Preserving and enhancing cultural heritage: the Digest project

Alessandra Cinini[1], Paola Marongiu[2], Eva Sassolini[3] & Monica Monachini[4]
[1]Istituto di Linguistica Computazionale "A. Zampolli", (CNR - Pisa), Italia, alessandra.cinini@cnr.it
[2]Istituto di Linguistica Computazionale "A. Zampolli", (CNR - Pisa), Italia, paola.marongiu@cnr.it
[3]Istituto di Linguistica Computazionale "A. Zampolli", (CNR - Pisa), Italia, eva.sassolini@cnr.it
[4]Istituto di Linguistica Computazionale "A. Zampolli", (CNR - Pisa), Italia, monica.monachini@cnr.it

ABSTRACT[1] (ENGLISH)
This paper aims to describe and analyze the activities related to the preservation and valorization of textual corpora of cultural-historical value, produced over a long period of time. We intend to share with the scientific community the problems related to the advancement of technologies used for data creation/processing, as well as the issues related to the transition from proprietary to standard formats to enable data sharing and interoperability. Finally, we will outline the complex transition to open data paradigms and the necessary steps for migrating data into international research infrastructures. Specifically, we will describe the translation process of Justinian's Digest, emphasizing how this has evolved alongside technological progress. Our contribution will offer the DH community a point of view on the evolution of the digitization/computerization of large textual resources, based on a real use-case. To illustrate this, we will examine how specialized translation from Latin to Italian can be enhanced using textual analysis tools. Furthermore, we will describe the transformation of both the original and translated texts into a parallel bilingual corpus; its conversion into XML TEI format; the steps taken for depositing the data in the CLARIN research infrastructure.
Keywords: texts corpora; standard TEI format; digital preservation

ABSTRACT (ITALIANO)
Valorizzare e custodire il patrimonio culturale: alla scoperta del progetto Digesto. Il contributo vuole discutere delle attività connesse alla salvaguardia e alla valorizzazione di corpora testuali di valore storico-culturale prodotti in un lungo arco temporale. Analizzeremo le problematiche legate al progresso delle tecnologie utilizzate per la creazione/elaborazione dei dati che sono mutate nel tempo, nonché l'arduo passaggio dai formati proprietari a quelli standard per una migliore condivisione e interoperabilità dei dati. Infine, descriveremo la non facile transizione verso paradigmi di dati aperti e i passaggi necessari per migrare i dati verso infrastrutture di ricerca internazionali.
In particolare, descriveremo le fasi del progetto di traduzione del Digesto di Giustiniano che hanno richiesto il supporto tecnologico del nostro gruppo di ricerca, e come questo contributo è cambiato di pari passo con il progresso tecnologico. Vogliamo offrire alla comunità DH un punto di vista reale sull'evoluzione della digitalizzazione/informatizzazione di grandi risorse testuali. Nel caso specifico, discuteremo del supporto alla traduzione specializzata dal latino all'italiano con strumenti di analisi testuale, e della trasformazione dei testi originali e tradotti in un corpus bilingue parallelo; di come questo è stato convertito in formato XML TEI; infine, delle operazioni necessarie al deposito dei dati all'interno dell'infrastruttura di ricerca CLARIN.
Parole chiave: corpora testuali; standard TEI; preservazione digitale

1. INTRODUCTION
The activities described in this paper are part of the TESTO project, which currently focuses on the development of a parallel Latin-Italian corpus of Justinian's Digest. The translation and digitization of the Digest began in the late 1990s and advanced gradually over an extended period (Schipani 1994[1996], 2005; Schipani & Lantella 2005a, 2005b, 2007, 2011). Initially, technological contributions were limited to digitization efforts. However, by the time the third book was completed, the need for more structured translation support had become evident. The fragmentary nature of the technological support provided led to challenges comparable to those encountered in many recent projects involving the digitization of archives encoded in outdated formats (Sassolini et al., 2014). The development of the project can be divided into three main phases, each closely tied to technological advancements. In the first phase, IT efforts focused on developing prototypes and stand-alone programs for acquiring the bilingual corpus and

