A Medieval Epigraphic Corpus and its Retro-Developments (CIFM-CBMA). The Exploratory Research of the COSME2 Consortium

Eliana Magnani, Nicolas Perreaux

A Medieval Epigraphic Corpus and its Retro-Developments (CIFM-CBMA): The Exploratory Research of the Cosme^2 Consortium

Eliana Magnani, Nicolas Perreaux

https://doi.org/10.1093/llc/fqaa069

Abstract

The digital 'Burgundian Epigraphic Corpus' is the result of a collaboration between two teams, the *Corpus of Inscriptions of Medieval France* (CIFM) and the *Corpus of Medieval Burgundian Texts* (CBMA), as part of the Cosme^2 (*Consortium Sources Médiévales*—linked to TGIR Huma-Num from CNRS—France), dedicated to digital approaches to historical corpora. This article explains how a complex set of documents mixing Latin, Greek, and Old French texts, accompanied by rich metadata, has been processed in order to allow new surveys by humanists. It shows how the corpus is constantly reinvested and how its exploitation, thanks to digital methods, generates new data and metadata that can be reinjected into the corpus and in turn operated, creating a kind of virtuous circle. Three retro-developments are briefly discussed here: (1) semantic web, connectivity, and named entities; (2) geographic information system (GIS) and automated extraction of new metadata; (3) lemmatization and automatic language detection.

A Framework: The Cosme^2's Consortium

Started in 2017, the Cosme^2 project (*Consortium Sources Médiévales*—linked to TGIR Huma-Num from CNRS) is dedicated to digital approaches to historical corpora, especially for French scholars working on the Middle Ages.[1] Created by medievalists, it is based on a scientific community broadly open to all disciplines: archaeology, history, art history, philosophy, linguistics, and literary studies. Its aim is to show how digital corpora can allow a 'return to the sources' (i.e. to medieval documentation), in order to build a community and propose frameworks, standards, and ideas around applied Digital

Humanities. Ultimately, the purpose of the consortium is to help with the creation, interoperability, and publication of standardized digital corpora online.

This article concerns more precisely a project developed within the 'Lemmas' working group (Magnani, 2019b).[2] This group aims to encourage the development of tools for the lemmatization of medieval textual sources, and the open access publication of lemmatized corpora.

There are relatively few lemmatized corpora in medieval history to this day, because this process is confronted with various technical difficulties. Beyond the heterogeneity of the available corpora, the lemmatization of mediolatin and vernacular texts must deal with the immense variety of forms of the medieval lexicon, attributable to declensions, regionalisms, and graphical variations. In this article, we want to show how a complex set of documents—the Burgundian epigraphic corpus—mixing Latin, Greek, and Old French texts, accompanied by rich metadata, has been processed in order to allow new surveys by humanists.

2 The Burgundian Epigraphic Corpus: From Stone to Digital

The digital 'Burgundian Epigraphic Corpus' is the result of the collaboration between two teams, in 2018 and 2019, with extensive experience in the elaboration of medieval textual corpora: the *Corpus of Inscriptions of Medieval France* (CIFM)[3] and the *Corpus of Medieval Burgundian Texts* (CBMA).[4] These two programs benefit from the important services of the TGIR Huma-Num: website hosting, data storage, sharing and archiving, server space, etc.

