CLARIAH-ES PressMint: Building Interoperable Corpora of Historical Press in Spain

Ainara Estarrona, Aritz Farwell, Xabier Goenaga, German Rigau
HiTZ Center - University of the Basque Country (EHU)
{ainara.estarrona, aritz.farwell, xabier.goenaga, german.rigau}@ehu.eus

Abstract

This paper describes CLARIAH-ES's contribution to PressMint in Spain as a distributed effort across regional nodes (e.g., Catalonia, Madrid, Basque Country, Galicia, Canary Islands, Alicante), each developing manageable corpora in partnership with key repositories such as ARCA, Patrimonio Digital Complutense, Euskariana, Jable, Galiciana, and the BVMC periodicals portal. A central technical challenge is heterogeneous legacy OCR quality, motivating experiments with AI/LLM-assisted OCR renewal, normalization layers, and linguistic enrichment (e.g., NER and entity linking). This effort is situated alongside ongoing dissemination and the EOSC Mesh "historical newspapers" use-case work aimed at scalable discovery, access, and federated computation over interoperable historical press data.

Keywords: CLARIAH-ES, corpus, historical newspapers, PressMint

1. Introduction

PressMint is a CLARIN Flagship initiative designed to produce a pan-European, interoperable, multilingual corpora of European historical newspapers (mostly from the late nineteenth to early twentieth centuries) that are comparable across countries and languages. A key motivation is that, although many national libraries already provide digitized newspaper collections, these datasets are typically not interoperable, which limits comparative "distant reading" approaches and broader European-scale analyses. PressMint addresses this by delivering a shared, FAIR-aligned resource and actively promoting its uptake among historians, linguists, media scholars, and other social sciences and humanties (SSH) communities. The project runs from June 2025 to May 2027 and is organized as a distributed effort involving multiple national consortia.

PressMint's objective is interoperability by design. Participating corpora are encoded to a common PressMint schema (a customisation of the TEI Guidelines) and supported by shared scripts and workflows—explicitly building on infrastructure and practices developed in the earlier ParlaMint project (Erjavec et al., 2023; Erjavec et al., 2025). In this manner, the same processing pipeline can be applied across corpora even when their source characteristics differ. In addition to TEI XML, PressMint plans to provide downstream formats (e.g., TSV, CoNLL-U, JSON) and to make the corpora openly available for download and through online corpus analysis tools, lowering the barrier for researchers to explore, compare, and reuse historical newspaper data at scale.

Within this European context, CLARIAH-ES[1], Spain's national research infrastructure for CLARIN and DARIAH, is actively contributing to PressMint in Spain by coordinating TEI encoding efforts, partnering with national and regional digital libraries, and experimenting with AI methods (including LLM-based approaches) to improve OCR and downstream processing. The following discussion describes CLARIAH-ES, its approach to building corpora for PressMint, and where the process of corpora creation for PressMint in Spain is at the present time.

2. CLARIAH-ES

CLARIAH-ES is a distributed digital research infrastructure created to coordinate Spain's participation in the two main European social sciences and humanities research infrastructure consortia, CLARIN and DARIAH (Riudavets et al., 2024). CLARIN's mission is to make language data, tools, and expertise accessible for research, while DARIAH focuses on digitally enabled research and teaching in the arts and humanities. Together, the two infrastructures form a complementary ecosystem of standards, services, and research communities.

CLARIAH-ES grew out of the earlier INTELE strategic network (2020–2022) and was a central player in Spain becoming a full member of CLARIN and DARIAH (Iruskieta et al., 2022). In practice, CLARIAH-ES exists to support the digital transformation of SSH research in Spain, enabling computational work with textual, visual, and audio materials by connecting researchers to shared infrastructures, methods, and tools. A core objective is to reduce the "digital divide" by promoting multilingualism, interoperability, resource sustainability and reuse, and open science practices through coordinated services and community building.[2]

[1] https://www.clariah.es/

[2] The network is funded by the Ministry of Science,

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 27–33

Organizationally, CLARIAH-ES brings together a multidisciplinary consortium that includes experts in language technologies, AI, HPC, library science, and SSH scholarship.[3] It comprises twelve nodes located across Spain, with its administrative and coordinating office based at HiTZ,[4] the Basque Center for Language Technology at the University of the Basque Country (EHU).