[1] As part of a joint elaboration, Eva Sassolini specifically oversaw the drafting of §§ 1 and 5, Alessandra Cinini of §§ 3, while Paola Marongiu focused on §§ 4 and 4.1. Paragraph 2 is a joint work. Monica Monachini participated as the project manager.

enabling textual searches. These early tools were created using then-available development environments and software systems, and were mainly developed *ad hoc*, later revealing compatibility issues with the new environments. In the second phase, the growing diffusion of the Internet prompted the transition of translation support tools to web-based platforms. This shift mitigated the issues related to the evolution of software environments and fostered collaborative and shared working practices. The third and ongoing phase addresses a new set of emerging challenges, primarily concerning data protection, standardization and interoperability of representation formats, and efficient data sharing strategies.

The creation of the bilingual parallel corpus of the Digest represents a peculiar initiative in the landscape of resources for legal studies and historical linguistics. Various corpora have been developed for legal studies, both mono- and multilingual. The CLARIN infrastructure hosts a "Legal Corpora" Resource Family,[2] which includes resources such as Europarl (Koehn 2005), a collection of European Parliament proceedings in 21 languages, and the JRC-Acquis corpus (Steinberger et al. 2006), a sentence-aligned parallel corpus in 22 languages. However, to our knowledge, there remains a notable absence of multilingual corpora specifically focused on historical legal texts. On the other hand, significant efforts have been made in the digitization of historical texts. Among other initiatives, we can mention the Perseus Digital Library (Crane 1987; Crane et al. 2006), which provides a large digital collection of Ancient Greek and Latin texts; the five treebanks for Latin, now aligned with the Universal Dependencies project annotation guidelines (Haug and Jøhndal 2008; Bamman and Crane 2011; Passarotti 2019; Cecchini et al. 2020; Cecchini et al. 2020); Corpus Corporum, another freely accessible digital collection of Latin texts (Roelli 2014); Musisque Deoque (Boschetti, Del Grosso & Spinazzè 2021), a digital collection of Latin poetry. Concerning parallel/comparable corpora of historical texts, it is worth mentioning the work by Yousef and Berti (2015), who automatically aligned Ancient Greek and Latin texts with their English translations at the sentence and word level. Our work on the Digest thus brings together the challenges and methodological considerations inherent in two research domains: the development of specialized parallel corpora and the encoding and alignment of historical texts. In this paper, we will describe the activities carried out in constructing the parallel corpus of the Digest, including the text alignment procedures (section 2), the rationale behind the selection and use of the XML-TEI encoding format (section 3), and the steps that will be taken to publish the corpus in compliance with Open Access principles (section 4). Section 5 outlines our plans for future developments in our work.

2. TEXT CORPUS OF CONCORDANCES

The Justinian's Digest, together with the *Codex Iustiniani*, the *Iustiniani Institutiones* and the *Novellae Constitutiones*, is part of the *Corpus Iuris Civilis*, a collection of Roman legal texts compiled under the order of Emperor Justinian the Great (527–566 CE) between 529 CE and 534 CE (Banchich et al. 2015; Dingledy, 2016). Justinian inherited from his predecessor Justin I a weak empire: the West Empire was now under the influence of Germanic tribes and Justinian saw in the project of the *Corpus Iuris Civilis* an opportunity to unite the empire under Christianity. The Digest represents the most important part of the *Corpus Iuris Civilis*, as it offers the wealthiest collection of Roman legal texts. The Digest encompasses selected excerpts from 1528 writings produced by 39 Roman jurists, starting from the Twelve Tables (449 BCE) until the jurist Hermogenianus (350 CE), for a total of 800 years of Roman legal history (Dingledy, 2016: 5; Ribary & McGillivray, 2020). The jurists' writings are organized in 50 books, divided into five groups depending on the topic: Public Law (Book I); Private Law (Books II–XLVII); Criminal Law (Book XLVIII); Appellate Procedure and Treasury (Book XLIX); Municipal Law, Specialized Law, and Definitions (Book L). The books are divided into 432 titles. Each title is divided into laws, and laws into paragraphs. Each law is preceded by an inscription stating the name of the jurist who wrote it, and the title and volume of the book from which the excerpt was taken (Banchich et al. 2015: 5). Studying the Digest nowadays is for jurists functional to the understanding and development of a system of principles on which each state's law is based, regardless of their specific juridical systems (Schipani 2005: xxvii).
For all the above, the Digest translation project concerns a specialized field (the legal domain), with all the challenges associated with the translation of specialist concepts/terms. The translators' first need was to consult past translations to facilitate their work and ensure coherent use of legal terminology throughout the text. Therefore, when our institute was involved, our main goal was to support the translators by providing them with a bilingual concordances query system, to be used as translation memories. To do this, the Latin texts of the Digest were automatically aligned with their Italian translations by using a

