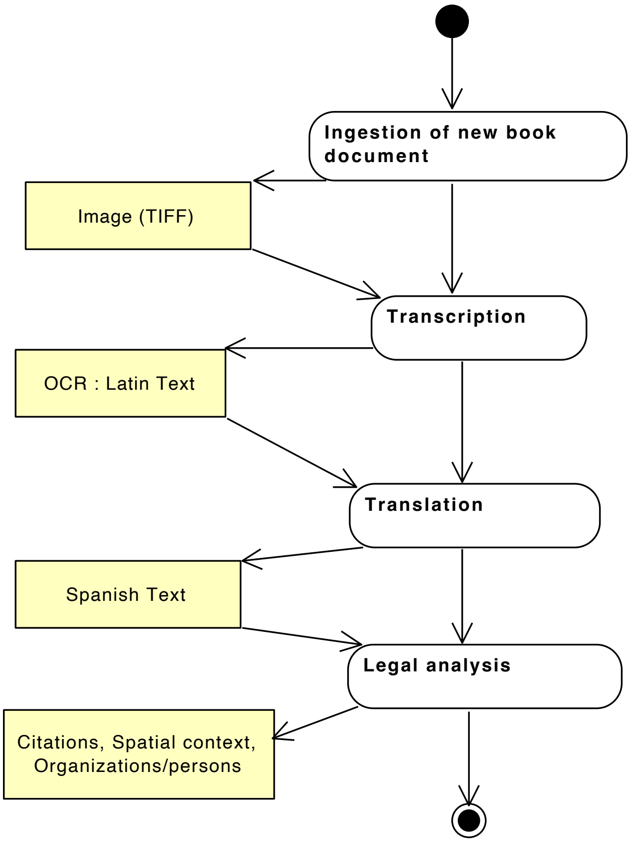
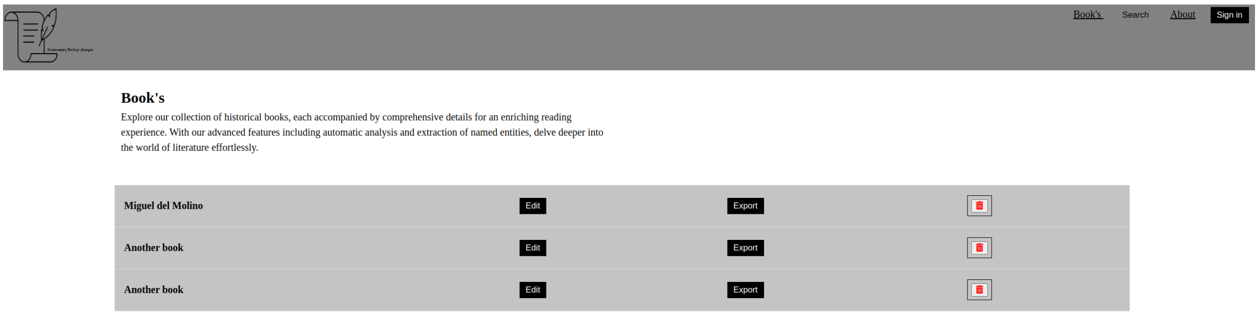

Ingestion of new book document
Image (TIFF)
Transcription
OCR : Latin Text
Translation
Spanish Text
Legal analysis
Citations, Spatial context, Organizations/persons

Figure 5. Workflow for the processing of civil law documents

On the one hand, the transcription component aims to convert the digitized images of book pages with printed text into machine-readable textual format. For this step, we have used Pero-OCR (Kišš et al. , 2023), a state-of-the-art optical character recognition (OCR) tool. The transcription process, apart from obtaining the OCR of every page in the book, also encompasses the division of text for the different civil law concepts. On the other hand, the translation component converts the original latin text into a modern language (e.g. Spanish). As a first approach for this translation task we have used the Google Translator API to translate the Latin text linked to each civil law concept into Spanish. Moreover, we have also components for citation extraction and named entity recognition. They contribute to the semantic understanding and analysis of document content, laying the groundwork for advanced information retrieval and manipulation. For the identification of citations, we have implemented the detection of some string patterns followed by the external citations. In the case of named entity recognition, we have integrated the use of *spacy*, an open-source Python library for Natural Language Processing the includes pre-trained models for the identification of locations and authorities in Spanish texts.