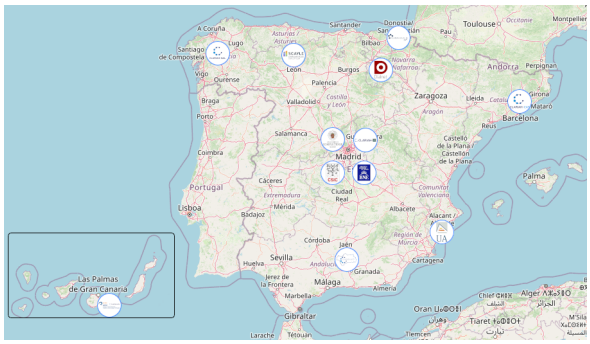

Figure 1: Map of CLARIAH-ES

3. Corpora For PressMint

CLARIAH-ES has taken the strategic decision to produce several separate corpora for PressMint according to regional nodes within the infrastructure, in part due to the presence of regional languages that were used in historical press in addition to Spanish. Currently, these nodes include CLARIAH-CAT (Catalonia), CLARIAH-CM (Madrid), CLARIAH-EUS (Basque Country), CLARIAH-GAL (Galicia), CLARIAH-IATEXT (Canary Islands), and CLARIAH-UA (Alicante). Other nodes are likely to participate in the future as well. Several of these nodes, in addition to CLARIAH-UNED (Madrid) and CLARIAH-SCAYLE (León),[5] will also experiment with AI techniques or provide expertise in language technology.

The distribution among these regional nodes in CLARIAH-ES has enabled several teams across the infrastructure to focus on building smaller, discrete, and more manageable corpora within the guidelines set out by PressMint. At the moment, these corpora are in different phases of development and experimentation, as described in the following subsections: 3.1. Compiled Corpora, 3.2. Corpora Under Construction, and 3.3. Language and AI Technology.

Innovation and Universities (RED2024-154077-E).

[3] https://www.clarin.eu/blog/introduction-clariah-es

[4] https://www.hitz.eus/en

[5] CLARIAH-SCAYLE will offer compute to perform OCR experiments.

3.1. Compiled Corpora

CLARIAH-GAL is currently the only node that already possesses a corpus of historical texts that may be adapted for use in PressMint[6] The collection, titled "Tesouro Informatizado da Lingua Galega" (TILG),[7] brings together over 3,000 documents written in Galician between 1612 and 2013, including historical press (which would be extracted for the PressMint corpus), totaling roughly 30 million word forms that are lemmatized and morphosyntactically tagged.

With respect to PressMint, one option under consideration is to take advantage of the existing annotation in the TILG and map its tags to Universal Dependencies (UD), which would avoid having to apply OCR and normalization from scratch. A drawback to this approach, however, is that TILG applies orthographic regularization, so the resulting corpus would be less faithful to the graphic and lexical variation of the original texts.

A second tentative alternative is to build a separate corpus from the historical press available in Galiciana,[8] the digital repository maintained by the Biblioteca de Galicia. This approach would preserve the original language as found in the historical texts, but also require CLARIAH-GAL to correct the largely unreviewed OCR for most of those texts, build a normalization layer, and perform any added annotation from scratch. The result would be a richer corpus from a historical perspective, but the effort would be considerably greater.

3.2. Corpora Under Construction

The latter option outlined above reflects the current status of the remaining corpora to a greater or lesser degree. As discussed in the following subsections, all other contributing nodes in CLARIAH-ES are at different stages in the process of constructing corpora in collaboration with regional repositories.

3.2.1. CLARIAH-CAT

CLARIAH-CAT,[9] coordinated by the Barcelona Supercomputing Center,[10] integrates research groups from every Catalan public university (UB, UAB, UPF,

[6] CLARIAH-GAL (https://www.clariah.gal/) is the Galician node of the Spanish CLARIAH-ES infrastructure, coordinated by the Instituto da Lingua Galega (https://ilg.usc.gal/en) with the collaboration of CiTIUS at the Universidade de Santiago de Compostela (https://citius.gal/).