[2] https://www.clarin.eu/resource-families/legal-corpora

statistical method for their synchronization, refined and guided by a network of linguistic anchors/items (Marinai et al. 1992). The semi-automatic alignment process relies on the morphological engines of the Italian and Latin languages, as well as on a bilingual dictionary, to establish a network of links between source and target text units within a predefined textual window. However, this process often breaks down when the structure of the translation diverges significantly from that of the original Latin text (Simard & Plamondon, 1998). To ensure accuracy, minimum thresholds are set for the number of detected equivalent translation units; when these thresholds are not met, manual intervention is required. This requirement has a significant impact on alignment times, which may increase considerably. Furthermore, an interface (or tool) is available to review the aligned texts, allowing for targeted adjustments in specific sections, typically used for aligning Ancient Greek and Italian.

Then, we transformed the results into bilingual parallel concordances accessible with the DBT (Data Base Testuale) textual analysis engine.[3] The tool, initially conceived as a PC software, was later transformed into a client-server application. The primary objective of the concordances corpus was to offer translators a resource where they could search for specific words or family of words in Latin and check their Italian correspondents in the previous translations.

Web development extended the range of users for these resources to scholars in other research domains. This required the implementation of new functionalities and a new interface for lower specific users, e.g. historical linguists interested in the study of non-literary Latin, semantic shift and the development of specialized lexicon.

As part of the project, we also plan to develop a bilingual digital glossary of Latin legal terms and their specialized Italian translations, drawn from the bilingual corpus. The creation of the glossary is still ongoing. We have lemmatized the books translated so far by using Latinpipe (Straka, Straková, & Gamba 2024) for the Latin texts, and Stanza (Qi et al. 2020) for the Italian translations. We are currently working on the evaluation of the lemmatization results and on devising the best strategies for the extraction of specialized/domain-specific Latin terms to be included in the glossary.

3. XML TEI FORMAT: WHY AND HOW

The process of computerizing data is a technological ecosystem and as such has specific characteristics. Nowadays, the objectives of a researcher working on this type of process cannot be exclusively oriented towards improving performance. In fact, the development of user support tools requires a different approach that takes the context into account. The aim is to make research more responsive to changes that are not only technological but also social and cultural. Achieving these goals requires continuous updates aligned with recent scientific developments, dealing with issues concerning the value of data and their dissemination. It is important for us to find new, multidisciplinary and adaptive approaches that can compete with the new challenges posed by the changing scenario. A computational linguist will work not only to improve the performance of the tools, but also their flexibility, so that they can be designed and implemented in an integrated way and respond to the varied users' information needs.

In addition to adapting/improving the tools already developed, we mapped in XML TEI the parallel corpus created by the bilingual concordances and all derived resources, with the aim of disseminating the project results, preserving them from obsolescence and making the data interoperable. We have chosen to deposit our resource in the CLARIN resource infrastructure due to its robust capabilities and commitment to expanding access to linguistic data.