Last, the *search/visualization* component facilitates the interaction between the users and our platform. There are two types of users: normal and expert users.

Normal users are researchers or general public interested in civil law. For these users the platform offers a direct access to a homepage displaying the list of books (Figure 6) and enabling the export of book contents in RDF format. In addition, the homepage allows the filtered search on books and specific contents of a book through facets for Concepts, Locations, Persons, and Organizations where the user can either type text or select a value from a list. The search results are presented in a web page which displays a list of all concepts that meet the search criteria specified on the search page. For example, Figure 7 presents a search result after filtering the concept name *Albarraneus*.

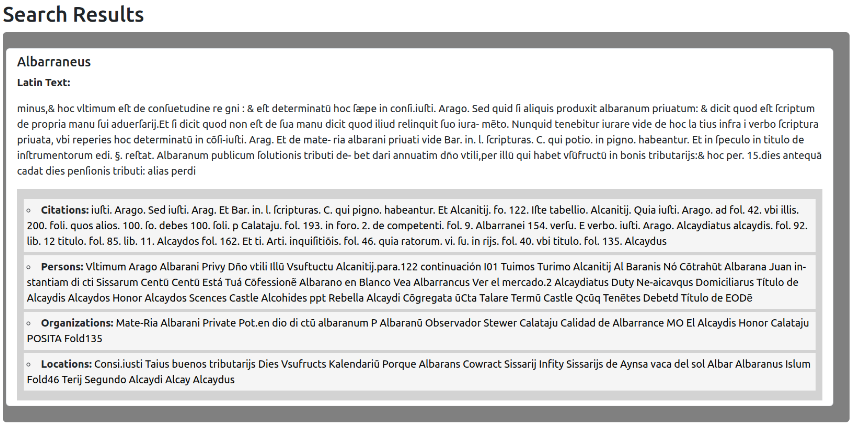


Book's

Explore our collection of historical books, each accompanied by comprehensive details for an enriching reading experience. With our advanced features including automatic analysis and extraction of named entities, delve deeper into the world of literature effortlessly.

Miguel del Molino	Edit	Export	🗑
Another book	Edit	Export	🗑
Another book	Edit	Export	🗑

Figure 6. Homepage displaying the list of books



Search Results

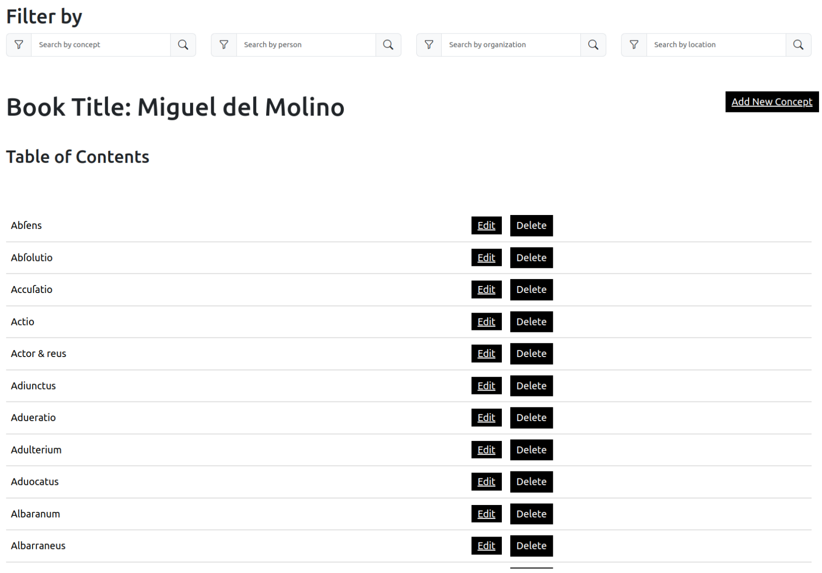
Albarraneus

Latin Text:

minus,& hoc vltimum eſt de conſuetudine re gni : & eſt determinatū hoc ſæpe in conſi.iuſti. Arago. Sed quid ſi aliquis produxit albaranum priuatum: & dicit quod eſt ſcriptum de propria manu ſui aduerſarij.Et ſi dicit quod non eſt de ſua manu dicit quod iliud relinquit ſuo iura- mẽto. Nunquid tenebitur iurare vide de hoc la tius infra i verbo ſcriptura priuata, vbi reperies hoc determinatū in cōſi-iuſti. Arag. Et de mate- ria albarani priuati vide Bar. in. l. ſcripturas. C. qui potio. in pigno. habeantur. Et in ſpeculo in titulo de inſtrumentorum edi. §. reſtat. Albaranum publicum ſolutionis tributi de- bet dari annuatim dn̄o vtili,per illū qui habet vſūfructū in bonis tributarijs:& hoc per. 15.dies antequā cadat dies penſionis tributi: alias perdi