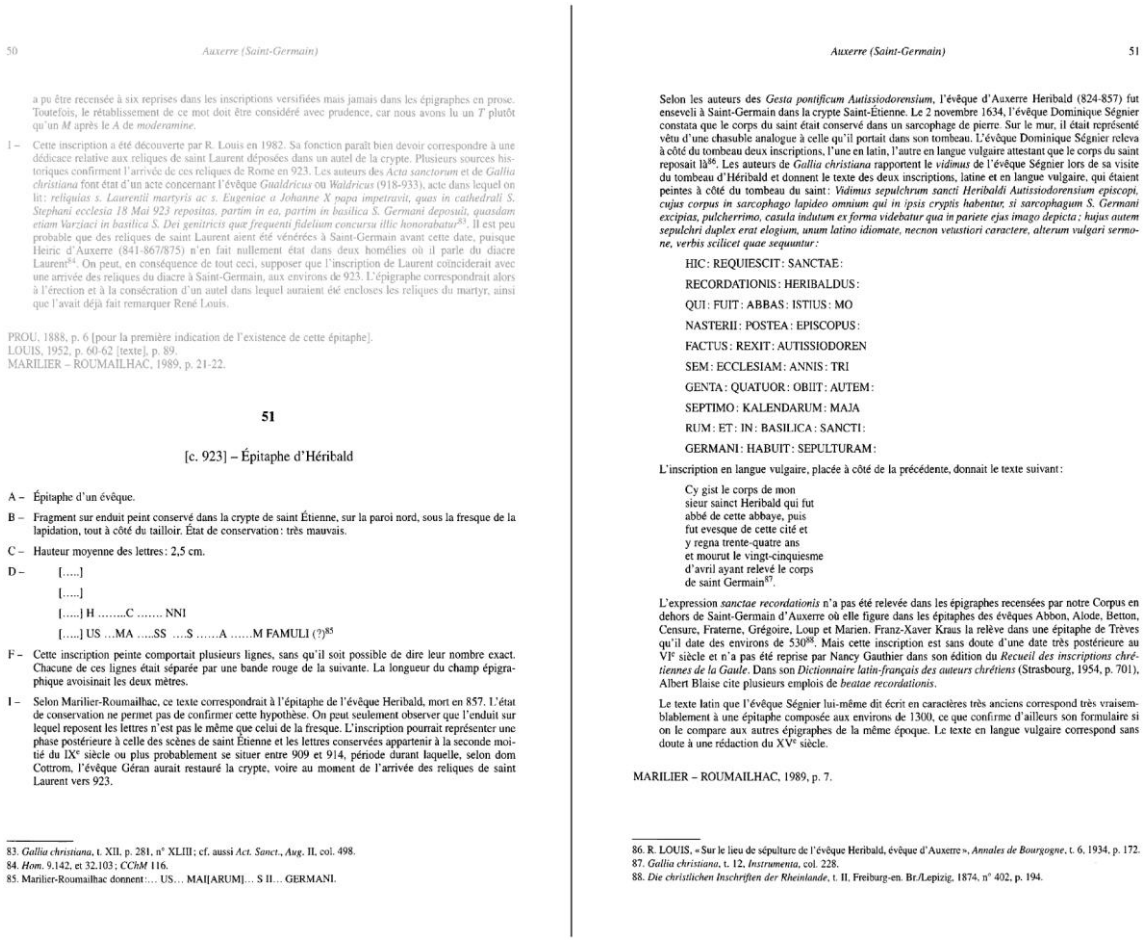
The CIFM, led by the *Centre d'études supérieures de civilisation médiévale* (CESCM) at Poitiers, is widely known for its work in the field of epigraphy.[5] Since the 1970s, it has published twenty-seven volumes of the '*Corpus des inscriptions de la France médiévale (Corpus of Inscriptions of Medieval France)*' (CIFM) (Favreau, 1982), now OCR'd and available online on the Persée platform.[6] The team is currently developing a program called

TITULUS for online publishing of inscriptions. The CBMA team, led by the *Laboratoire de médiévistique occidentale de Paris* (LaMOP), develops a platform and database that mainly contains more than 29,000 diplomatic documents, but also more than 300 hagiographic texts, and now, thanks to this current project, more than 1,400 epigraphic inscriptions. Initially conceived as a corpus of diplomatic texts (i.e. charters), CBMA has developed a more global perspective, and seeks to include documents of very different types (Magnani, 2019a). This corpus of Burgundian epigraphic inscriptions is the very purpose of the joint project with the Poitiers team,[7] and is at the heart of our project: epigraphy is indeed an ideal point of comparison, because it uses formulas (such as diplomatic texts), makes extensive use of named entities, but also uses a lexicon that is usually found in narrative texts. However, this epigraphic corpus raises various difficulties linked to its heterogeneity. It is indeed edited in paper format, as we have said, but it is also largely supplemented by the scholarly observations of CIFM collaborators. These have accumulated in the form of complex sheets. The use of this corpus thus requires us to take into account a double reality: on the one hand, the edited volumes, which are fairly easy to formalize; on the other hand, these numerous and heterogeneous records, which require the intervention of humans if we wish to integrate them into the corpus (which is necessary, given the amount of information added). The formalization of the corpus and its subsequent historical analysis therefore required the gathering of this information and its articulation, in particular via metadata.

Our first goal was thus to extract the texts and their metadata from the printed and OCR'd volumes of the CIFM (Favreau *et al.*, 1997, 1999; Favreau and Michaud, 2000) (Fig. 1), and incorporate them into an existing database, the CBMA. Within the corpus, the inscriptions are presented according to a more or less standardized framework, which allows precisely the extraction of information by digital process. In addition to the usual metadata for historical corpora (chronology, place of origin, place of conservation, etc.), there is in the CIFM a variety of information ranging from the nature of the medium (stone, wood, etc.), the height of the letters and more palaeographic elements, and historiographic analyses. In the case of the 10th-century inscription taken as

an example in Fig. 1, we note that the 'average height of the letters' is 2.5cm; that it is an episcopal epitaph; which is in a very poor state of preservation, because it was painted on plaster in the crypt of Saint Étienne's cathedral in Auxerre. Since 2017, the CBMA team wished to develop its corpus, in order to make it multitypological and multilingual, which will eventually allow cross queries on different types of documents. This approach is fundamental for Medieval History, as the scribes who wrote the theological manuscripts were often the same ones who produced the charters, also commissioning the paintings, sculptures, and inscriptions. The presence of very rich metadata within the CIFM makes it possible precisely to create links between the different documentary typologies and is therefore a crucial issue for our new approach, but also a major challenge (because metadata sets do not match from one documentary type to another). From this point of view, the CIFM is not a simple textual corpus, in the sense that significant information is not exclusively contained in the text: the materiality of the inscription, its very history, conveys information that must be taken into account in the analyses. While this situation applies more or less to all historical corpora, it is particularly true here because of the 'memorial' nature of the objects included in the CIFM, which makes the metadata surrounding the text particularly meaningful.

