Umanistica Digitale – ISSN: 2532-8816 – n. 22, 2026: Special Issue
A. Esuli, V.R. Imperia, G. Puccetti – *Automatic Annotation of Legal References (Allegationes) in the Liber Extra's Ordinary Gloss*
DOI: http://doi.org/10.60923/issn.2532-8816/22163

Automatic Annotation of Legal References (Allegationes) in the Liber Extra's Ordinary Gloss

Andrea Esuli

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" – Pisa, Italy
andrea.esuli@isti.cnr.it

Vincenzo Roberto Imperia

Università degli Studi di Palermo - Dipartimento di Giurisprudenza – Palermo, Italy
vincenzoroberto.imperia@unipa.it

Giovanni Puccetti

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" – Pisa, Italy
giovanni.puccetti@isti.cnr.it

Abstract

The study of normative corpora of the past is a key activity in the fields of Religious Studies and Legal History. The development of intelligent software tools that support this activity is of paramount importance to support the digital transformation of the community. We present an interdisciplinary activity that leads to an accurate automatic annotation of legal references in the *Liber Extra*'s Ordinary Gloss. An index of legal references has been derived from the annotations enabling the creation of novel navigation and data analysis tools. The contribution of this work is twofold: the actual index is already by itself a valuable resource for the discipline, and we detail the process that leads to its production, showing that an effective result can be delivered by a small team with limited resources. Both the index and the code are made publicly available.

Keywords: Legal references, Allegationes, Information Extraction, Conditional Random Fields, Dataset, IRCDL2025

Lo studio dei corpora normativi del passato costituisce un'attività centrale negli ambiti degli studi religiosi e della storia del diritto. Lo sviluppo di strumenti software in grado di supportare tale

attività riveste un'importanza fondamentale ai fini della trasformazione digitale della comunità scientifica. Il contributo presenta un'attività interdisciplinare finalizzata all'annotazione automatica accurata dei riferimenti giuridici presenti nella Glossa ordinaria al Liber Extra. A partire da tali annotazioni è stato elaborato un indice dei riferimenti giuridici (allegationes), che permette lo sviluppo di nuovi strumenti di navigazione e di analisi dei dati. Il lavoro offre un duplice contributo: da un lato, l'indice costituisce di per sé una risorsa di significativo valore per la disciplina; dall'altro, viene illustrato nel dettaglio il processo che ne ha consentito la realizzazione, dimostrando come sia possibile ottenere risultati efficaci anche attraverso il lavoro di un gruppo ristretto con risorse limitate. Sia l'indice sia il codice sono resi pubblicamente disponibili.

Parole chiave: Riferimenti Giuridici, Allegationes, Estrazione di Informazioni, Conditional Random Fields, Dati, IRCDL2025

1. Introduction

ITSERR (2024) (Italian Strengthening of the ESFRI RI RESILIENCE) is an interdisciplinary and distributed Research Infrastructure for Religious Studies. The development of innovative AI-based tools that support the digital transformation of the community is one of the many goals of the project. In this context we present a contribution to the "GNORM" software, which aims to provide users with a set of tools and functionalities to facilitate research on large normative corpora of the past. One case of study is to enable the consultation of the *Corpus Iuris Canonici* [23] text and its *Glossa Ordinaria* (Ordinary gloss) [5], enabling users to access the network embodied by the visual model in which legal texts (manuscripts and printed books) were structured, as described in Section 1.1.

The contributions toward this goal we present in this paper are:

- A reproducible pipeline for the automatic annotation of legal references (allegationes) in the *Liber Extra*'s *Glossa Ordinaria*, combining expert annotation with supervised machine learning.
- A complete, openly available index of legal references linking glosses to specific norms across the *Corpus Iuris Canonici* and *Corpus Iuris Civilis*.
- Demonstrating that high-accuracy results can be obtained with limited computational and human resources, making the approach suitable for small research teams in the humanities.

The remainder of this section provides the historical and conceptual background necessary to understand the role of glosses and allegationes in medieval legal culture.

1.1. Glossa and Allegationes, from Medieval Books to Computer Science

Before examining the specific case study presented here and evaluating the concrete applications in the field of legal history resulting from the development of automatic annotation techniques for legal allegations, it is necessary to clarify the meaning of terms such as gloss, ordinary gloss, and allegations, and in particular to consider the nature of allegations in the intellectual context

of medieval jurists. The term 'glossa' (gloss) refers to "a brief annotation composed and written to explain a text and addressing either its terminology and its exterior trappings or its animating spirit and its underlying principles" [4]. From the late 11th century and for several centuries thereafter, the gloss served as the principal paratextual tool through which law masters in law schools, later Universities, explained and commented on Roman and Canon law compilations, enabling students and practitioners to engage with technically complex texts.


303 Decretalium Gregorij Lib. I. 304

CAP. I.

CAP. II.

CAP. III.

CAP. IIII.

Figure 1: A page from the *Liber Extra*'s *Ordinary Gloss* (*Decretales D. Gregorii Papae IX* 1582). The box of text in the upper center of the page is the normative text from *Liber Extra*. The surrounding text are the glosses that also contain the legal references.

The term 'Glossa Ordinaria' (ordinary gloss) refers to the apparatus of glosses compiled by some eminent jurists, renowned for their exceptional quality and thoroughness. Their widespread

acceptance ensured their consistent reproduction alongside the normative reference text, establishing them as the ordinary apparatus [10]. The phenomenon of normative texts accompanied by an ordinary gloss is evident both in the manuscript production of the late Middle Ages, in the earliest printed works of the late 15th century, and continuing into the significant editorial initiatives of the 16th century.