As a first step, we studied a representation model in TEI. Based on the chosen model, we implemented the automatic mapping process. This resulted in the output of separate files for each book, available in both the Italian and Latin language versions. Each book has a tree structure divided into sections that has been encoded with the <DIV> tag. The ordering and organization follow the Digest's textual units in their original printed form. As shown in Figure 1, each section's tag is assigned attributes that specify its type: title; fragment; principium/paragraph; an identifier for text alignment in the two languages. More specifically, Figure 1 shows how an inscription is encoded in the TEI model.

[3] Text analysis system developed by Eugenio Picchi at CNR-ILC with various software configuration possibilities, including the contrastive consultation of digital corpora.

```
<div type="frammento" resp="POMPONIUS" corresp="#libro nono ad Sabinum">
<p>
<seg xml:id="D.18.I.13_L" ana="D.18.I.13">IDEM libro nono ad Sabinum. Sed si servo meo vel ei cui mandavero
vendas sciens fugitivum illo ignorante, me sciente, non teneri te ex empto verum est.</seg>
</p>
</div>
<div type="frammento" resp="ULPIANUS" corresp="#libro vicesimo octavo ad Sabinum">
<p>
<seg xml:id="D.18.I.14_L" ana="D.18.I.14">ULPIANUS libro vicesimo octavo ad Sabinum. Quid tamen dicemus, si in
materia et qualitate ambo errarent? ut puta si et ego me vendere aurum putarem et tu emere, cum aes esset? ut puta
coheredes viriolam, quae aurea dicebatur, pretio exquisito uni heredi vendidissent eaque inventa esset magna ex
parte aenea? venditionem esse constat ideo, quia auri aliquid habuit. nam si inauratum aliquid sit, licet ego aureum
putem, valet venditio: si autem aes pro auro veneat, non valet.</seg>
</p>
</div>
```

Figure 1. Encoding of *inscriptiones*

Regarding alignment procedures, some considerations can be made concerning the linguistic anchors that connect specific words with their potential translations. These elements, initially extracted to improve the contrastive query of bilingual concordances, are generated by automatic software procedures with a significant rate of 'noise'. Since in the context of specialized translation a literal approach to the source text is considered insufficient, the translator's intent must be interpretative. For this reason, the translated text may in some cases contain integrations/additions of various kinds (explanatory, stylistic, morpho-syntactic, etc.) that make the alignment less accurate. In fact, the translator leaves out the formal correspondence between the two languages but interprets and reformulates what is necessary to achieve an equivalence of meaning (Angelucci, 2024). In our case, this scenario requires synchronization of the texts at the fragment level rather than the word level, while simultaneously opening the possibility of extending the alignment to other languages into which the Digest has been or is currently being translated. For example, the translation of the Digest in Chinese has been in progress for years (Schipani, 2010; Luo & Colangelo, 2013; Li & Colangelo, 2016). Furthermore, we are studying methods to analyze the linguistic anchors and codify them in resources to be shared.
The textual parts added by the translators, due to their peculiarity, have been encoded with a specific TEI tag (<ADD>, as shown in Figure 2) and linked to an external resource that describes and documents with examples the various types of textual additions.

```
<seg xml:id="D.01.II.2.6_I" ana="D.01.II.2.6">
        In seguito, pressappoco nel medesimo periodo, sulla base di queste leggi furono composte le azioni
<add>processuali</add> con le quali gli uomini contendessero tra loro; si volle che tali azioni fossero certe e solenni,
affinché il popolo non ne istituisse come volesse. Questa parte del diritto viene chiamata: "azioni di legge", cioè
"azioni legittime". E così, pressoché nel medesimo periodo, nacquero questi tre diritti: le Leggi delle Dodici Tavole; da
queste cominciò a fluire il diritto civile; sulla base delle medesime Tavole vennero composte le azioni di legge. Di tutto
ciò, tuttavia, sia la scienza dell'interpretare sia le azioni erano nell'ambito di competenza del collegio dei pontefici, tra i
quali si statuiva chi in ciascun anno fosse preposto a <add>rispondere alle domande de</add>i privati; e il popolo si
avvalse di questa consuetudine pressappoco per cento anni.
</seg>
```

Figure 2. Example of textual additions introduced by the translators

Finally, we worked on reconstructing the history of the project, the actors, methods and technologies used. The information reconstructed is included in the header of the TEI corpus, but also in the Data Management Plan, required for depositing the resource in CLARIN, which we will discuss in detail in the next paragraph. Once the translation is completed, it will be possible to specialize the TEI model and annotate the texts at a lower level.