- **Citations:** iuſti. Arago. Sed iuſti. Arag. Et Bar. in. l. ſcripturas. C. qui pigno. habeantur. Et Alcanitij. ſo. 122. Iſte tabellio. Alcanitij. Quia iuſti. Arago. ad ſol. 42. vbi illis. 200. ſoli. quos alios. 100. ſo. debes 100. ſoli. p Calataju. ſol. 193. in ſoro. 2. de competenti. ſol. 9. Albarranei 154. verſu. E verbo. iuſti. Arago. Alcaydiatus alcaydis. ſol. 92. lib. 12 titulo. ſol. 85. lib. 11. Alcaydos ſol. 162. Et ti. Arti. inquiſitiōis. ſol. 46. quia ratorum. vi. ſu. in rijs. ſol. 40. vbi titulo. ſol. 135. Alcaydus
- **Persons:** Vltimum Arago Albarani Privy Dn̄o vtili Illū Vsuftuctu Alcanitij.para.122 continuación I01 Tuimos Turimo Alcanitij Al Baranis Nó Cōtrahūt Albarana Juan instantiam di cti Sissarum Centū Centū Está Tuá Cōfessionẽ Albarano en Blanco Vea Albarrancus Ver el mercado.2 Alcaydiatus Duty Ne-aicavqus Domiciliarus Título de Alcaydis Alcaydos Honor Alcaydos Scences Castle Alcohides ppt Rebella Alcaydi Cōgregata ūCta Talare Termū Castle Qcūq Tenẽtes Debetd Título de EODẽ
- **Organizations:** Mate-Ria Albarani Private Pot.en dio di ctū albaranum P Albaranū Observador Stewer Calataju Calidad de Albarrance MO El Alcaydis Honor Calataju POSITA Fold135
- **Locations:** Consi.iusti Taius buenos tributarijs Dies Vsufructs Kalendariū Porque Albarans Cowract Sissarij Infity Sissarijs de Aynsa vaca del sol Albar Albaranus Islum Fold46 Terij Segundo Alcaydi Alcay Alcaydus

Figure 7. Search Results web page



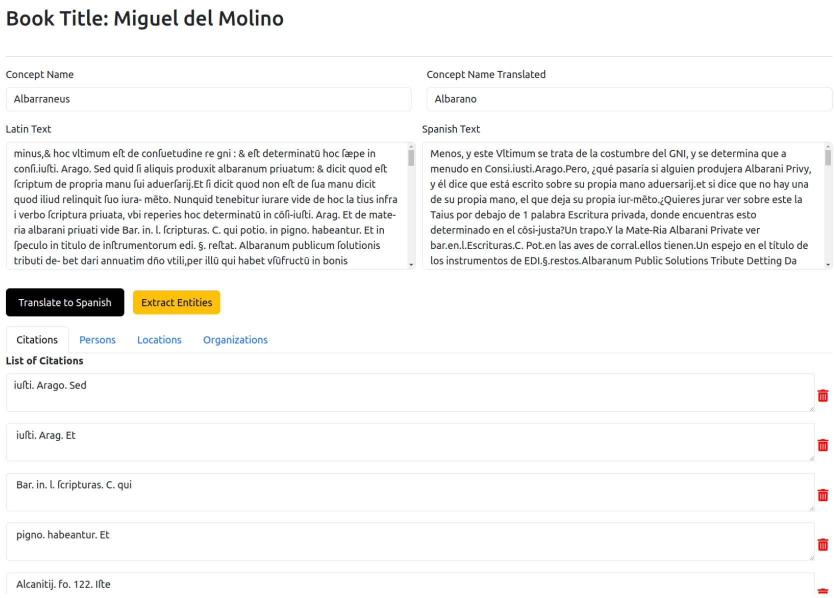
Filter by

Book Title: Miguel del Molino

Table of Contents

Abſens	Edit	Delete
Abſolutio	Edit	Delete
Accuſatio	Edit	Delete
Actio	Edit	Delete
Actor & reus	Edit	Delete
Adiunctus	Edit	Delete
Adueratio	Edit	Delete
Adulterium	Edit	Delete
Aduocatus	Edit	Delete
Albaranum	Edit	Delete
Albarraneus	Edit	Delete