Manuscripts and printed texts retained the standard page layout essentially unchanged, with the normative text and the ordinary gloss arranged according to a model that has been defined, in its essential features, as the "agora template" [14], as shown in Figure 1. This spatial configuration, with the normative text at the center framed by the glosses, mirrored the dialogical nature of the legal interpretation method: the authoritative text was accompanied by commentary that explained, questioned, or contradicted it. This visual arrangement symbolized a lively and interrelated debate, where text and commentary coexisted in a symbiotic relationship. Based on this, it can be argued that a number of important considerations regarding Accursius' apparatus to the *Corpus iuris civilis* also apply to any other apparatus of glosses. As noted in Speciale [28]: "alla radice del successo dell'apparato accursiano è la struttura ipertestuale dell'opera: ipertesto *ante litteram*, l'*apparatus* si caratterizza per la struttura frammentaria e non lineare e per l'inscindibile collegamento, anche sul piano della contiguità fisica nei manoscritti e nelle stampe, tra glossae e testo normativo"[1].

To clarify the meaning of this comparison, it can be affirmed that it is the very content of the apparatus—whether ordinary or otherwise—that qualifies them as "hypertexts" [28], since they demand the active participation of the reader. The margins of the normative text contain various components, each serving a specific function, but collectively acting as tools to help the user understand the main text, making it more accessible and consultable. Particular attention will be paid here to one of these components, the *allegationes*, defined by Hermann Kantorowicz [15] in the last century as references to *auctoritates* that must justify every assertion or, at least, deal with those that at first glance appear to contradict it. These references are, in the vast majority of cases, citations of legal norms, presented in a highly abbreviated form, using abbreviations and numbers that correspond to the conventional criteria for identifying the legal source containing the norm. These references, initially sporadic in the earliest layers of glosses, gradually increased in number until they became quantitatively predominant, alongside the refinement of interpretive techniques developed by jurists. Before delving into the analysis of the specific case study presented here, and before evaluating the practical applications derived from the development of automatic annotation techniques for legal allegations—and testing their utility for legal history—it is essential to emphasize the intrinsic value of the nature of allegations.

Recently, what Kantorowicz had already emphasized in his contribution [15], albeit without much elaboration, has been authoritatively reaffirmed and extensively clarified: allegations are not merely citations.

Many recent works ([24];[19];[20];[21]) have authoritatively reaffirmed and extensively clarified that allegations are not merely citations: the use of a reference constitutes an appeal to an

[1] "at the root of the success of Accursian apparatus is the hypertextual structure of the work:a hypertext ante litteram, the apparatus is characterised by its fragmentary and non-linear structure and by the inseparable connection, also on the level of physical contiguity in manuscripts and prints, between glossae and normative text"

authority that not only lends greater solidity to the legal argument, but forms its very foundation. For this reason, when interpreting and commenting on a normative text, it is not uncommon for medieval jurists to refer to *auctoritates* that the modern reader might consider irrelevant to the legal discourse, including entire quotations or simple references to passages from religious, literary, poetic, or philosophical works.

Of course, this type of allegation was much less common than references to strictly legal sources.

The use of legal references by medieval jurists had thus a complex significance. It demonstrated that interpretative operations were carried out on legitimate grounds, based, on the one hand, on applicable and valid legal norms and, on the other, on the consistent use of legal categories shared within a common cultural and intellectual context [19]. Therefore, the primary aim of the jurists was to establish general rules capable of binding together various legal texts which had the force of law but, due to the peculiar genesis of the collections or compilations containing them, could conflict with other texts—either within the same collection or externally, but endowed with the same authority, within or outside the same collection [7].

These conceptual premises are closely related and provide the backdrop for considerations concerning more practical aspects. The first pertains to the citation style of allegationes, which presupposed a normative text that was now stable, fixed, and unalterable. This allowed the reader to pinpoint the referenced passage with certainty [21]. The second concerns the nature and purpose of these references. The primary aim was to create a network in which the various sedes materiae were organized and linked, thus enabling the reader-user to navigate vast normative compilations [28]. Moreover, the use of legal references could take on an even more incisive argumentative function if, in addition to a principle or rule derived from the normative text, allegations *pro* and *contra* were included. This technique quickly evolved into a distinct literary genre, with many works constructed according to these specific criteria [33].

The formal and substantive nature of the allegationes in the works of the jurists underscores the need to exploit the potentials of computer science, and specifically machine learning, to develop accurate tools capable of preserving their peculiarities in future digital editions of medieval legal texts [21]. This requirement is particularly relevant to the purposes of the GNORM project. One of the objectives is to enable the consultation of the *Corpus Iuris Canonici* text and its ordinary gloss.

The interface, currently in the design and development phase, will allow users to access the network embodied by the visual model in which legal texts (manuscripts and printed books) were structured, as discussed earlier. To facilitate navigation, the various text components will be organized into levels, ideally mirroring the interpretative process followed by medieval jurists. The first level will contain the legal text, structured according to the internal organization of each *Corpus Iuris Canonici*'s component. The second level will include the text of the *Glossa Ordinaria*. The third level will provide an explanation of the legal allegations contained in the ordinary gloss following modern citation standards, along with further detail, and, finally, the text of the norm to which the allegation refers. In this way, the dynamic nature of the reader–text interaction characterizing apparatus as hypertexts will be preserved in a structured and accessible manner. The design and implementation of an automatic annotation system for *allegationes*, understood narrowly as legal references, is therefore an essential step in advancing the concept behind GNORM.

2. Annotation of Legal References

The complex and stratified genesis of the ordinary apparatus to normative compilations, closely linked to the concept of the authoritative text itself in the Middle Ages, prevented the preparation of critical editions of them in the modern sense. Consultation of them still requires recourse to manuscripts or early modern printed editions [19]. This is also the case for the *Corpus iuris canonici*. For this reason, the *Glossa Ordinaria* (Ordinary Gloss) to the *Decretales Gregorii IX*, best known as the *Liber Extra*, was chosen for the design and development of an automatic annotation system for the legal allegations.