4. DATA PRESERVATION: MIGRATING THE DATA INTO THE CLARIN INFRASTRUCTURE

The original purpose of the project was to provide a valuable tool to be used as a translation memory, to ensure translation normalization. Since then, the goal of the project has gradually broadened to ensure data sharing with the scientific community. For this reason and to guarantee the preservation of data and resources created in the long history of the project, we decided to convert the aligned bilingual corpus into

the XML TEI interoperable format (§3) and to share it with the humanities and social sciences communities (DHs) via the CLARIN infrastructure, all by adhering to the best practices underlying the FAIR[4] principles. Within this distributed digital infrastructure, the individual centers are the network nodes. For our project we addressed ILC4CLARIN, a type B CLARIN center, which is part of CLARIN-IT, the Italian CLARIN ERIC node.[5]
By doing so, we aim not only to preserve our developed resources but also to align them to the CLARIN framework and facilitate interoperability with linguistic services offered by the infrastructure. This strategic decision aligns with CLARIN's mission to enhance data discoverability and facilitate interoperability across linguistic and cultural studies. Additionally, the CLARIN infrastructure offers us the opportunity to publish our textual data within the evolving landscape of Linked Open Data (LOD), which is an important aspect of contemporary digital scholarship. In particular, CLARIN-IT is developing a robust Linked Open Data (LOD) platform, designed to enhance data sharing and collaborative research in the context of the H2IOSC project (Horizon 2020 International Open Science Cloud).[6]
The LOD platform aims to integrate diverse datasets from various fields into a cohesive framework, thereby promoting an interconnected approach to research practices. By utilizing Linked Open Data principles, the platform provides a means to interlink datasets, allowing researchers to navigate through related information seamlessly, regardless of the original source or format. This not only enriches the context of individual datasets but also facilitates interdisciplinary research by bridging gaps between disparate fields. By positioning our data in the 'cloud' of linguistic Linked Open Data in the future, we can ensure that it is not only accessible but also interconnected with other resources. This interlinking will enable richer, more complex queries and analyses, thus enhancing the overall utility of our data for the community.

4.1. The data management plan

As part of the process for ensuring adherence to the best practices in terms of Open Science and data FAIRification, we have included in the project activities the creation and continuous update of a Data Management Plan (DMP).
A DMP is a document that follows the life cycle of the data, from data collection, organization, documentation and quality control to data sharing, archiving and preservation (Michener 2015). The DMP is supposed to be a 'living document' that evolves together with the advancement of the project. Drafting the first version of the DMP at the very beginning of the project has been a fundamental step to ensure a structured view of our data cycle. It allowed us to think ahead of the nature of our data, including e.g. copyright and privacy-related issues, expected outcomes, overall costs throughout the project's lifespan, best practices and decisions in terms of licenses that will be applied to the data.
In our DMP we included, among other types of information, legal implications of the data creation/annotation process; long-term data preservation (type of infrastructure or repository where our data will be archived); actors involved in the data life cycle; expected costs of carrying out the project activities (Giglia 2023).
To fill out the DMP for the TESTO project we relied on ARGOS,[7] a platform powered by OpenAIRE[8] that allows for creating, modifying and archiving machine-actionable DMPs. The completion of the first version of our DMP posed various challenges, due to the very peculiar nature of the data and to the history of the TESTO project.
As mentioned earlier, the project started in the late 90s and is still underway to this day. There have been various changes throughout the years, including the project funding, directors, partners, assistants (especially translators) and technologies used to process the data. Moreover, the project as it is conceived today involves different university partners, together with our institute. Delivering the DMP proved more challenging than initially anticipated, as it required reconstructing the project's history to ensure proper

[4] https://www.go-fair.org/fair-principles/

[5] https://ilc4clarin.ilc.cnr.it/, hosted by our Institute.