Figure 8. Management of the list of concepts in a book



Book Title: Miguel del Molino

Concept Name: Albarraneus

Concept Name Translated: Albarano

Latin Text:

minus,& hoc vltimum eſt de conſuetudine re gni : & eſt determinatū hoc ſæpe in conſi.iuſti. Arago. Sed quid ſi aliquis produxit albaranum priuatum: & dicit quod eſt ſcriptum de propria manu ſui aduerſarij.Et ſi dicit quod non eſt de ſua manu dicit quod iliud relinquit ſuo iura- mẽto. Nunquid tenebitur iurare vide de hoc la tius infra i verbo ſcriptura priuata, vbi reperies hoc determinatū in cōſi-iuſti. Arag. Et de mate- ria albarani priuati vide Bar. in. l. ſcripturas. C. qui potio. in pigno. habeantur. Et in ſpeculo in titulo de inſtrumentorum edi. §. reſtat. Albaranum publicum ſolutionis tributi de- bet dari annuatim dn̄o vtili,per illū qui habet vſūfructū in bonis

Spanish Text:

Menos, y este Vltimum se trata de la costumbre del GNI, y se determina que a menudo en Consi.iusti.Arago.Pero, ¿qué pasaría si alguien produjera Albarani Privy, y él dice que está escrito sobre su propia mano adversarij.et si dice que no hay una de su propia mano, el que deja su propia iur-mẽto.¿Quieres jurar ver sobre este la Taius por debajo de 1 palabra Escritura privada, donde encuentras esto determinado en el cōsi-justa?Un trapo.Y la Mate-Ria Albarani Private ver bar.en.l.Escrituras.C. Pot.en las aves de corral.ellos tienen.Un espejo en el título de los instrumentos de EDI.§.restos.Albaranum Public Solutions Tribute Detting Da

List of Citations

- iuſti. Arago. Sed
- iuſti. Arag. Et
- Bar. in. l. ſcripturas. C. qui
- pigno. habeantur. Et
- Alcanitij. ſo. 122. Iſte

Figure 9. Edit concept web page

The experts are the authorized users that can add, modify or delete the books and their associated concepts (Figure 8, in previous page). In addition, they have access to the advanced functionalities for automatic transcription, translation, and information extraction in the web page for editing concepts (Figure 9, in previous page).

4. Experimental results

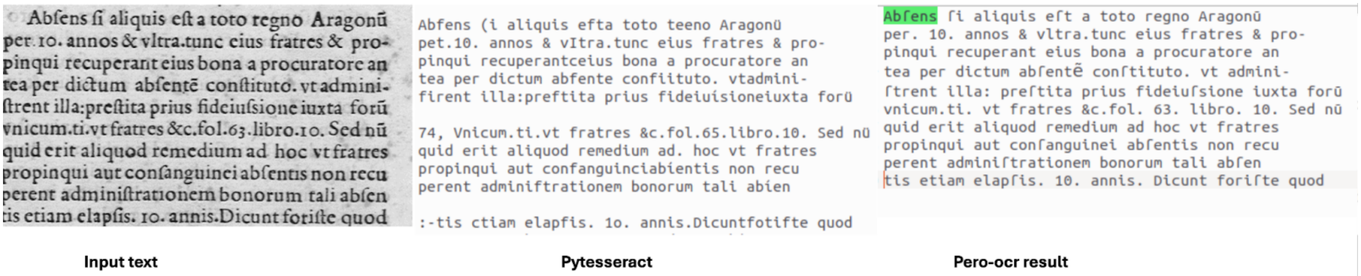
The proposed platform for the analysis of documentary heritage, whose design was presented in Section 3, has been implemented using the Django framework for web development. In addition, the components for content management have been implemented in Python integrating different libraries for OCR (*Pero-OCR*), and named entity recognition (*spacy*).