Fig. 1



50 Auxerre (Saint-Germain)

a pu être recensée à six reprises dans les inscriptions versifiées mais jamais dans les épigraphes en prose. Toutefois, le rétablissement de ce mot doit être considéré avec prudence, car nous avons lu un *T* plutôt qu'un *M* après le *A* de *moderamine*.

I – Cette inscription a été découverte par R. Louis en 1982. Sa fonction paraît bien devoir correspondre à une dédicace relative aux reliques de saint Laurent déposées dans un autel de la crypte. Plusieurs sources historiques confirment l'arrivée de ces reliques de Rome en 923. Les auteurs des *Acta sanctorum* et de *Gallia christiana* font état d'un acte concernant l'évêque *Gualdricus* ou *Waldricus* (918-933), acte dans lequel on lit: *reliquias s. Laurentii martyris ac s. Eugeniae a Johanne X papa impetravit, quas in cathedrali S. Stephani ecclesia 18 Mai 923 repositas, partim in ea, partim in basilica S. Germani deposuit, quasdam etiam Varziaci in basilica S. Dei genitricis quæ frequenti fidelium concursu illic honorabatur*[83]. Il est peu probable que des reliques de saint Laurent aient été vénérées à Saint-Germain avant cette date, puisque Heiric d'Auxerre (841-867/875) n'en fait nullement état dans deux homélies où il parle du diacre Laurent[84]. On peut, en conséquence de tout ceci, supposer que l'inscription de Laurent coïnciderait avec une arrivée des reliques du diacre à Saint-Germain, aux environs de 923. L'épigraphe correspondrait alors à l'érection et à la consécration d'un autel dans lequel auraient été enclosses les reliques du martyr, ainsi que l'avait déjà fait remarquer René Louis.

PROU, 1888, p. 6 [pour la première indication de l'existence de cette épitaphe].
LOUIS, 1952, p. 60-62 [texte], p. 89.
MARILIER – ROUMAILHAC, 1989, p. 21-22.

51

[c. 923] – Épitaphe d'Héribald

A – Épitaphe d'un évêque.

B – Fragment sur enduit peint conservé dans la crypte de saint Étienne, sur la paroi nord, sous la fresque de la lapidation, tout à côté du tailloir. État de conservation: très mauvais.

C – Hauteur moyenne des lettres: 2,5 cm.

D – [.....]
[.....]
[.....] H ……..C ……. NNI
[.....] US …MA …..SS ….S ……A ……M FAMULI (?)[85]

F – Cette inscription peinte comportait plusieurs lignes, sans qu'il soit possible de dire leur nombre exact. Chacune de ces lignes était séparée par une bande rouge de la suivante. La longueur du champ épigraphique avoisinait les deux mètres.

I – Selon Marilier-Roumailhac, ce texte correspondrait à l'épitaphe de l'évêque Heribald, mort en 857. L'état de conservation ne permet pas de confirmer cette hypothèse. On peut seulement observer que l'enduit sur lequel reposent les lettres n'est pas le même que celui de la fresque. L'inscription pourrait représenter une phase postérieure à celle des scènes de saint Étienne et les lettres conservées appartenir à la seconde moitié du IX[e] siècle ou plus probablement se situer entre 909 et 914, période durant laquelle, selon dom Cottrom, l'évêque Géran aurait restauré la crypte, voire au moment de l'arrivée des reliques de saint Laurent vers 923.

83. *Gallia christiana*, t. XII, p. 281, n° XLIII; cf. aussi *Act. Sanct., Aug.* II, col. 498.
84. *Hom.* 9.142, et 32.103; *CChM* 116.
85. Marilier-Roumailhac donnent: … US… MAI[ARUM]… S II… GERMANI.

Auxerre (Saint-Germain) 51

Selon les auteurs des *Gesta pontificum Autissiodorensium*, l'évêque d'Auxerre Heribald (824-857) fut enseveli à Saint-Germain dans la crypte Saint-Étienne. Le 2 novembre 1634, l'évêque Dominique Ségnier constata que le corps du saint était conservé dans un sarcophage de pierre. Sur le mur, il était représenté vêtu d'une chasuble analogue à celle qu'il portait dans son tombeau. L'évêque Dominique Ségnier releva à côté du tombeau deux inscriptions, l'une en latin, l'autre en langue vulgaire attestant que le corps du saint reposait là[86]. Les auteurs de *Gallia christiana* rapportent le *vidimus* de l'évêque Ségnier lors de sa visite du tombeau d'Héribald et donnent le texte des deux inscriptions, latine et en langue vulgaire, qui étaient peintes à côté du tombeau du saint: *Vidimus sepulchrum sancti Heribaldi Autissiodorensium episcopi, cujus corpus in sarcophago lapideo omnium qui in ipsis cryptis habentur, si sarcophagum S. Germani excipias, pulcherrimo, casula indutum ex forma videbatur qua in pariete ejus imago depicta; hujus autem sepulchri duplex erat elogium, unum latino idiomate, necnon vetustiori caractere, alterum vulgari sermone, verbis scilicet quae sequuntur:*

HIC: REQUIESCIT: SANCTAE:
RECORDATIONIS: HERIBALDUS:
QUI: FUIT: ABBAS: ISTIUS: MO
NASTERII: POSTEA: EPISCOPUS:
FACTUS: REXIT: AUTISSIODOREN
SEM: ECCLESIAM: ANNIS: TRI
GENTA: QUATUOR: OBIIT: AUTEM:
SEPTIMO: KALENDARUM: MAJA
RUM: ET: IN: BASILICA: SANCTI:
GERMANI: HABUIT: SEPULTURAM:

L'inscription en langue vulgaire, placée à côté de la précédente, donnait le texte suivant:

Cy gist le corps de mon
sieur sainct Heribald qui fut
abbé de cette abbaye, puis
fut evesque de cette cité et
y regna trente-quatre ans
et mourut le vingt-cinquiesme
d'avril ayant relevé le corps
de saint Germain[87].