Promulgated by the Pope with the bull *Rex Pacificus*, the legal collection is subdivided into 5 books, 185 titles, 1971 chapters, with a total of 9872 *lemmas*. The gloss of each lemma contains a variable number of legal references. From the text of the 1582 *Editio Romana* [5], which is the common reference edition, a complete digital transposition was made by Edward A. Reno III, as part of the project "The Digital Decretals" [26]. The files containing the digital text can be found on the project's website, together with specific references to the sections of the apparatus not included in the transcription work, as well as the interventions made to standardize elements of spelling, abbreviation, punctuation and numbering in the text. Figure 2 illustrates the process that led to an accurate automatic annotation of the whole *Liber Extra*'s Ordinary Gloss based on the annotation of a small subset by a human expert, followed by the creation of an index of all the annotations in which every legal reference, pointing to a specific norm, is linked to a lemma, chapter, and title of the *Liber Extra*. The next Sections will detail the process.

2.1 The Annotation Schema

An annotation is a span of text marked in the text to identify some relevant information of a specific type. The annotation schema we defined identifies the four types of entities:

Annotation of type glossed lemma, "Lemma glossato", indicates the specific terms that are glossed. Legal references are included in the gloss text.

Annotations of type legal reference, "Allegazione normativa" annotate the references to legal norms or regulations.

The last two annotation types are title, "Titolo", and chapter, "Capitolo", which together with the lemma precisely identify the position of the legal reference in the *Liber Extra*.

The annotations of these entities thus enable building a link between the *Liber Extra* and the legal norms and regulations that are crucial for the interpretative framework of the book. The annotation data also include the exact character position of the beginning and the end of the annotated text in the digital text.

All annotations in the present study are restricted to contiguous spans. More complex scenarios, such as non-contiguous or overlapping annotations, which may occur in other historical corpora, are not addressed here, as they were not required by the objectives of the GNORM project. These cases would require the customization of the annotation schema to the goals of the specific documents being studied, possibly including non-contiguous or overlapping annotations which are supported by the annotation platform employed (see Section 2.2).

2.2 Expert Annotation and Annotation Platform

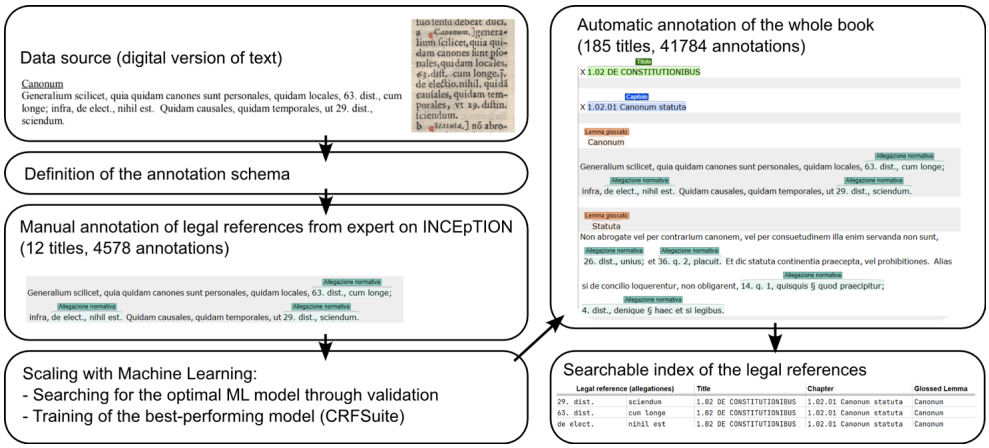

Figure 2: An illustration of the process that led to construction of the index of all the legal references in the *Liber Extra*'s *Glossa Ordinaria*.

An expert of the domain performed the annotation of the legal references on a subset of the book. The expert annotated 12 titles out of 185, making a total of 4578 annotations. The titles have been randomly sampled across the whole book to better cover the different content of each section of the book, i.e., following Reno's notation [26]: 1.02, 1.11, 1.33, 2.02, 2.23, 3.02, 3.26, 4.17, 4.19, 5.01, 5.03, 5.23. The expert focused only on annotating the legal references, as the annotations of the other three types have been made successively in a completely automatic way, as described in Section 2.3.

The tool used by the expert to perform the manual annotation is INCEpTION [16]. INCEpTION is a popular annotation platform designed for collaborative and efficient text annotation. It supports a wide range of tasks, including named entity recognition, relation extraction, and classification. INCEpTION helps the manual annotation with a dedicated GUI design, and also by providing automated suggestions for spans of text that are evaluated as potential new annotations, which the expert can validate with a single mouse click. The automated suggestions are based on the definition of a recommender, i.e., a machine-learning algorithm that trains a model by continuously observing the annotations made by the expert. The automatic annotation model trained by INCEpTION is a valuable help for the annotator, yet it is not of sufficient accuracy to be used to perform a complete automatic annotation of the rest of the book. We thus used the 12 annotated titles as a training set for a proper batch training process of more accurate models, as detailed in Section 2.3.

To situate this choice within the broader landscape of digital humanities annotation tools, it is worth briefly mentioning two related initiatives. CATMA [6] is a platform oriented toward literary text annotation and analysis, emphasizing interpretive and exploratory annotation practices. EUPORIA [3] focuses on collaborative annotation and interpretation of ancient texts, particularly in philological and exegetical contexts. While these tools address different scholarly needs, INCEpTION was selected for this project because of its strong support for supervised

machine learning, iterative recommender systems, and its support to multiple export standard formats.