Moreover the feasibility of the implementation has been tested with the processing of a doctrinal civil law book written by Miguel del Molino and printed in 1585: *Repertorium Fororum et Observantiarum Regni Aragonum: una pluribus cum determinationibus consilii iustitiae Aragonum practicis atquae cautelis eisdem fideliter annexis*, available at

> *https://derechoaragones.aragon.es/es/consulta/registro.do?id=600036*

In particular, we were interested in the performance of the automated assitants for transcription, translation and information extraction.

With respect to the transcription, we compared the results of Pero-OCR with the results obtained after applying Tesseract, a widely used open-source OCR tool (https://github.com/tesseract-ocr/tesseract). The experience showed that Pero-OCR is more efficient. For instance, Figure 10 depicts the transcription of a sample text with both Tesseract and Pero-OCR. There are letters that were not recognized by Tesseract , such as the letter "s" (*ſ* contained in *abſens*), which in the Latin language is very similar to "f". In contrast, it was correctly detected by Pero-OCR.



Abſens ſi aliquis eſt a toto regno Aragonũ
per.10. annos & vltra.tunc eius fratres & pro-
pinqui recuperant eius bona a procuratore an
tea per dictum abſentē conſtituto. vt admini-
ſtrent illa:preſtita prius fideiuſsione iuxta forũ
vnicum.ti.vt fratres &c.fol.63.libro.10. Sed nũ
quid erit aliquod remedium ad hoc vt fratres
propinqui aut conſanguinei abſentis non recu
perent adminiſtrationem bonorum tali abſen
tis etiam elapſis. 10. annis.Dicunt foriſte quod

Input text

```
Abfens (i aliquis efta toto teeno Aragonū
pet.10. annos & vItra.tunc eius fratres & pro-
pinqui recuperantceius bona a procuratore an
tea per dictum abfente confiituto. vtadmini-
firent illa:preftita prius fideiuísioneiuxta forū

74, Vnicum.ti.vt fratres &c.fol.65.libro.10. Sed nū
quid erit aliquod remedium ad. hoc vt fratres
propinqui aut confanguinciabientis non recu
perent adminiftrationem bonorum tali abien

:-tis ctiam elapfis. 1o. annis.Dicuntfotifte quod
```

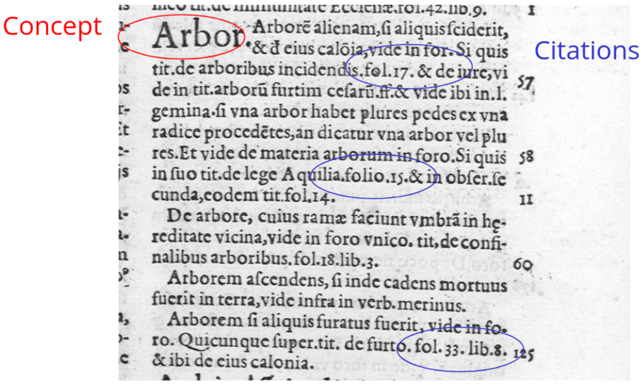
Pytesseract

```
Abſens ſi aliquis eſt a toto regno Aragonū
per. 10. annos & vltra.tunc eius fratres & pro-
pinqui recuperant eius bona a procuratore an
tea per dictum abſentẽ conſtituto. vt admini-
ſtrent illa: preſtita prius fideiuſsione iuxta forū
vnicum.ti. vt fratres &c.fol. 63. libro. 10. Sed nū
quid erit aliquod remedium ad hoc vt fratres
propinqui aut conſanguinei abſentis non recu
perent adminiſtrationem bonorum tali abſen
tis etiam elapſis. 10. annis. Dicunt foriſte quod
```

Pero-ocr result

Figure 10. Comparison between Tesseract and Pero-OCR to generate a text

With respect to the translation, we compared qualitatively the results obtained after invoking the API of several online translators: Google Translator, Yandex, DeepL, Translateking, imTranslator and Translateking. After comparing the results obtained with some sample texts directly generated by Pero-OCR, we considered that the Google Translator API was providing the best results.