[6] https://www.h2iosc.cnr.it/

[7] ARGOS offers a simplified, guided step-by-step process to complete the DMP and makes sure the user has entered all information relevant to their specific project data (https://argos.openaire.eu/splash/about/how-it-works.html).

[8] "OpenAIRE is a Non-Profit Partnership of 50 organisations, established in 2018 as a legal entity, OpenAIRE A.M.K.E, to ensure a permanent open scholarly communication infrastructure to support European research" (from https://www.openaire.eu/about).

credit was given to all former directors and collaborators. The information related to the project history, including funding, editors of the published volumes, translators for each paragraph in the published volumes, and current partners of the project, constitutes the metadata for the bilingual corpus in XML TEI, and will be accessible as an external resource linked from the header of the TEI files. The DMP details the type of metadata and information included in the TEI header.

The long history of the project has also affected the type of technologies used for data processing and analysis. As mentioned in section 2, the alignment of the Latin-Italian texts was carried out with DBT, a proprietary text analysis system that also enables the online contrastive consultation of the bilingual corpus. In the perspective of sharing the data with the research community, and adhering to the FAIR principles, we decided to convert the data and resources generated through DBT into the XML TEI format. The same workflow will be applied to the books that will be published until completion of the translation process. To ensure full transparency on the workflow and data processing, we decided to include this information in the DMP, providing details about the type of software used and the process developed to obtain the TEI files.

Additional issues stem from the fact that the project is still ongoing, such as determining the appropriate licensing for the texts and finalizing the bilingual glossary. One of the milestones of the project is a first deposit of the data, i.e., the bilingual corpus in TEI format, on the platform ILC4CLARIN. However, the translation of the Digest is still incomplete, with books 36–50 yet to be translated. Additionally, the TEI model developed from previous translations has not yet been tested against the new material; it's possible that the Latin or Italian texts in the remaining books contain types of information not present in earlier ones, which may require different treatment. Therefore, we decided to apply a restricted license to this first deposit, which only allows the people working on the project to access the data. We explained the reasons for our decision in the first version of our DMP.

Finally, as mentioned earlier, one of the outputs envisaged for the completion of the project is a bilingual glossary of Roman legal terminology. However, as the translation is still ongoing, this output can only exist in a provisional form. Therefore, we decided to announce the publication of the bilingual glossary as a future outcome of the project, leaving the details for the updated version of the DMP.

5. FUTURE PERSPECTIVES

The project is still underway, and finishing the translation remains the primary goal. We have defined the TEI model of the corpus but, as mentioned in § 3, various options exist for the representation of textual data in this format. It is our intention to study them further. The nearest future perspective is to archive all 50 books in Latin-Italian into the CLARIN infrastructure under an open license.

We also plan to experiment with integrating services and resources into centralized platforms, such as the CLARIN Switchboard. Further testing will be devoted to the publication of the project data as Linguistic Linked Open Data, incorporating them into the resource cloud for historical languages. This step will broaden the potential for sharing, hopefully fostering synergies with similar resources, projects and initiatives, e.g. by bridging translations of the Digest into other languages. This scenario is a concrete possibility, especially considering the ongoing translation of the Digest into Chinese.

ACKNOWLEDGEMENTS

Research Project PRIN 2022 PNRR P20224NJLK – SH2 – TESTO "Translating, Encoding, Sharing The Origins". From the Littera Florentina to an open-access Italian translation of Justinian's Digest". Piano Nazionale Ripresa e Resilienza (PNRR) M4C2 – Investimento 1.1 "Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN)- funded by the European Union – NextGeneration EU" – CUP B53D23032480001.