L'expression *sanctae recordationis* n'a pas été relevée dans les épigraphes recensées par notre Corpus en dehors de Saint-Germain d'Auxerre où elle figure dans les épitaphes des évêques Abbon, Alode, Betton, Censure, Fraterne, Grégoire, Loup et Marien. Franz-Xaver Kraus la relève dans une épitaphe de Trèves qu'il date des environs de 530[88]. Mais cette inscription est sans doute d'une date très postérieure au VI[e] siècle et n'a pas été reprise par Nancy Gauthier dans son édition du *Recueil des inscriptions chrétiennes de la Gaule*. Dans son *Dictionnaire latin-français des auteurs chrétiens* (Strasbourg, 1954, p. 701), Albert Blaise cite plusieurs emplois de *beatae recordationis*.

Le texte latin que l'évêque Ségnier lui-même dit écrit en caractères très anciens correspond très vraisemblablement à une épitaphe composée aux environs de 1300, ce que confirme d'ailleurs son formulaire si on le compare aux autres épigraphes de la même époque. Le texte en langue vulgaire correspond sans doute à une rédaction du XV[e] siècle.

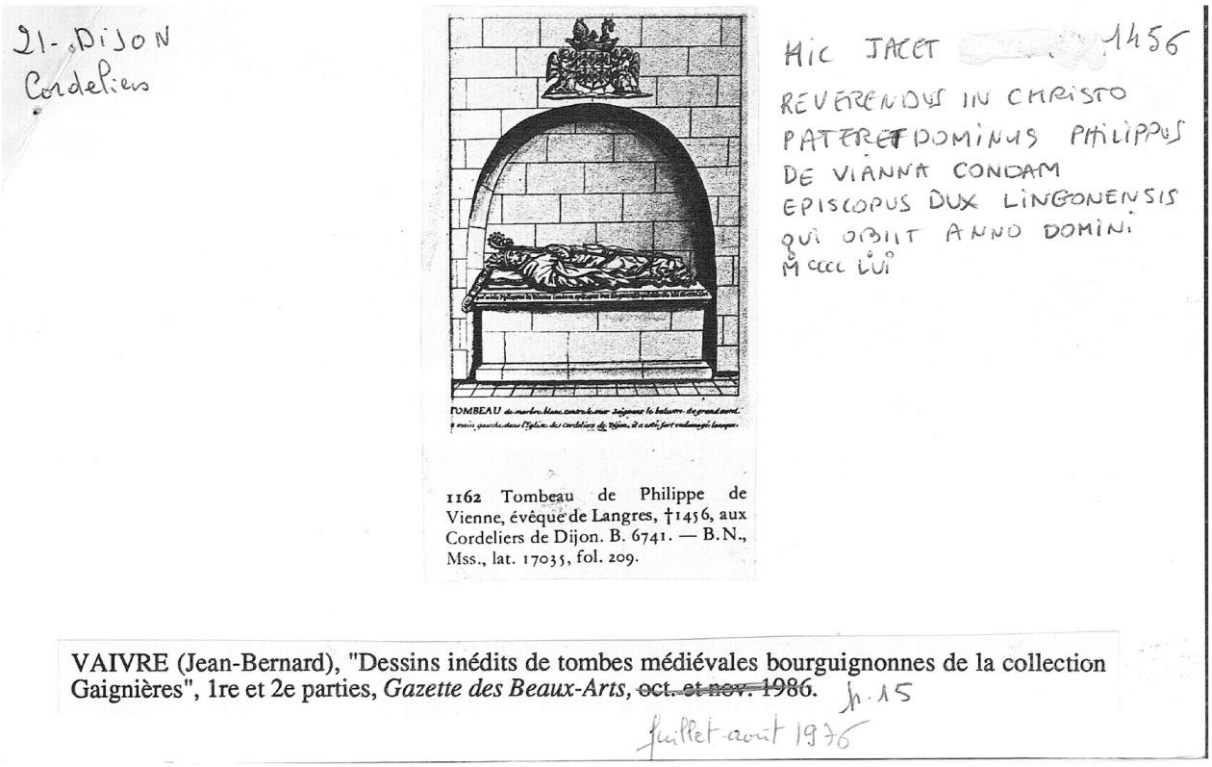
MARILIER – ROUMAILHAC, 1989, p. 7.