Concept
Citations

meo tit.de immunitate Ecclesiæ.fol.42.lib.9. 1
Arbor. Arborē alienam,ſi aliquis ſciderit,
&đ eius calõia,vide in for. Si quis
tit.de arboribus incidendis.fol.17. & de iure,vi 57
de in tit.arborũ furtim ceſarũ.ff.& vide ibi in.l.
gemina.ſi vna arbor habet plures pedes ex vna
radice procedētes,an dicatur vna arbor vel plu
res.Et vide de materia arborum in foro.Si quis 58
in ſuo tit.de lege Aquilia.folio.15.& in obſer.ſe
cunda,eodem tit.fol.14. 11
De arbore, cuius ramæ faciunt vmbrã in hę-
reditate vicina,vide in foro vnico. tit,de confi-
nalibus arboribus.fol.18.lib.3. 60
Arborem aſcendens,ſi inde cadens mortuus
fuerit in terra,vide infra in verb.merinus.
Arborem ſi aliquis furatus fuerit, vide in fo-
ro. Quicunque ſuper.tit. de furto. fol. 33. lib.8. 125
& ibi de eius calonia.

Figure 11. Example of concept and citations

Regarding the extraction of citations, we concluded that this processing task must be customized to the special features of each book. Figure 11 shows part of the text associated to the concept "Arbor" and we highlighted some external citations to *"Fori Regni Aragonum"* (https://derechoaragones.aragon.es/es/consulta/registro.do?id=600013), a collective work printed in 1496 that compiles the official regulation books about civil law by the end of the XV century. Even in this small example we can identify some patterns to cite a book in this collective work. For instance, from the text:

> […] vide in foro. Quicunque super.tit. de furto.fol.33.lib.8.

we can derive the following citation pattern:

> […] vide in foro. <beginning words of a paragraph within a title>.tit. de <title name>.fol.<page number>.lib.<book number>.

Figure 12 shows an excerpt of the cited page of the external book.

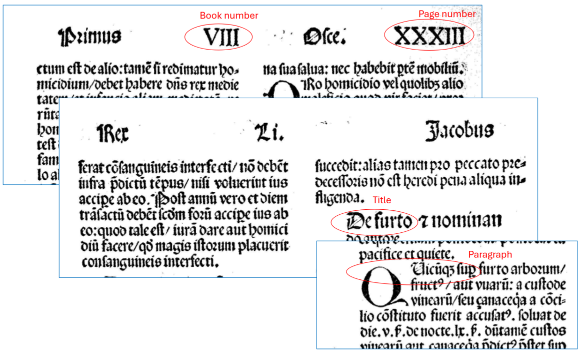

Book number
Page number
Primus
VIII
Osee.
XXXIII
Rex
Li.
Jacobus
Title
De furto ꝛ nominan
Paragraph

Figure 12. Example of cited title and paragraph in page 33 of book 8.

About the direct applicability of the *spacy* library for named entity recognition (NER), the quality of results is not very satisfactory because the translation into Spanish contains frequent mistakes, derived in turn by some errors found in the Latin transcription. Although the results of NER could be improved once a better translation into Spanish is provided, we believe that the incorporation of knowledge bases such gazetteers of historical place names or authority files would help to identify better these named entities.

5. Conclusions

This paper has presented a first prototype for the development of a web platform to facilitate the analysis of civil law documentary heritage. The design of the platform has been customized to the specific needs of the experts in this domain. First, we have designed a conceptual framework to represent the information required for a deep analysis of civil law contents. In addition, we have integrated open-source tools to assist in tasks such as transcription, translation, and information extraction, which usually require high human resources if performed completely manually.

Although the results obtained by automated tasks could be clearly improved, they provide an appropriate input in each step that can be revised by experts to ameliorate the performance. For instance, if the Latin transcription of a text is revised by an expert in the platform, the automated translation will provide better results.

As future work, we will continue with the development of the platform and the testing of more alternatives for assisting in the automated tasks of transcription, translation, and information extraction. This further development will be accompanied with a detailed analysis of the results obtained with the ingestion of a more extensive corpus of civil law books in the platform. In addition, a version manager to control variants among different manuscripts, editions and transcriptions is being considered. Finally, a system for collaborative annotation and linking of the documents to support advanced academic research is also envisaged (García-Marco, 2020).

Acknowledgements

This paper is part of the R&D projects T59_23R and S15_23R supported by the Aragon Regional Government in cooperation with the “Cátedra de Derecho civil y foral de Aragón”.