[7] https://ilg.usc.es/TILG/

[8] https://biblioteca.galiciana.gal/gl/inicio/inicio.do

[9] Website currently under construction (https://clariah.cat/).

[10] https://www.bsc.es/

UdG, UdL, URV, UPC, UOC), as well as research centers such as the Computer Vision Center (CVC), the Artificial Intelligence Research Institute (IIIA-CSIC), and the Catalan Institute of Classical Archaeology (ICAC). In addition, the network includes the Biblioteca de Catalunya, the civil society organization SoftCatalà, and the companies LaTempesta and Avenir-Cultura.

CLARIAH-CAT has contacted the Arxiu de Revistes Catalanes Antigues (ARCA)[11] about the opportunity to build a corpus utilizing its collection of historical press. ARCA is a collaborative portal promoted by the Biblioteca de Catalunya[12] that provides centralized and open access to digitized Catalan historical newspapers and magazines.

3.2.2. CLARIAH-CM

CLARIAH-CM,[13] led by the Universidad Complutense de Madrid (UCM), acts as a regional coordination hub that brings together researchers and research groups from the six public universities in the Madrid area (the Universidad Politécnica, the Universidad de Alcalá), the Universidad Autónoma de Madrid), the Universidad Rey Juan Carlos), the Universidad Carlos III), and the Complutense). CLARIAH-CM expects to build its corpus with material provided by the Patrimonio Digital Complutense (PDC) and Biblioteca Digital de la Comunidad Madrid.

The PDC[14] is the UCM Library's portal for digitized cultural heritage. Its wide-ranging holdings include historical periodicals from the nineteenth and twentieth centuries. The Biblioteca Digital de la Comunidad Madrid[15] aggregates holdings from the Biblioteca Regional de Madrid, the Real Academia Española, the Real Academia de la Historia, and the Fundación Lázaro Galdiano. It provides access to mostly public-domain works dating roughly from the fifteenth to the twentieth century, with particular emphasis on materials related to Madrid and its culture.

3.2.3. CLARIAH-EUS

CLARIAH-EUS,[16] led by HiTZ, seeks to strengthen Basque language- and Basque culture–related digital humanities research (Alkorta et al., 2025). In addition to HiTZ, the node is made up of various research groups from the University of the Basque Country and from other institutions and organizations, such as the Soziolinguistika Klusterra[17], UEU-GOI[18], Badalab[19], and the Research Centre for Basque Language and Texts (Iker–administered by the CNRS, the University Bordeaux Montaigne, and the University of Pau and Pays de l'Adour).[20] CLARIAH-EUS is collaborating with Euskariana to construct its PressMint corpus and hopes to expand this collaboration to other regional libraries and repositories in the near future. Euskariana, the Basque Government's collaborative digital portal for Basque culture and cultural heritage, is managed by the Biblioteca Digital de Euskadi and gathers together contributions from various partners, including public administrations, municipalities, universities, cultural institutions, archives, libraries, and museums.

This material encompasses a large collection of historical press and Euskariana has agreed to provide CLARIAH-EUS and HiTZ with close to 180 periodicals –images, pdf, txt, and metadata– that cover the period 1874-1939. Published mostly in the Basque Country, these periodicals run the gamut of political ideologies and include newspapers printed in Basque, English, French, and Spanish. All the objects are OCRed, but it is likely that many will benefit from a renewed OCR effort given that most were processed over fifteen years ago. HiTZ is currently considering how best to create datasets for evaluation (a gold standard) and exploring methods to apply OCR using LLMs. The node also expects to enrich the corpus with normalization, NERC, and entity linking.