2.3 Training a Model for the Automatic Annotation of Legal References

The code to replicate the automatic annotation is published with an open source license [12]. We focus here only on the legal references, as the annotation of the other types of entities has been solved using different methods, as detailed in Section 2.3.1. The automatic annotation process is modeled as a word labeling problem.

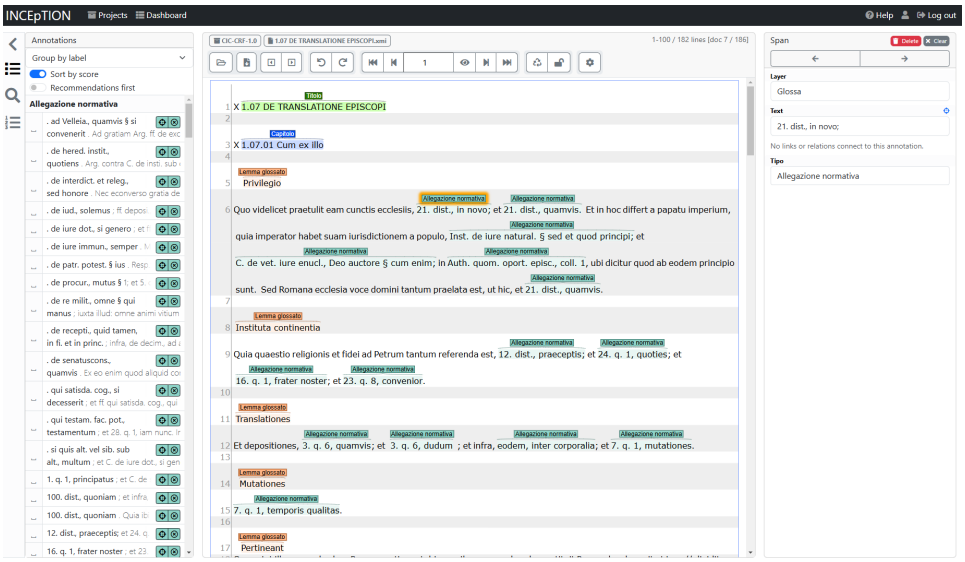

Figure 3: The annotation interface in INCEpTION. The document being annotated is shown in the center of the page. Text highlighted in different colors indicates the annotated entities of different types: orange for the "Lemma glossato" (glossed lemma), green for the "Allegazione normativa" (legal reference). The left-side bar lists all the annotations in the document, the right-side bar shows the details of the currently selected annotation.

We adopted the BILOU labeling schema [25] in which each word can be assigned to one of a set of labels, either if it is part of a legal reference or not (O label). For words that are part of a legal reference, different labels are used if the word is at the beginning of the annotation (B label), it is the last (L label), it is inside the sequence of words that form the legal reference (I label), or the legal reference is composed of a unique word (U label).

We tested two approaches to train the automatic annotation model: using a "traditional" statistical machine learning algorithm, i.e., Conditional Random Fields [29], and fine-tuning a transformer-based Large Language Model (LLM) [31]. A preliminary test of few-shot prompting [35] of a generative LLM reported a very low accuracy and was discarded.

Conditional Random Fields (CRFs) are probabilistic graphical models that are used to model sequential or structured data by defining conditional probabilities of a set of target variables given

a set of observed variables, allowing for the incorporation of context and interdependencies among the variables in the target space. In natural language processing (NLP), CRFs are often used for tasks like named entity recognition or part-of-speech tagging, where the output labels (e.g., tags) are interdependent. We tested two configurations for the CRFs graph: a simple configuration in which the observed variables are all the word bigrams in a context window of three words before and after the one to be annotated, and a rich configuration that considered bigrams and trigrams in a context window of size seven.

The process of *fine-tuning* of an LLM consists of taking a pre-trained transformer language model, such a BERT [9], and further training its parameters using task-specific annotated data, in this case to predict word-level labels corresponding to legal references. For the fine-tuning of LLMs we tested two models: the original BERT [9], as a baseline, and LatinBERT [1], a state-of-the-art model for Latin. The fine-tuning of LLMs was made using the Transformers python package [34], training for 10 epochs, using a learning rate of $2 \cdot 10^{-5}$, a weight decay value of 0.01, and a batch size of 32. Sequences longer than 512 tokens were split into multiple sequences. The comparison of the methods is based on a 10-fold cross validation on the expert-annotated data. The annotated data is split into ten parts. The splits are kept the same across all the tested methods. For every fold one tenth of the annotated data is considered to be the test data and the remaining nine tenths are the training data. The process repeats for all the folds, collecting for each tested method an automatic annotation of all data. The accuracy of the automatic annotation is determined by comparing it with the annotations from the expert, using a token-and-blank evaluation model [11].

Method	Accuracy	Training Time	Hardware	Model size
CRFs basic	0.891	7m	CPU (20 cores)	0.5MB
CRFs rich	0.978	21m	CPU (20 cores)	1.1MB
BERT [23]	0.837	13m	GPU (4 A40)	411.8MB
Latin BERT [24]	0.924	13m	GPU (4 A40)	423.3MB

Table 1: Results of cross-validation experiments (10 folds) on the expert annotated data. "Training Time" reports the time required to train an automatic annotation model on the entire set of expert-annotated data.

Results in Table 1 show that CRFs in the rich configuration performed with close to perfect accuracy. The more complex graph of the rich configuration required three times the training time of the basic configuration, yet the training time is still short and the improvement is worth the additional cost. The comparison between BERT and LatinBERT shows the impact of the main training language on a task. The tokenizer of BERT obviously struggles with Latin words.