86. R. LOUIS, « Sur le lieu de sépulture de l'évêque Heribald, évêque d'Auxerre », *Annales de Bourgogne*, t. 6, 1934, p. 172.
87. *Gallia christiana*, t. 12, *Instrumenta*, col. 228.
88. *Die christlichen Inschriften der Rheinlande*, t. II, Freiburg-en. Br./Lepizig, 1874, n° 402, p. 194.

Example of a presentation form for an entry in the CIFM. This is the epitaph of Heribald (ca. 923) (Favreau and Michaud, 2000, pp. 50–1) (with CESCM-CIFM's permission)

Several operations of linking, enriching, and recording between the two structures, CIFM and CBMA, have thus been carried out, but we cannot detail them all here. However, two processes are important to mention: (1) the first is the manual intervention of an epigrapher, Aurore Menudier, to add the 'unpublished' texts that were in the old paper catalogue in Poitiers (Fig. 2)[8]; (2) the other is the enrichment of metadata, in particular with the geolocation coordinates of inscriptions, done in an automated way, since the location indications were already filled in. All the documentation collected and the files generated are available on the project web page and its GitLab.[9] The texts of the inscriptions are searchable with Philologic4,[10] but they also have been lemmatized and can be imported directly into the TXM software, and soon into NoSketchEngine.[11]

Fig. 2



Example of a record in the old CIFM paper catalogue. Funeral inscription for Philip of Vienne, bishop of Langres (1456) (with CESCM-CIFM's permission)

3 Retro-Developments: How Metadata Can Be a New Start?

During the development of this corpus, the team faced several challenges. One of the most important is in fact very common: while many medieval corpora exist nowadays, they are still seldom used (Perreaux, 2014). So, how to move from the state of 'production/accumulation' of textual (or iconographic) databases to the writing of new scientific articles, taking advantage of the benefits offered by digital methods, but also corresponding to the disciplinary standards and expectations? This challenge is not easy, because the consultation of digital corpora often differs from those in paper editions, bringing new opportunities but also difficulties (Guerreau, 2011). One of the main epistemological problems raised by the use of corpora in history is probably (and paradoxically) the mass of occurrences that the method allows to identify. If the historical method is adapted to the treatment of small series

of texts/objects/images, it is much less so when it comes to analyzing thousands of them (Guerreau, 2001, pp. 163–90; Perreaux, 2014)—which is particularly clear in the case of texts such as those mentioned above. This difficulty is not only a question of scale: it imposes a profound change in the historical method, which must rethink its heuristics (i.e. how to present documents, how to process them, how to analyze the results, how to include graphs and tables in the demonstration and writing, etc.). How can one go from these hundreds of documents to a historical narrative that corresponds to the traditional academic framework? The linearity of writing, in particular, requires that the material be ordered in a certain way. Finally, the digital approach of the corpus makes it possible to consider the CIFM as a very dense network of relationships (social, lexical, material), where all the inscriptions potentially respond to each other. If we may use a metaphor, we move here from a two- or three-dimensional approach to an *n*-dimensional/structural approach.[12] More often than not, this situation leads scholars to consider that the digital corpora are the final result of the research, whereas (in our opinion) they should rather appear as the starting point of new investigations. Of course, the acquisition/creation of a digital corpus is always the product of numerous past inquiries, the result of numerous operations and choices made by our predecessors. And experience shows that a corpus can be constantly reinvested (Magnani and Ingrand-Varenne, 2018). Its exploitation, for example, thanks to digital methods, generates new data that can be reinjected into the corpus and in turn operated—creating a kind of virtuous circle.

In the case of the CIFM-CBMA, this retro-development is possible because of the properties of textual data, i.e. named entities, and the restitution of the results of research carried out with the corpus and its metadata with different software, applications, or ad hoc scripts. We will briefly discuss three of these retro-developments here:

1. Semantic web, connectivity, and named entities;
2. Geographic information system (GIS) and automated extraction of new metadata;
3. Lemmatization and automatic language detection.

3.1 Semantic web, connectivity, and named entities

One of the characteristics of the epigraphic corpus is that 63% of it consists of funeral inscriptions and epitaphs. These memorial inscriptions contain, for the most part, the names of deceased persons and dating information, most often linked to the date of death. Once the texts were transcribed, these elements could have been detected automatically.

However, this information had already been extracted manually in printed edition and was included in the metadata set under the 'title' containing the date, type of inscription, and name of person, as in Fig. 1: '[c. 923] - Épitaphe d'Héribald'. There was therefore no individualization of people's names for this corpus. Using the resources of the semantic web—the Resource Description Framework (RDF) files—Pierre Brochard, via Python scripts and a French model of the spaCy library,[13] has automatically detected these named entities in the titles and created a prosopographic database, that can be cross-referenced with other prosopographic resources developed by historians or already present on the Internet (Wikidata, for example).[14] The aim is two-fold: to link resources together, through cross-formalization; but also to write historical studies, because these prosopographical data contained in the inscriptions have been used very little until now. This first enrichment of the corpus will also make it possible to search for existing and little-known links between the characters contained in the Burgundian charters and those included in the epigraphic documents. In the end, these additions open up little explored fields for documentary comparison.[15]