3.2.4. CLARIAH-IATEXT

CLARIAH-IATEXT[21] is the Canary Islands node of CLARIAH-ES, led by the Instituto Universitario de Análisis y Aplicaciones Textuales (IATEXT) at the Universidad de Las Palmas de Gran Canaria (ULPGC). In similar fashion to its counterparts, CLARIAH-IATEXT serves as a regional infrastructure for Canarian research communities working in the fields of digital humanities and social sciences. CLARIAH-IATEXT expects to build its PressMint corpus in collaboration with the Museo Canario and the Biblioteca de la ULPGC.

The Museo Canario's repository of historical press[22] contains a continuously growing archive

[11] https://arca.bnc.cat/arcabib_pro/en/inicio/inicio.do

[12] The Biblioteca de Catalunya is Catalonia's national library (https://www.bnc.cat/).

[13] https://www.ucm.es/clariah-cm-en

[14] https://patrimoniodigital.ucm.es/s/patrimonio/page/inicio

[15] https://bibliotecavirtualmadrid.comunidad.madrid/bvmadrid_publicacion/es/inicio/inicio.do

[16] https://www.clariah.eus/en

[17] https://soziolinguistika.eus/en/

[18] https://www.goi-institutua.eus/

[19] https://badalab.eus/

[20] https://iker.cnrs.fr/?lang=en

[21] https://iatext.ulpgc.es/es/clariah_iatext

[22] https://www.elmuseocanario.com/en/

that aims to gather all periodicals published in the Canary Islands, complemented by titles produced by Canarian communities abroad. It is regarded as the most complete repository in the archipelago, holding periodicals dating from the eighteenth century to the present day, including fin-de-siècle newspapers, when the Canary Islands became one of the five Spanish provinces with the highest newspaper production. The periodical repository at the Biblioteca de la ULPGC is largely delivered through Jable,[23] the Canary Islands Digital Press Archive, a long-running initiative that provides access to historical and modern periodicals published in the Canary Islands.

3.2.5. CLARIAH-UA

CLARIAH-UA[24] is the University of Alicante's research infrastructure node for digital humanities within the CLARIAH-ES consortium. The node is closely associated with the Biblioteca Virtual Miguel de Cervantes (BVMC), whose mission includes advancing DH research, designing technologies for the humanities, and providing access to Hispanic cultural material. The BVMC curates a dedicated periodicals portal[25] that gives researchers and the general public structured access to pre-1930 historical press, comprising about 275 titles. CLARIAH-UA will draw on this collection to construct its corpus for PressMint.

3.3. Language and AI Technology

Much of the material that will be utilized to build corpora for PressMint across CLARIAH-ES's nodes possess OCR that was applied several years before. Recent advances in OCR technology that leverages LLMs have significantly improved and facilitated the application of OCR to historical texts, including periodicals. In addition to other AI-related tools and approaches, CLARIAH-ES is experimenting with how these OCR techniques may be applied to the respective corpora it is producing for PressMint.

Although several nodes within CLARIAH-ES will participate in this initiative, CLARIAH-UNED and CLARIAH-EUS have already begun to explore how best to approach the OCR problem. CLARIAH-UNED,[26] is doing so as part of the GRESEL-UNED project,[27] which investigates Spanish-language historical press between 1850 and 1945 by using LLMs to analyze discourses about nation, identity, feminism, literature, and international relations. The project's core goal is to build a RAG research assistant that is reliable for scholarly inquiry, but to make the assistant effective, the project is developing and adapting linguistic resources and NLP/IR workflows, including training or tailoring models, extracting and classifying historically meaningful entities, and improving factual grounding through retrieval so researchers may run better question-answering and exploratory analyses over the corpus.

CLARIAH-EUS and HiTZ are investigating how Latxa,[28] the largest and best-performing LLM available for Basque, may be harnessed to help build its corpus for PressMint. Early experiments with OCR involve tests utilizing PERO OCR (Kodym and Hradiš, 2021) and ScribbleSense.[29] For normalization, both statistical (Phonetisaurus,[30] cSMITiser) and AI techniques are under evaluation (Ljubešić et al., 2016; Scherrer and Ljubešić, 2016).

4. PressMint-Related CLARIAH-ES Events and Dissemination

Members of CLARIAH-ES have organized or participated in several events that have included a focus on historical press or PressMint. Below is a list of some of these as well as several more that will take place in the coming months.