For example, the five-word expression "Quae est radix omnium malorum" is tokenized by BERT into 19 tokens, whereas LatinBERT produces exactly five tokens. This allowed LatinBERT to better identify the relevant elements of the language that are related to the expression of legal references, whereas BERT struggled with very short sub-word tokens, which are evidently less statistically related to the concept of legal reference. The comparison of CRFs with LLMs highlights that in this annotation task CRFs have many advantage points. The obvious one is that CRFs get the best accuracy. No less important is that CRFs require fewer computational

resources. The training time of CRFs is based on using a desktop with a single Intel i9 CPU, while the LLMs' training time is based on a dedicated server with 4 A-40 NVIDIA GPUs, costing roughly ten times the desktop computer.

Similarly, the sizes of the final models show a clear advantage for CRFs. The differences in training time are not very relevant, considering the differences in hardware, and the fact that even the longest training time is relatively quick. CRFs make it possible to train an annotation model on personal hardware, enabling the adoption of machine learning to small research groups with limited computational resources. More experiments on the fine-tuning of LLMs may have reduced the gap with CRFs, yet the high accuracy obtained by CRFs satisfied our goals, thus not justifying the additional computational costs.

2.3.1 Automatic Annotation of Other Entities

The annotation of chapters and titles has been made using a regular expression, exploiting the specific format used in The Digital Decretals [26]. For the annotation of the glossed lemmas, we had only an ordered list of the glossed lemmas, leaving us to identify their position in the text. We solved this with a search algorithm that considered the constraints imposed by the list. For example, the word "Omnipotens" occurs 31 times in the whole text, but the only instance as a glossed lemma occurs after the glossed lemma "Incomprehensibilis" and before "Ineffabilis."

Solving all the constraints for all of the glossed lemmas allowed us to find the exact position for all of them.

2.4 Building the Index of Legal References

The automatic annotation of the whole text identified 41,784 legal references [13], each linked to a glossed lemma, a chapter, a title, and a book part (among the five parts of the Ordinary Gloss relative to the five books of the *Liber Extra*). The automatic annotation thus had a multiplier effect of approximately 9.1 with respect to the number of annotations from the expert. We split each legal reference in two parts, the "title" or "section" that contains the referenced norm, and the string that identifies the norm. For example, the legal reference "22.q.4, incommutabilis" points to Gratian's *Decretum*, *Causa* 22, *Quaestio* 4, and its chapter (or canon) "incommutabilis" (according to today's citation standard: C.22 q.4 c.9). This allowed us to identify a total of 1,795 referenced titles or sections from various collections, i.e., the *Corpus Iuris Canonici* and the *Corpus Iuris Civilis*. The index enables the researchers to perform the activities presented in Section 1 and detailed with examples in Section 4.

2.5 Data Formats, Standards, and Interoperability

In addition to their inclusion in the GNORM software, the index and the annotated corpus from which the index is derived are published with a long-term preservation and reuse perspective [13].

The index is published in a simple tabular form (in CSV format), in which each row is composed of six fields that identify the source and target of the legal reference, i.e., the position of the legal allegation in the *Liber Extra* defined by the Book, Title, Chapter, and Glossed Lemma (e.g., "Book 4", "4.19 DE DIVORTIIS", "4.19.05 Ex litteris tuis", "Redire compellas"), and the target of the legal allegation defined by the Legal reference (title/section, e.g., "de consuet."), and Legal reference (chapter/norm, e.g., "ex parte").

The annotated corpus is published in two versions: the expert's annotations, and the full annotation that merges the expert's annotations and the automatic annotations.

Annotations are published in multiple formats, including the native INCEpTION format, and various export format supported by the software, i.e., CoNLL, UIMA CAS, NIF, TEI XML, and WebAnno TSV formats. These formats facilitate interoperability with existing digital humanities infrastructures and support alignment with established best practices in digital text encoding ([30];[27]).

3. Qualitative Analysis

A qualitative evaluation of the results of the automatic annotation process is particularly relevant when considering the degree of correctness and accuracy in relation to the amount of data processed. Regarding the correctness of the annotation, the system's ability to distinguish legal references from the rest of the text—which is usually written in a discursive and argumentative style, according to the explanatory function inherent in the gloss—is impressive. Concerning the accuracy of the annotation, the system has demonstrated, in most cases, its ability to delimit the exact scope of legal references.

There are a few exceptions when the text of the allegation is unusually long or contains overly specific references. There are no particular issues in identifying allegations that refer to the *Liber Extra* or to the three partes of Gratian's *Decretum*, as each is clearly defined in its citation form.

The same holds true for the allegations to the Code or the Institutes of the Corpus Iuris Civilis, which are usually clearly preceded by their respective symbols (C. or Inst.). The automatic annotation has a recurring imprecision when identifying references to the Digest, preceded by the abbreviation "ff.", which is erroneously omitted in many cases. In this case the automatic annotation model shows a preference for shorter annotations, considering that in most cases both versions of the annotations, with or without "ff.", are potentially correct yet referring to different sources. The specific nature of this error made it easy to solve with a simple automatic post processing of the annotations, checking for any eventual "ff." preceding them and adding it to the annotation, as we found no cases in which the "ff." expression preceding an annotation was to be excluded. The published version of the index. [11] is thus not affected by this issue.