3.2 GIS and automated extraction of new metadata

Since the 1970s, the inscriptions are organized by CIFM's team according to the French districts, also called the '*départements*'. Although these '*départements*' are convenient for organizing the paper publication, they are very anachronistic for the Middle Ages and can mislead scientific analysis. For its part, the CBMA project usually uses the medieval ecclesiastical districts, that is to say the dioceses. Even if these entities varied during the Middle Ages, it was important to be able to propose a distribution of the corpus-based on medieval criteria since the whole medieval social structure was based on the

ecclesiastical spatial structure. But while the epigraphic corpus did not record the dioceses, it already contained the indication of the place of production, provenance, or conservation of the inscriptions. During our project, this geographical information has been automatically associated with the geolocation coordinates (WGS 84—EPSG 4326), resulting in the creation of a shapefile (.shp) that can be used in GIS software.[16]

From this file, several GIS analyses were made by Davide Gherdevich and projected on a map, where he recently improved the boundaries of the medieval Burgundian dioceses as part of the ANR Col&Mon (*Collégiales et Monastères*) project (Gherdevich, 2017). The operation carried out is technically quite simple but scientifically fundamental. From the place of each inscription in this map, a table was extracted with their situation in the corresponding diocese. The diocese of each inscription was thus able to be retro-injected from the maps into the metadata of the corpus. As there is no correspondence between present-day '*départements*' and medieval dioceses, places within a '*département*' may be located in different dioceses. For example, the localities located today in the '*département*' of Yonne were located in the former medieval dioceses of Sens or Auxerre. For the first time, inscriptions can be queried by their diocese. This method of automatically assigning dioceses had never been used before, at least to our knowledge, in medieval studies. Yet its interest seems central because, once again, these diocesan spatial entities were the basic framework for the action of the bishops—themselves being the central figures of this system, alongside the saints. At a practical level, this information is essential for the cross-exploitation of different types of CBMA sources, because it allows, for example, surveys on the local specificities of the lexicon, the circulation of formulas, the circulation of important characters, and their social reproduction strategies. If the enrichment of a database by metadata is a practice that is now more widespread in the digital humanities, the interest of our approach also lies in the fact that it is not static. On the contrary, we take advantage of the constitution of the corpus in 'layers' (records, editions, then new records), to add others and link them all—but without ever considering this digital work as finished. On the contrary, the corpus is open to additions and

transformations, which avoids falling back into the pitfalls of the paper format. This approach is obviously of interest in terms of historical analysis: the different medieval documents and corpora were once articulated by different links, chronological, material, and ideal. However, their reclassification as 'archives' during the 18th–20th centuries partially destructured these links, by proposing new relationships between documents, this time according to scientific, library, and archival criteria. Digital methods, by offering the possibility to rearticulate these documents, offer another way of looking at them, by making it possible to revive some of the links that were once woven between them—first of all because the corpus in itself makes the documents copresent and coconsultable. In this way, the complexity of a documentary material that has never been static, but that has constantly evolved, been incremented and revised, can be better rendered. Digital technology thus provides an opportunity to take into account documentation and scholarship simultaneously, which seems to us to be a fundamental challenge in current history—partly accentuated by the digitization of documents.

3.3 Lemmatization and automatic language detection

One of the other aims of the project was to propose a lemmatized corpus. However, the inscriptions present various difficulties: they are short, stylized, full of proper names and dates, but above all multilingual. Latin, Old French, and even Greek can be found in varying proportions. In the first place, inscriptions sometimes contain several languages (Latin can be manifested, for example, by the simple presence of an '*Amen*'). The first lemmatization experiments conducted by Nicolas Perreaux therefore aimed to refine the formalization of the corpus, to test existing parameters, and to propose possible automated processing methods. It should be noted that no medieval multilingual corpus had been lemmatized before this project. After the formalization of the corpus, the challenge was to observe to what extent the set of linguistic tags assigned by epigraphers/philologists to inscriptions[17] corresponded (or not) to the granularity resulting from an analysis of the texts by the computer whose parameters are established by language. In other words, it is a question of observing to what extent our algorithmic chain allows (1) to predict the languages used in a given

document; (2) to see to what degree of refinement (i.e. granularity) this process takes into account very subtle or punctual language switch (as in the case of the Latin lemma '*Amen*', already mentioned). Some natural language processing (NPL) libraries already offer similar functions in R or Python (e.g. spaCy), but they are poorly adapted to historical languages. In addition, we wanted to control the entire chain of lemmatization—which allowed us to better understand the pitfalls and difficulties encountered by the lemmatizer of the CIFM.

Overall good results were obtained with the Latin inscriptions tagged with the TreeTagger software[18] and the parameters for medieval Latin OMNIA (dir. Alain Guerreau) (Bon, 2009, 2010, 2011),[19] using the tokenizer developed for the CEMA corpus (Perreaux, 2015). However, these parameters frequently face difficulties when dealing with named entities, which they still have trouble detecting due to a lack of example within the training set. In order to overcome this problem, we chose to automatically assign the value NAM (=NAMed entity) to words beginning with a capital letter that were not at the beginning of the sentence. This combination of different methods for Latin document (classical training on known parameters, then ad hoc post-processing) allowed us to improve the results of the lemmatization—which is essential for future historical analyses, especially semantic ones. Concerning inscriptions in Old French, two sets of tools were used: the parameters for TreeTagger from the work on the New Corpus of Amsterdam (Achim Stein)[20] and the Pandora lemmatizer (Mike Kestemont, Jean-Baptiste Camps, Thibault Clerice) (Kestemont *et al.*, 2017; Manjavacas *et al.*, 2019),[21] which require different adaptations of the tokenizer, with a slightly better recognition by the first. In particular, the segmentation of the texts, necessary before any lemmatization, required the development of new scripts in the case of a lemmatization via TreeTagger. The syntactic structures specific to Old French obviously require the use of its own rules, for example, with regard to apostrophes—which must be separated from the terms they precede.