4.1. PastReader

PastReader[31] was an IberLEF 2025 shared task designed to enable automatic transcription of digitized Spanish historical press using newspapers housed at the digital repository of Spain's National Library. The PDFs that where utilized often include OCR but the extracted text can be unreliable because of degraded scans, complex layouts, and historical typography. To address these challenges, PastReader organized evaluation around two core problems: (1) OCR error correction and (2) end-to-end "curated" text extraction directly from scanned images, encouraging the use of multimodal approaches as well as robust post-processing. Overall, the task's goal was to reduce human effort in mass digitization workflows and improve the accessibility, retrieval, and

newspaper/

[23] https://jable.ulpgc.es/

[24] https://clariah-ua.cervantesvirtual.com/

[25] https://www.cervantesvirtual.com/portales/hemeroteca/r

[26] https://clariah.uned.es/

[27] The GRESEL initiative (PID2023-151280OB-C22) is funded by Spain's Ministry of Science, Innovation and Universities (https://gresel-uned.hypotheses.org/).

[28] https://www.hitz.eus/en/node/340

[29] https://scribblesense.cz

[30] https://github.com/AdolfVonKleist/Phonetisaurus

[31] https://sites.google.com/view/pastreader2025

long-term preservation of Spanish newspaper heritage by benchmarking and promoting more accurate, efficient transcription systems.

4.2. FDS Seminar “Humanidades Digitales en Acción"

The Fundación Duques de Soria seminar “Humanidades digitales en acción: herramientas para el análisis de prensa histórica en español,”[32] held in Soria on July 2-4, 2025, was an award-winning international initiative that gathered together a multidisciplinary team to discuss how to make Spanish historical newspapers easier to access, digitize, and analyze through digital methods, Conceived as a bridge between three professional communities —library science specialists, humanists, and computer scientists— the seminar was structured around core thematic axes that combined historical and literary approaches to newspapers with hands-on digital workflows for digitization/OCR, discoverability, and computational analysis.

4.3. Fourth CLARIAH-EUS Workshop

The Fourth CLARIAH-EUS Workshop (“Humanitate Digitalak eta Gizarte Zientziak gaur egungo Hizkuntza Teknologiak aplikatuta”)[33] took place on November 28, 2025 in Vitoria-Gasteiz. A CLARIAH-EUS community event focused on applying current language technologies to digital humanities and the social sciences, it included a poster session that showcased tools, corpora, and RAG-oriented work undertaken by members of the infrastructure. A poster dedicated to PressMint, “PRESSMINT: Egunkari Historikoen Corpus Elkarreragingarriak,” introduced the project to the Basque audience.

4.4. II Ciclo CLARIAH-CM

The “II Ciclo CLARIAH-CM: formación en Prensa Histórica (Herramientas y metodologías digitales),”[34] an in-person training series launched by the CLARIAH-CM node, provides regular, hands-on workshops plus short theoretical introductions on digital methods for humanities research. The current cycle focuses on building, processing, and analyzing Spanish historical press corpora. The aim is to help researchers streamline workflows and deepen methodological expertise in this field and consists of three sessions that combine guided practice with the option to work on participants’ own materials. The sessions include Label Studio for OCR dataset creation (February 27, 2026), historical press corpus processing with Sketch Engine (March 23, 2026), and AI-based exploration of historical press (April 27 2026).

4.5. I Jornada CLARIAH-ES en Bibliotecas

The "I Jornada CLARIAH-ES en Bibliotecas: Infraestructuras Digitales y Ciencia Abierta"[35] (Madrid, March 3, 2026) is an outreach-oriented event hosted at the Instituto Cervantes (Madrid) that brings together librarians, archivists, and researchers to discuss how digital research infrastructures can help transform traditional collections into open, reusable, and more visible resources within the broader ecosystem of Open Science. The program is organized into three thematic blocks: an introduction to infrastructures (CLARIN, DARIAH, CLARIAH-ES, SSH Open Marketplace, and EOSC), a section devoted to workflows and projects, and a closing block focused on data-driven library services and community building. A key highlight is the PressMint session, which introduces PressMint as an example of how libraries and infrastructures can collaborate to improve large-scale access, processing, and reuse of historical newspaper collections.