REFERENCES

Angelucci, M. (2024). Computabilità e traduzione interlinguistica: considerazioni sull'impiego dell'Intelligenza Artificiale in ambito traduttivo. *DigItalia*, 19(1), 183–197.

Bamman, D. & Crane, G. (2011). The Ancient Greek and Latin Dependency Treebanks. *Language technology for cultural heritage: Selected papers from the LaTeCH Workshop Series*. Sporleder, C., Bosch, A., Zervanou, K. (Eds.), Berlin, Heidelberg: Springer. 79–98.

Banchich, T. M., Marenbon, J., & Reid, C. J. (2015). The Revival of Roman Law and Canon Law. *A Treatise of Legal Philosophy and General Jurisprudence. Volume 6: A History of the Philosophy of Law from the Ancient Greeks to the Scholastics*. Miller Jr, F. D., & Biondi, C. A. (Eds.), Dordrecht: Springer. 251–265.

Boschetti, F., Del Grosso, A. M., & Spinazzè, L. (2021). La galassia Musisque Deoque: storia e prospettive. *ANTICHISTICA. ARCHEOLOGIA*, 32, 405–419.

Branco, A., Eskevich, M., Frontini, F., Hajič, J., Hinrichs, E., de Jong, F., Kamocki, P., König, A., Lindén, K., Navarretta, C., Piasecki, M., Piperidis, S., Pitkänen, O., Simov, K., Skadina, I., Trippel, T., Witt, A., & Zinn, C. (2023). The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond. *Language Resources & Evaluation*, 1–32. DOI: 10.1007/s10579-023-09658-z

Cecchini, F. M., Korkiakangas, T. & Passarotti, M. (2020). A New Latin Treebank for Universal Dependencies: Charters between Ancient Latin and Romance Languages. *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France 11–16 May 2020. Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., & Piperidis, S. (Eds.). 933–942.

Cecchini, F. M., Sprugnoli, R., Moretti, G., & Passarotti, M. (2020). UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works. *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020,* Bologna, Italy, March 1–3, 2021. Dell'Orletta, F., Monti, J. & F. Tamburini (Eds.). 1–7.

Dingledy, F. W. (2016). The Corpus Juris Civilis: A Guide to its History and Use. *Legal Reference Services Quarterly,* 35(4), 231–255. DOI: 10.1080/0270319X.2016.1239484

Giglia, E. (2023). Open Science cafè - Come scrivere un Data Management Plan (DMP). *Zenodo*. DOI: 10.5281/zenodo.7526256

Haug, D. & Jøhndal, M. (2008). Creating a Parallel Treebank of the Old Indo-European Bible Translations. *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH)*, Prague, Czech Republic, June 28 2007. Sporleder, C. & Ribarov, K. (Eds.). 27–34.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *Machine Translation Summit*. Phuket: Thailand. http://people.csail.mit.edu/koehn/publications/europarl/

Li, C., & Colangelo, L. (2016). *Luoma fa Minfa Daquan fanyi xilie: Xueshouo Huizuan di shiliu juan. Dixiao yu jituo = Iustiniani Augusti Digesta seu Pandectae, libro XVI. De Compensationibus et de deposito*. Beijing: Zhongguo Zhengfa Daxue Chubanshe.

Luo, G., & Colangelo, L. (2013). *Luoma fa Minfa Daquan fanyi xilie: Xueshouo Huizuan di' erishiasan juan. Hunyin yu jiazi = Iustiniani Augusti Digesta seu Pandectae, libro XXIII. De Matrimonio et dote*. Beijing: Zhongguo Zhengfa Daxue Chubanshe.

Marinai E., Peters C., & Picchi E. (1992). A Project for Bilingual Reference Corpora. *Acta Linguistica Hungarica,* 41, 1/4, 191–204. http://www.jstor.org/stable/44308294

Michener, W. K. (2015). Ten simple rules for creating a good data management plan. *PLoS computational biology*, 11(10), e1004525.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* Celikyilmaz, A. & Wen, T-H. (Eds.). 101–108.