Liber Extra		Book 1		Book 2		Book 3		Book 4		Book 5	
de elect.	1217	de elect.	739	de appell.	499	de praeben.	225	de spons.	200	de sent. excom.	405
de appell.	930	de offi. deleg.	366	de testib.	363	12. q. 2	181	de despon. impub.	100	de simon.	223
de offi. deleg.	721	de rescript.	361	de offi. deleg.	219	de iure patron.	177	de eo qui dux.	63	de accusat.	172
de sent. excom.	689	de appell.	296	de elect.	207	de elect.	164	qui fil. sint legit.	59	11. q. 3	150
de rescript.	585	de praeben.	158	de praescrip.	153	16. q. 1	162	de eo qui cog. cons.	55	de homic.	127
de testib.	545	ff. de procur.	121	de restit. spol.	145	de decim.	160	de cond. appos.	48	de privileg.	116
de praeben.	439	de re iudic.	104	de iureiur.	135	de conver. coniug.	108	de frig. et malef.	45	1. q. 1	114
12. q. 2	388	ff. de recepti.	103	de re iudic.	132	16. q. 7	108	de divort.	37	de haeret.	107
de simon.	379	de sent. excom.	91	2. q. 6	129	de censib.	103	27. q. 2	36	50. dist.	102
11. q. 3	359	de renunciat.	91	de probat.	124	de concess. praeben.	95	de cons. et affin.	34	de elect.	87

Table 2: The ten sections or titles with most references in the Ordinary Gloss, considering the whole *Liber Extra*, or each book of the *Liber Extra* separately. Numbers indicate the number of references.

Quantitatively insignificant are the cases in which the system detects text passages as legal allegations when they are not; this typically results from the peculiar structure of the text in question. Particularly interesting are the cases where the system, due to the syntactic structure of the text, has detected actual allegations that are "non-legal," such as references to Gospel passages. In conclusion, the large amount of correct annotations and the limited effort required to correct the erroneous ones—which will be the next step of the project's activities—confirms the validity and potential of the approach to annotation we presented.

4. Statistics on the Index

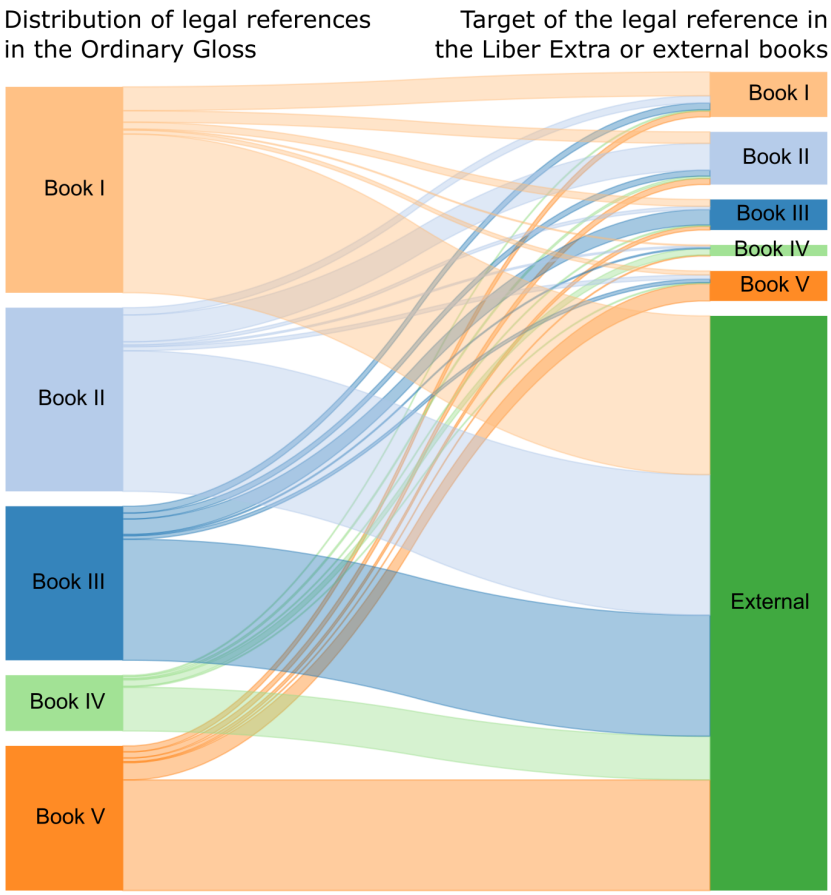

Figure 4: Sankey diagram showing the distribution of legal references in the Ordinary Gloss (left part) in relation to the distribution of the targets of the legal references, either in the *Liber Extra* or external books (right part).

The statistical analysis presented in this section is based on the final version of the index, which incorporates corrections derived from the qualitative analysis described in Section 3, including post-processing adjustments to Digest references ("ff.").

In addition to the use as a browsing resource, the index itself can be the subject of analysis. Table 2 lists the ten titles or sections with the most references in the Ordinary Gloss, considering the whole *Liber Extra* and each of its books separately. Just from this table the structure of the content in the *Liber Extra* emerges. For example, it is known that Book IV of the *Liber Extra* is dedicated to marital matters and in fact the relevant column in the table contains numerous references to norms about marriage.

Figure 4 shows the number of legal references in each book of the Ordinary Gloss and to which book of the *Liber Extra*, or external book, they refer. We can see how roughly one third of the legal references are references to other norms of the *Liber Extra* and thus can be considered to be internal references, while the other two thirds are referring to other books. With respect to internal references, in each book the references to the book itself are always the most frequent, with Book I as the second source of internal references for all the books.