4.6. IA y Humanidades Digitales para la Prensa Histórica

The CLARIAH-UNED and CLARIAH-EUS organized event, “IA y humanidades digitales para la prensa histórica”[36] (Madrid, April 20, 2026) is a forum devoted to how AI and digital humanities can improve the digitization, transcription, and analysis of historical newspapers. The morning sessions are dedicated to the AI-driven exploration of historical press (particularly methodological challenges such as verification, prompting strategies, and AI analysis of multilingual press) and a library roundtable that brings together perspectives from major cultural heritage institutions, including the "Hemeroteca Digital" at Spain’s National Library and the "Biblioteca Virtual de Prensa Histórica," maintained by Spain’s Ministry of Culture. The afternoon features a discussion on research design for multimodal OCR inference, domain-specific entity recognition, and knowledge-graph/semantic exploration approaches, followed a session dedicated to PressMint, connecting the event to broader

[32] https://fds.es/seminario-humanidades-digitales-en-accion-herramientas-para-el-analisis-de-prensa-historica-en-espanol

[33] https://www.clariah.eus/eu/4.workshopa

[34] https://www.ucm.es/clariah-cm/ii-ciclo-clariah-cm-formacion-en-prensa

[35] https://www.clariah.es/en/node/50

[36] https://gresel-uned.hypotheses.org/jornada-ia-y-humanidades-digitales-para-la-prensa-historica

European efforts to build interoperable historical newspaper corpora.

4.7. III Xeira CLARIAH-GAL

The III Xeira CLARIAH-GAL (May 31, 20026) is an annual event organized by the Instituto da Lingua Galega and supported by CiTIUS. It is designed as a networking and dissemination space to connect Galician projects and research groups working in digital humanities, arts, social sciences, and language-technology supported research. This year, one of the talks will be devoted to PressMint.

4.8. First CLARIAH-ES Summer School

The first CLARIAH-ES Summer School, "Impulsando las Humanidades Digitales en la era de la IA generativa," to take place in June in Donostia-San Sebastian, will also devote space to historical press. The two-day course will feature a keynote talk on new LLM-based approaches for digital humanities using historical newspapers.

5. EOSC Mesh Use Case

CLARIN, DARIAH, HiTZ (CLARIAH-EUS), and the Universidad de Jaén (CLARIAH-AND) are all involved in the EOSC Mesh project. EOSC Mesh is a Horizon Europe project designed to reduce fragmentation in Europe's research data and service landscape by expanding the EOSC Federation with seven interoperable nodes working together as an EOSC Mesh Hub. Overall, its objective is to make EOSC more resilient and scalable, accelerate node onboarding via a Node Operator Framework, and enable large-scale discovery and reuse of research objects across domains.

Several use cases have been designed within the project. One, "Discovery, access and integration of data from historical newspapers," focuses on enabling datafication and computational analysis workflows for textual cultural heritage, specifically digitized historical newspapers. Part of the objective is to overcome several of the main obstacles with respect to interoperability caused by the disparate nature of the collections, such as discovering relevant content across multiple trusted repositories, gaining access, integrating heterogeneous sources, and selecting suitable tools/execution environments. This use case targets cultural heritage institutions (libraries, archives, museums) and humanities/DH researchers (estimated at 500,000 in Europe) and anticipates infrastructure needs of around 150 TB to store and process corpora alongside key knowledge bases (e.g., Wikidata and GeoNames). It also leverages EOSC Mesh core capabilities, generic capabilities (cloud compute, notebooks, online storage), and the SSHOC Marketplace as the thematic entry point, with development steps that include deploying workflows for new service creation and building federated Retrieval-Augmented Generation (RAG) systems over indexed datasets produced by the datafication pipelines.