At the end of the process, one of the possible solutions to lemmatize inscriptions and overcome the difficulties posed by the different degrees of

linguistic mixing would be to let lemmatizers separate languages by comparing the number of unidentified terms (*unknows*) in order to retain, in the end, only the most 'effective' version. To go further, one could even consider letting an algorithm decide, sentence by sentence, what is the language of the sequence, and thus go down to more refined levels than the document as a whole.

Other automatic recognition operations are currently being implemented in the other CBMA components, but have already been tested in CEMA (a corpus of European charters, developed by Nicolas Perreaux, in conjunction with other programs on which the corpus is based) (Perreaux, 2015): in particular, automatic detection of duplicates or attempts to automatically tag documentary types (bullae, diplomas) by machine learning (SVM and neural networks). These ongoing enhancements are not just about creating metadata. Each new field could be the subject of specific scientific investigations, both historiographical (for example, around the question of duplicates, and therefore multiedited acts), and historical in the strict sense (with, for example, the question of the chrono-geographical distribution of textual genres in the Middle Ages, which should be taken up as a whole).

4 Conclusion

Thanks to the various operations carried out, the Burgundian epigraphic corpus is the first multilingual lemmatized, geolocalized, and freely available medieval corpus on the web. In that perspective, we would like to stress that in order to link corpora, not only between them, but also to improve their 'citability' and analytical potential, it is necessary to generate recent and reliable metadata, which is what digital methods allow. This 'quest for metadata', its semiautomated production and integration, seems to be at the heart of many current digital humanities projects. Indeed, academics have realized that while bringing together documents from different origins, languages, and chronologies was interesting and could change our perception of history, the juxtaposition of documents without elements to make the link between them was not only counterproductive but also hazardous. Methods for semiautomated generation of metadata for efficient querying of medieval

documentary collections are thus one of the goals pursued by Cosme[2] and the teams/projects brought together in the consortium. By generating new data/metadata on the data, by crosslinking them, we create new ways of approaching the documents and integrating them into scientific analyses. In this way, we aim to take full advantage of a dynamic in which the acquisition, documentation, and exploitation of corpora are not successive stages to be connected but coactive operations that are constantly reinvested and refueled.[22]

References

Bon B.(2009). OMNIA – Outils et Méthodes Numériques pour l'Interrogation et l'Analyse des textes médiolatins. *Bulletin du centre d'études médiévales d'Auxerre - BUCEMA*, 13: 291-2. http://journals.openedition.org/cem/11086 (accessed 10 October 2019).

Bon B.(2010). OMNIA: outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (2). *Bulletin du centre d'études médiévales d'Auxerre - BUCEMA*, 14: 251–2 http://journals.openedition.org/cem/11566 (accessed 10 October 2019).

Bon B. (2011). OMNIA: outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (3). *Bulletin du centre d'études médiévales d'Auxerre - BUCEMA*, 15: 333–4 http://journals.openedition.org/cem/12015 (accessed 10 October 2019).

Favreau R. (1982). Le corpus des inscriptions de la France médiévale. InKloos R. M. (ed),*Fachtagung für lateinische Epigraphik des Mittelalters und der Neuzeit*.Kallmünz:M. Lassleben, pp.61–72

Favreau R., Michaud J., Mora B.(1997). *Jura, Nièvre, Saône-et-Loire. Corpus des inscriptions de la France médiévale 19*. Paris: CNRS Editions.

Favreau R., Michaud J., Mora B. (1999). *Côte-d'Or. Corpus des inscriptions de la France médiévale 20*. Paris: CNRS Editions

Favreau R., Michaud J. (2000). *Yonne. Corpus des inscriptions de la France médiévale 21*. Paris: CNRS Editions

Gherdevich D. (2017). Les limites des diocèses au Moyen Âge: sources historiques et outils d'interprétation SIG. Carnet de recherche COL&MON. https://colemon.hypotheses.org/21 (accessed 11 December 2020).

Guerreau A. (2001). *L'avenir d'un passé incertain. Quelle histoire du Moyen Âge au XXIe siècle*. Paris: Seuil

Guerreau A. (2011). Pour un corpus de textes latins en ligne. *Bulletin du centre d'études médiévales d'Auxerre—BUCEMA—Collection CBMA - Les journées d'études*. http://journals.openedition.org/cem/11787 (accessed 10 October 2019).