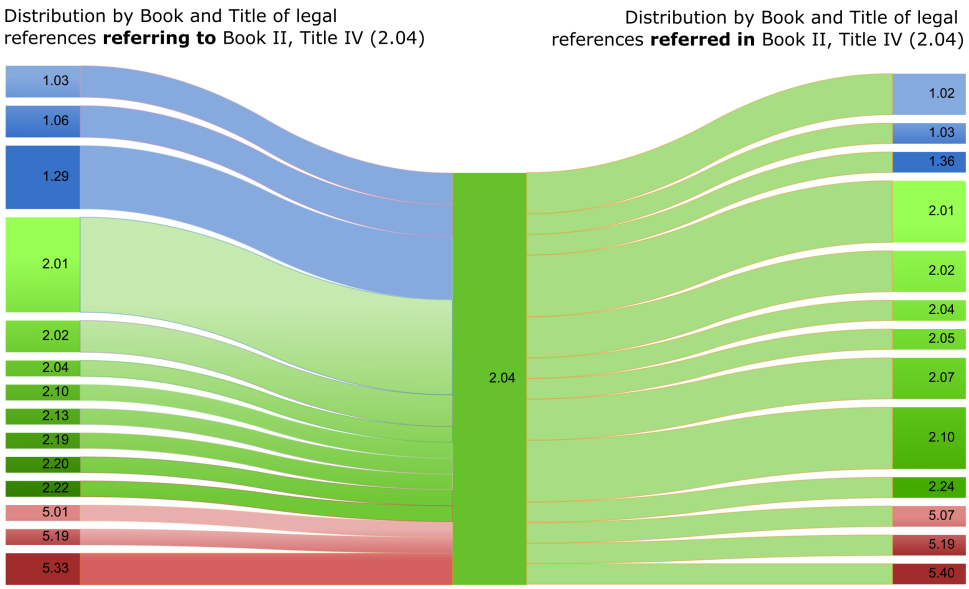

Figure 5: Sankey diagram showing the distribution of incoming and outgoing legal references in the Ordinary Gloss for a specific book and title (2.04). Books and Titles are indicated with decimal numbers (book.title). Only references inside the Ordinary Gloss are shown.

As shown in Figure 5, the analysis of the structure of the references can go into more detail by focusing on specific parts of the Ordinary Gloss. For example, Figure 5 shows the Books and Titles of the Ordinary Gloss containing legal references to the norms of Book II Title IV. It also shows the opposite case, i.e., to which Books and Titles the legal references of Book II Title IV point.

This visualization presents an effective way to show the relations between the norms of the *Liber Extra* and can also be a visual tool to guide the navigation across the norms.

5. Future Work

A promising direction for future development is the alignment of legal references with existing taxonomies, ontologies, or knowledge graphs ([2];[18]). Such integration would enable semantic web–based exploration and reasoning, further strengthening the role of annotated glosses within interoperable digital legal history infrastructures. In this perspective, the index produced in this work can be seen as a first step toward a formally modelled legal knowledge graph.

The proposed annotation pipeline can be straightforwardly transferred to other glossed legal corpora provided that certain prerequisites are met: the availability of a stable reference text, consistent citation conventions, and a representative subset of expert annotations. Applying the approach to corpora in other languages would require language-specific feature engineering or pre-trained language models.

More complex annotation scenarios, including non-contiguous or overlapping spans, may arise in other traditions and they would require adaptations of both the annotation schema, which is supported by the current annotation platform, and the learning models, for which there are already valid solutions ([17];[22];[32])

6. Conclusions

In the field of Legal History, the possibility of using an automatic annotation system for legal allegations is proving to be a valuable tool. The development of coordinated and interconnected databases of normative texts and their commentary apparatus, as well as the creation of digital editions of fundamental works of medieval jurists, are just a few examples of potential applications. Moreover, the creation of specific reference ontologies could enable the training of increasingly complex retrieval systems, capable of extracting and sorting a vast amount of data that would otherwise be unmanageable, making larger projects practically infeasible.

In the case study presented in this work, we have been able to annotate a large amount of text by exploiting the annotation of a human expert on a tenth of the corpus and then a very effective machine learning setup to achieve a complete, accurate annotation of the whole text. The outcome of this research activity is twofold: we produced a valuable resource, the index, that will contribute to the GNORM software and support studies on the Ordinary gloss and the *Liber Extra*; and we have proved an effective low-resource pipeline that can be replicated for similar activities in the field of Religious Studies, Legal History, and many related disciplines in the humanities.

Acknowledgements

This work was supported by project "Italian Strengthening of ESFRI RI RESILIENCE" (ITSERR) funded by the European Union under the NextGenerationEU funding scheme (CUP: B53C22001770006).

Author Contributions

Esuli and Puccetti wrote the code, ran the experiments, and wrote Section 2. Imperia made the manual annotation, and wrote Sections 1 and 3. All the authors wrote abstract, conclusions, and revised the final version of the paper.

References

[1] Bamman, D., and P. J. Burns. 2020. "Latin BERT: A Contextual Language Model for Classical Philology." arXiv preprint arXiv:2009.10053.

[2] Barzaghi, S., Palmirani, M., & Peroni, S. (2020). Development of an ontology for modelling medieval manuscripts: the case of Progetto IRNERIO. Umanistica Digitale, 4(9), 117–140. https://doi.org/10.6092/issn.2532-8816/11187

[3] Boschetti, F., Bambaci, L., Del Grosso, A. M., Mugelli, G., Khan, A. F., Bellandi, A., & Taddei, A. (2023). Collaborative and Multidisciplinary Annotations of Ancient Texts: The Euporia System. The Ancient World Goes Digital, 6, 172-223.

[4] Bellomo, M. 1995. The Common Legal Past of Europe, 1000–1800.

[5] Bernard of Parma. 1582. Glossa Ordinaria to Decretals Gregory IX. In Decretales D. Gregorii Papae IX. suae integritati una cum glossis restitutae. Cum privilegio Gregorii XIII. Pont. Max. et aliorum Principum. Roma: In Aedibus Populi Romani.

[6] Gius, E., Meister, J. C., Meister, M., Petris, M., Gerstorfer, D., Akazawa, M., & Messner, S. (2025). CATMA (7.2.0). Zenodo. https://doi.org/10.5281/zenodo.1470118

[7] Conte, E. 2021. "The Centre and the Margins of the Jungle of Glossed Manuscripts." Rivista Internazionale di Diritto Comune: 55–73.