6. Conclusion

PressMint provides a timely and structured response to the rapid growth of digitized historical newspapers in Europe. It seeks to make these historical newspaper collections comparable across countries and languages by pursuing interoperability by design. PressMint does so through a shared TEI-based schema, reusable scripts/workflows, and dissemination in multiple downstream formats to lower barriers for "distant reading" and large-scale reuse. Overall, PressMint is not merely as a corpus-building project, but also a shared European methodology for transforming historical newspapers into an interoperable network, unlocking the potential for new comparative research questions while encouraging collaboration among technologists, humanists, and cultural heritage institutions.

In this context, CLARIAH-ES contributes to Spain's involvement in PressMint by aligning partners, repositories, and technical practices across its distributed national infrastructure. The decision to develop several regional corpora has enabled teams to work with manageable collections while still converging on common standards: assembling corpora through partnerships with major digital libraries and experimenting with AI-based improvements to OCR and normalization. Across these efforts, the main technical challenge remains the heterogeneity and quality of legacy OCR, often produced many years ago, making robust correction, normalization layers, and consistent linguistic enrichment central to producing reliable, comparable corpora.

Ongoing experimentation with AI tools and techniques will ideally help improve transcription quality and scholarly utility, while dissemination activities will aid in increasing PressMint's uptake across library, humanities, and language technology communities. Finally, alignment with broader initiatives such as the EOSC Mesh use case on historical newspapers points toward scalable discovery, access, and federated computation over interoperable historical press data, supporting reproducible, cross-regional and cross-lingual research on Europe's cultural and political history at an unprecedented scale.

7. Acknowledgements

PressMint is funded by CLARIN ERIC. The CLARIAH-ES infrastructure is funded by the Ministry of Science, Innovation and Universities (RED2024-154077-E). EOSC Mesh is funded by the European Commission.

8. Bibliographical References

Jon Alkorta, Aritz Farwell, Joseba Fernandez de Landa, Begoña Altuna, Ainara Estarrona, Mikel Iruskieta, Xabier Arregi, Xabier Goenaga, Jose Mari Arriola, Inma Hernáez, and David Lindemann. 2025. CLARIAH-EUS: A strategic network helping basque country researchers to participate in european research infrastructures. In *Selected papers from the CLARIN Annual Conference 2024. Linköping Electronic Conference Proceedings 216 (eds. Vincent Vandeghinste and Thalassia Kontino)*.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Matyáš Kopp, Steinþór Steingrímsson, Sigrún Helgadóttir, Črtomir Grobol, et al. 2023. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57(1):415–448.

Tomaž Erjavec, Maciej Ogrodniczuk, Agnes Pisanski Peterlin, Simon Krek, and Andrej Pančur. 2025. Parlamint ii: advancing comparable parliamentary corpora across europe. *Language Resources and Evaluation*, 59(3):1–25.

Mikel Iruskieta, Ainara Estarrona, Aritz Stephen Farwell, and Germán Rigau. 2022. INTELE: promoviendo la participación en las infraestructuras eric clarin y dariah. *Boletín de la ANABAD*, 72(2):63–91.

Oldřich Kodym and Michal Hradiš. 2021. Page layout analysis system fornbsp;unconstrained historic documents. In *Document Analysis and Recognition – ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II*, page 492–506, Berlin, Heidelberg. Springer-Verlag.

Nikola Ljubešić, Katja Zupan, Darja Fišer, and Tomaž Erjavec. 2016. Normalising Slovene data: historical texts vs. user-generated content. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 146–155.

Francisco J. Carreras Riudavets, Ainara Estarrona, Aritz Farwell, Mikel Iruskieta, Manuel Marco Such, Maite Melero, Arturo Montejo-Ráez, Daniel Riaño, German Rigau, Dolores Romero, Salvador Ros, Elena Sánchez, and Xulio Sousa. 2024. CLARIAH-ES: Strategic network for the integration in the european research infrastructures in social sciences and humanities. In *SEPLN (Projects and Demonstrations)*, volume 3729 of *CEUR Workshop Proceedings*, pages 30–35. CEUR-WS.org.

Yves Scherrer and Nikola Ljubešić. 2016. Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 248–255.