Lévi-Strauss C. (1964–71). *Mythologiques*, 4 vols. Paris: Plon

Kestemont M., de Pauw G., van Nie R., Daelemans W. (2017). Lemmatization for variation-rich languages using deep learnin. *Digital Scholarship in the Humanities*, 32(4): 797–815

Magnani E. (2019a). Des chartae au corpus: la plateforme des CBMA - Chartae/Corpus Burgundiae Medii Aevi. In Balouzat-Loubet C. (ed), *Digitizing Medieval Sources. Challanges and Methodologies. L'édition en ligne de documents d'archives médiévaux. Enjeux, méthodologie et défis*. Turnhout: Brepols, pp. 57–67

Magnani E. (2019b). Lemmes: un groupe de travail sur les outils de lemmatisation et les corpus de textes médiévaux lemmatisés. *Archivum Latinitatis Medii Aevi - ALMA*, 76(2): 340–4

Magnani E., Ingrand-Varenne E. (2018). Le corpus épigraphique bourguignon (VIII^{e}-XV^{e} siècle). Des catalogues aux applications numériques. *Bulletin du centre d'études médiévales d'Auxerre—BUCEMA—Collection CBMA - Les journées d'études*. http://journals.openedition.org/cem/15591 (accessed 10 October 2019).

Manjavacas E., Kadar A., Kestemont M. (2019). Improving lemmatization of non-standard languages with joint learning. NAACL-HTL 2019 - North American Chapter of the Assosiation for Computational Linguistics: Human Language Technologies, arXiv: 1903.06939v1 (accessed 11 December 2020).

Perreaux N. (2014). De l'accumulation à l'exploitation? Expériences et propositions pour l'indexation et l'utilisation des bases de données diplomatiques,In Ambrosio A., Barret S., Vogeler G. (eds), *DigitalDiplomatics. The Computer as a Tool for the Diplomatist*? Köln-Weimar-Wien: Böhlau Verglag, pp. 187–210

Perreaux N. (2015). L'écriture du monde (I). Les chartes et les édifices comme vecteurs de la dynamique sociale dans l'Europe médiévale (VIIe-milieu du XIVe siècle). *Bulletin du centre d'études médiévales d'Auxerre - BUCEMA*, 19(2). http://journals.openedition.org/cem/14264. DOI: 10.4000/cem.14264 (accessed 10 October 2019).

Footnotes

1
Cosme[2] is directed by Paul Bertrand: https://cosme.hypotheses.org/ (accessed 10 October 2019). TGIR is a 'Très Grande Infrastructure de Recherche' (Very Large Research Infrastructure) of the French National Centre for Scientific Research (CNRS) https://www.huma-num.fr/ (accessed 10 October 2019).

2
The consortium is organized in five different working groups, working on: (1) The alignment of manuscript shelfmarks, (2) Lemmas, (3) Named Entities, (4) Dates and Formulas, (5). 'Values' and Measurements.

3
http://titulus.huma-num.fr (accessed 10 October 2019).

4
http://www.cbma-project.eu (accessed 10 October 2019).

5
https://cescm.labo.univ-poitiers.fr/ (accessed 10 October 2019).
6
https://www.persee.fr/collection/cifm (accessed 10 October 2019).
7
The team includes Mathieu Beaud, Pierre Brochard, Davide Gherdevich, Estelle Ingrand-Varenne, Eliana Magnani, Aurore Menudier, Nicolas Perreaux, Coraline Rey.
8
The printed volumes of the CIFM cover a period from the 8th to the mid-13th century. However, several inscriptions from the late 13th, 14th, and 15th centuries have been catalogued in a paper catalog. These 'unpublished' inscriptions have also been included in the digital corpus produced in 2018–19. Specifically, 471 inscriptions were taken from the printed OCR'd volumes and 947 from the paper catalogue, in a total of 1,418 documentary units.
9
http://www.cbma-project.eu/%C3%A9ditions/textes_epigraphiques.html (accessed 10 October 2019); https://gitlab.huma-num.fr/lamop/cbma-epigraphie (accessed 10 October 2019).
10
http://philologic.lamop.fr/epigraphie/ (accessed 10 October 2019).
11
http://textometrie.ens-lyon.fr/ (accessed 10 October 2019); https://www.sketchengine.eu/ (accessed 18 October 2019).
12
In a way, the problem of the corpus has largely been addressed by structuralist anthropologists, who have sought to bring order to series of narratives. Although these were essentially qualitative approaches, the logic seems to us similar (see in particular Lévi-Strauss, 1964–71).
13
https://spacy.io/ (accessed 23 September 2020).
14
Approximately 20% of the automatically detected named entities had, however, to be corrected manually.

15
The team is still developing research in this field.
16
Two GIS software packages have been used, QGIS (https://www.qgis.org, accessed 23 September 2020) and ArcGIS (https://www.arcgis.com/index.html, accessed 23 September 2020).
17
Some example of the tagset that was used: 'latin' = Latin, 'français' = French, 'lat_français' = predominance of Latin, 'fr_latin' = predominance of French, etc.
18
http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (accessed 10 October 2019).
19
http://glossaria.eu/ (accessed 10 October 2019).
20
https://sites.google.com/site/achimstein/research/resources (accessed 10 October 2019).
21
https://github.com/hipster-philology/pandora (accessed 10 October 2019).
22
We thank Charles West for proofreading our English text. We also thank Kouky Fianu and the anonymous reviewer for their very useful comments.