References

Aljalbout, S.; Falquet, G. (2017). Un modèle pour la représentation des connaissances temporelles dans les documents historiques: Applications sur les manuscrits de F. Saussure. // Proc. 28es Journées francophones d'Ingénierie des Connaissances (IC 2017): Caen, France, July 2017.

Erdmann, A.; Brown, C.; Joseph, B.D; Janse, M.; Ajaka, P.; Elsner, M.; de Marneffe, M. (2016). Challenges and solutions for Latin named entity recognition. // COLING 2016: 26th International Conference on Computational Linguistics. Association for Computational Linguistics. 85–93.

Fischer, L.; Scheurer, P.; Schwitter, R.; Volk, M. (2022). Machine translation of 16th century letters from Latin to German. // Proceedings of 2nd Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) at LREC-2022, Marseille.

García Marco, Francisco Javier. Knowledge Organization in Historical Information Systems Revisited: Changes in Society, Technology and Expectations 25 Years Later. // Knowledge Organization at the Interface. Proceedings of the Sixteenth International ISKO Conference 6-8 July 2020 Aalborg, Denkmark. Würzburg: Ergon-Verlag GmbH, 2020. 474-478.

Gupta, A.; Gutierrez-Osuna, R.; Christy, M.; Capitanu, B.; Auvil, L.; Grumbach, L.; Furuta, R.; Mandell, L. (2015). Automatic Assessment of OCR Quality in Historical Documents. // Proc. of 29th AAAI Conference on Artificial Intelligence. 1735-1741.

Hamdi, A.; Jean-Caurant, A.; Sidere, N.; Coustaty, M.; Doucet, A. (2019). An analysis of the performance of named entity recognition over ocred documents. // 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE. 333–334.

Hubkova, H. (2019). Named-entity recognition in Czech historical texts: Using a CNN-BiLSTM neural network model. Ph.D. thesis.

Kišš, M.; Hradiš, M.; Beneš, K.; Buchal, P.; Kula, M. (2023). SoftCTC: semi-supervised learning for text recognition using soft pseudo-labels. // International Journal on Document Analysis and Recognition (IJDAR). 2, 1-17.

Kodym, O.; Hradiš, M. (2021). Page layout analysis system for unconstrained historic documents. // Proc. of 16th International Conference on Document Analysis and Recognition–ICDAR 2021: Lausanne, Switzerland, September 5–10, 2021. Part II, 492-506).

Lacasta, J.; Nogueras-Iso, J.; Zarazaga-Soria, F.J.; Pedraza-Gracia, M.J. (2022). Tracing the origins of incunabula through the automatic identification of fonts in digitised documents. // Multimedia Tools and Applications. 81:28, 40977-40991.

Li, J.; Sun, A.; Han, J.; Li, C. (2020). A survey on deep learning for named entity recognition. // IEEE Transactions on Knowledge and Data Engineering. 34:1, 50-70.

Martínez Garcia, E; García Tejedor, Á. (2020). Latin-Spanish Neural Machine Translation: From the Bible to Saint Augustine. // Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages: Marseille, France. European Language Resources Association (ELRA). 94–99.

Neji, H.; Ben Halima, M.; Nogueras-Iso, J.; Hamdani, T.M.; Lacasta, J.; Chabchoub, H.; Alimi, A.M. (2024). Doc-Attentive-GAN: attentive GAN for historical document denoising. // Multimedia Tools and Applications. 83, 55509–55525.

Rodriquez, K.J.; Bryant, M.; Blanke, T.; Luszczynska, M. (2012). Comparison of named entity recognition tools for raw OCR text. // 11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, Vienna, Austria, September 19-21, 2012. Scientific series of the OGAI. 5, 410–414.

Tiedemann, J. (2012). Parallel data, tools and interfaces in Opus. // Calzolari, Nicoletta (Conference Chair); et al., (eds). Proc. of 8th Int. Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey. European Language Resources Association (ELRA).

van Strien, D.; Beelen, K.; Coll Ardanuy, M.; Hosseini, K.; McGillivray, B.; Colavizza, G. (2020). Assessing the impact of OCR quality on downstream NLP tasks. // Proceedings of the 12th International Conference on Agents and Artificial Intelligence. 1, 484-496. https://doi.org/10.5220/0009169004840496

Enviado: 2024-06-07. Aceptado: 2024-06-11.