Passarotti, M. (2019). The Project of the Index Thomisticus Treebank. *Digital Classical Philology*, 10, 299–320.

Ribary, M., & McGillivray, B. (2020). A Corpus Approach to Roman Law Based on Justinian's Digest. *Informatics*, 7(4), 44. https://doi.org/10.3390/informatics7040044

Roelli, P. (2014). The Corpus Corporum, a New Open Latin Text Repository and Tool. *Archivum Latinitatis Medii Aevi (ALMA): Bulletin Du Cange*, 72, 289–304.

Sassolini, E., Cucurullo, S., & Sassi, M. (2014). Methods of textual archive preservation. *Proceedings of CLiC-it 2014 – First Italian Conference on Computational Linguistics*, vol. I. Pisa, Italy, 9–10 December 2014. 334–338.

Schipani, S. (1994[1996]). Primo rapporto sulla attività della ricerca: "Il latino del diritto e la sua traduzione. Traduzione in italiano dei Digesta di Giustiniano". *Studia et documenta historiae et iuris, 60. Studi in Memoria di G. Lombardi.* Falchi, G. A. (Dir.). Roma: Pontificia Università Lateranense. 553–568.

Schipani, S. (2005). Premessa. *Iustiniani Augusti Digesta seu Pandectae. Digesti o Pandette dell'Imperatore Giustiniano. Testo e traduzione.* Volume I (libri 1–4). Schipani, S. (Ed.) & Lantella, L. Milano: Giuffrè Editore.

Schipani S. (Ed.), & Lantella, L. (2005a). *Iustiniani Augusti Digesta seu Pandectae. Digesti o Pandette dell'Imperatore Giustiniano. Testo e traduzione*. Volume I (libri 1–4). Milano: Giuffrè Editore.

Schipani S. (Ed.), & Lantella, L. (2005b). *Iustiniani Augusti Digesta seu Pandectae. Digesti o Pandette dell'Imperatore Giustiniano. Testo e traduzione*. Volume II (libri 5–11). Milano: Giuffrè Editore.

Schipani S. (Ed.), & Lantella, L. (2007). *Iustiniani Augusti Digesta seu Pandectae. Digesti o Pandette dell'Imperatore Giustiniano. Testo e traduzione.* Volume III (libri 12–19). Milano: Giuffrè Editore.

Schipani, S. (2010). 桑德羅. 斯奇巴尼教授文集 = Sandro Schipani, *scritti di diritto romano pubblicati in cinese.* 中國政法大學出版社. *Corporis Iuris Civilis. Digesta*. Schipani, S. (Ed.) (2001–2016). Beijing: CUPL. (books published in the collection: 1, 3–4, 6, 8–9, 12–13, 16–18, 22–24, 41, 48).

Schipani S. (Ed.), & Lantella, L. (2011). Iustiniani Augusti Digesta seu Pandectae. Digesti o Pandette dell'Imperatore Giustiniano. Testo e traduzione. Volume IV (libri 20–27). Milano: Giuffrè Editore.

Simard, M., & Plamondon P. (1998). Bilingual Sentence Alignment: Balancing Robustness and Accuracy. *Machine Translation,* 13, 59-80. DOI: https://doi.org/10.1023/A:1008010319408

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. & Varga, D. (2006). The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J. & Tapias, D. (Eds.).

Straka, M., Straková, J. & Gamba, F. (2024). ÚFAL LatinPipe at EvaLatin 2024: Morphosyntactic Analysis of Latin. *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*. Torino: ELRA and ICCL. 207–214.

Yousef, T., & Berti, M. (2015). The Digital Fragmenta Historicorum Graecorum and the Ancient Greek-Latin Dynamic Lexicon. *Proceedings of the Workshop on Corpus-Based Research in the Humanities (CRH)*, Warsaw, Poland, 10 December 2015. Mambrini, F., Passarotti, M., Sporleder, C. Warszawa (Eds.). 117–123.