[8] Decretales D. Gregorii Papae IX. ~~1582~~. Suae Integritati Una Cum Glossis Restitutae. Cum Privilegio Gregorii XIII. Pont. Max. et aliorum Principum. 1582. Roma: In Aedibus Populi Romani. Accessed 2024. https://digital.library.ucla.edu/catalog/ark:/21198/zz0014rx7w?cv=35.

[9] Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), edited by J. Burstein, C. Doran, and T. Solorio, 4171–4186. Minneapolis: Association for Computational Linguistics. https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[10] Dolezalek, G. 2021. "Glosses and the Juridical Genre Apparatus glossarum in the Middle Ages." Rivista Internazionale di Diritto Comune 32: 9-54.

[11] Esuli, A., and F. Sebastiani. 2010. “Evaluating Information Extraction.” In International Conference of the Cross-Language Evaluation Forum for European Languages, 100–111. Springer.

[12] Esuli, A., and G. Puccetti. 2024. “A Machine Learning Pipeline to Automatically Annotate Legal References (Allegationes) in the Liber Extra’s Ordinary Gloss.” https://github.com/aesuli/CIC_annotation. doi:10.5281/zenodo.14381817.

[13] Esuli, A., V. R. Imperia, and G. Puccetti. 2025. “Automatic Annotation of the Legal References in the LiberExtra’s Ordinary Gloss (Version 2.0) [Data set].” doi:10.5281/zenodo.17953666.

[14] Hespanha, A. M. 2008. “Form and Content in Early Modern Legal Books.” Rechtsgeschichte-Legal History 12: 12–50.

[15] Kantorowicz, H. U. 1935. “Die Allegationen im Späteren Mittelalter.” Archiv für Urkundenforschung: 15–29.

[16] Klie, J.-C., M. Bugert, B. Boullosa, R. E. de Castilho, and I. Gurevych. 2018. “The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation.” In Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, 5–9. Association for Computational Linguistics. http://tubiblio.ulb.tu-darmstadt.de/106270/.

[17] Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3036–3046, Minneapolis, Minnesota. Association for Computational Linguistics.

[18] García-Menéndez, Á., Labra-Gayo, J. E., & Gayo-Avello, D. (2024). Unlocking Historical Knowledge: A Semantic Web Approach to Medieval Notarial Document Analysis. CEUR Workshop Proceedings, 3967.

[19] Menzinger, S. 2017. “Reflections on the Connection between Author and Text in Medieval Juridical Production.” Historia et Ius 11.

[20] Menzinger, S. 2019. "The Past, the Others, Himself: The Open Dialogue of a Medieval Legal Author with His Text." S. Boodts, P. De Leemans, S. Schorn (eds.), *Sicut dicit: editing ancient and medieval commentaries on authoritative texts,* Turnhout, Brepols, 2019, pp. 273-299.

[21] Menzinger, S. 2022. "Interazione tra Testo e 'Citazione' nella Dottrina Giuridica Civilistica: Secoli XII e XIII." In Juristische Glossierungstechniken als Mittel Rechtswissenschaftlicher Rationalisierungen, 15–26. Erich Schmidt Verlag.

[22] Aldrian Obaja Muis and Wei Lu. 2016. Learning to Recognize Discontiguous Entities. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 75–84, Austin, Texas. Association for Computational Linguistics.

[23] Pennington, K. 2012. "Corpus iuris canonici." In *Diccionario General de Derecho Canónico*, edited by J. Otaduy, A. Viana, and J. S. Rueda, 757–765. Thomson Reuters Aranzadi.

[24] Quaglioni, D. 2019. "Licet Allegare Poetas: Formanti Letterari del Diritto fra Medioevo ed Età Moderna." in F. Meier, E. Zanin, *Poesia e diritto nel Due e Trecento italiano*, Ravenna, Longo Editore, 2019, pp. 209-220.

[25] Ratinov, L., and D. Roth. 2009. "Design Challenges and Misconceptions in Named Entity Recognition." In Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), 147–155.

[26] Reno, E. 2024. The Digital Decretals. Accessed November 1, 2024. https://www.digitaldecretals.com/.

[27] Sanderson, R., Ciccarese, P., & Van de Sompel, H. (2017). Web Annotation Data Model (W3C Recommendation). World Wide Web Consortium. https://www.w3.org/TR/annotation-model/

[28] Speciale, G. 2001. "Apparatus: Ipertesto Vivo e Aperto." Ius Commune. Zeitschrift für Europäische Rechtsgeschichte 28: 47–59.

[29] Sutton, C., and A. McCallum. 2012. "An Introduction to Conditional Random Fields." Foundations and Trends in Machine Learning 4: 267–373.

[30] TEI Consortium. (2025). TEI P5: Guidelines for Electronic Text Encoding and Interchange (version P5 4.10.2). The TEI Consortium. *https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html*

[31] Vaswani, A. 2017. "Attention Is All You Need." In Advances in Neural Information Processing Systems.

[32] Yu Wang, Hanghang Tong, Ziye Zhu, and Yun Li. 2022. Nested Named Entity Recognition: A Survey. ACM Trans. Knowl. Discov. Data 16, 6, Article 108 (December 2022), 29 pages. https://doi.org/10.1145/3522593

[33] Weimar, P. 1967. “Argumenta Brocardica.” Studia Gratiana 14: 89–123.

[34] Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, and M. Fun-Towicz. 2020. “Transformers: State-of-the-Art Natural Language Processing.” In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45.

[35] Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. 2019. “Language Models Are Unsupervised Multitask Learners.” OpenAI Blog 1: 9.