



LREC 2026

**First Workshop on Creating Interoperable Corpora of
Historical Newspapers**

Workshop Proceedings

Editors

Maciej Ogrodniczuk, Petya Osenova and Tanja Wissik

May 16, 2026

©ELRA Language Resources Association (ELRA), 2026
These proceedings are licensed under a Creative Commons Attribution-
NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-71-5

Preface

Welcome to the 1st Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint), held on May 16, 2026, in Palma de Mallorca, Spain, as part of the 15th biennial Language Resources and Evaluation Conference (LREC 2026).

While historical newspapers are a goldmine for researchers in linguistics, history, and social sciences, they often suffer from a lack of interoperability. This workshop represents a milestone for the PressMint project – a CLARIN flagship initiative – which aims to build a multilingual, TEI-encoded, and standardized set of corpora covering European press from around the start of the 20th century. By bringing together over 20 partners across Europe, the project aims to bridge the gap between isolated national archives and a unified, transnational research infrastructure.

For this inaugural edition, we received 17 submissions. Following a rigorous double-blind peer review process, the program committee decided to accept 14 papers, resulting in an 82% acceptance rate. These contributions reflect a high standard of scholarship, with an average score of approximately 3.81/5 across the accepted submissions. Three papers were presented orally, including the project overview paper, one paper representing data modelling and another one discussing data processing. 11 papers were presented as posters.

The program of the workshop highlights the geographic and thematic breadth of the PressMint community. The authors represent a diverse array of countries, including Austria, Bulgaria, Denmark, Germany, Hungary, Italy, Poland, Portugal, Slovenia, Spain, Switzerland and Ukraine. The presented papers explore critical dimensions of historical data processing, including corpus construction, interoperability and standards, technical challenges and user perspectives.

We are especially honored to welcome our invited speaker, Maud Ehrmann from Ecole Polytechnique Fédérale de Lausanne (EPFL), whose pioneering work in the digital humanities provides a vital context for our efforts in newspaper processing.

The success of this workshop is due to the hard work of many. We are deeply grateful to our Program Committee for their detailed and constructive reviews, which ensured the high quality of these proceedings. We also thank CLARIN ERIC and the PressMint project for their ongoing support. Finally, we would like to thank the workshop participants for joining us.

We hope these proceedings serve as a valuable resource for the community and inspire further collaboration in the preservation and analysis of our shared European history.

We hope you enjoyed the workshop and Mallorca as much as we did!

— Maciej Ogrodniczuk, Petya Osenova and Tanja Wissik

Organizers and Reviewers

Organizing Committee

Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences, PL

Tanja Wissik, Austrian Academy of Sciences, AT

Petya Osenova, Sofia University "St. Kl. Ohridski" & IICT-BAS, BG

Program Committee

Emanuela Boros, University of La Rochelle, FR

Jesse de Does, Instituut voor de Nederlandse Taal, Leiden, NL

Maud Ehrmann, École Polytechnique Fédérale de Lausanne, CH

Tomaž Erjavec, Jožef Stefan Institute, SI

Aritz Farwell, HiTZ Basque Research Center for Language Technology, EHU, ES

Maria Gavrilidou, Institute for Language and Speech Processing, Athena Research Center, GR

Normunds Grūzītis, University of Latvia, LV

Maarten Janssen, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. Charles University, CZ

Matyáš Kopp, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. Charles University, CZ

Noémi Ligeti-Nagy, Hungarian Research Centre for Linguistics, HU

Nikola Ljubešić, Jožef Stefan Institute, SI

Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences, PL

Petya Osenova, Sofia University "St. Kl. Ohridski" and IICT-BAS, BG

Adam Pawłowski, University of Wrocław, PL

Stelios Piperidis, Athena Research Centre, GR

Claudia Resch, Austrian Academy of Sciences, AT

German Rigau, HiTZ Basque Research Center for Language Technology, EHU, ES

Maria Shvedova, National Technical University "Kharkiv Polytechnic Institute" UA / University of Jena, DE

Inguna Skadiņa, Institute of Mathematics and Computer Science, University of Latvia, LV

Steinþór Steingrímsson, The Árni Magnússon Institute for Icelandic Studies, IS

Egon W. Stemle, Institute for Applied Linguistics, Eurac Research, IT

Tanja Wissik, Austrian Academy of Sciences, AT

Invited Talk

Beyond Borders: Connecting Historical Media at Scale with Impresso

Maud Ehrmann

Abstract

Mass digitisation has transformed historical research by producing vast machine-readable corpora – notably historical newspapers and broadcasts – making full-text search and text mining-based enrichments standard tools for retrieval and exploration. Yet digitised collections remain fragmented, siloed by media type, language, institution, and national boundary, and the transformative potential of studying them together, at scale and across borders, remains unrealised.

Overcoming this fragmentation requires more than aggregation. Connectivity is the core ambition of *Impresso — Media Monitoring of the Past*: not simply consolidating collections, but semantically enriching and linking them, and ensuring legally grounded access through interfaces designed for historical research.

This talk presents Impresso’s strategy for connecting historical media collections from over twenty European cultural heritage institutions along four interdependent dimensions. First, data governance: cross-border access to copyright-protected collections demands frameworks that balance institutional control with researcher needs, amid concerns around AI and data reuse. Second, data interoperability: processing and indexing diverse collections at scale requires reconciling heterogeneity in format, levels of granularity and quality, and archival structures. Third, data processing: connecting information requires semantic enrichment and linking across languages and modalities within a multilingual embedding space. Fourth, interface design: transnational and transmedia discovery requires versatile research affordances – from human-centred exploration to fully programmatic, data-driven inquiry – as provided by the Impresso WebApp and Datalab.

Together, these efforts outline a model for connecting historical media responsibly at scale, ensuring that access extends beyond institutional borders, semantic enrichment spans languages and modalities, and scholarly inquiry is supported by versatile interfaces.

Speaker Bio

Maud Ehrmann is a research scientist and lecturer at the Digital Humanities Laboratory (DH LAB) of the École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. Her research sits at the intersection of natural language processing, digital humanities, and historical document processing, with a focus on named entity recognition and linking, semantic enrichment, and the large-scale processing of historical media collections. She works on Impresso, a series of two interdisciplinary research projects that uses machine learning to advance the processing, semantic enrichment, exploration, and study of historical media across modalities, time, languages, and national borders. She has co-initiated and co-organises the series of HIPE shared tasks on historical document processing, originally centered on named entity processing, and now extending to relation extraction and OCR post-correction.

Table of Contents

<i>PressMint: Towards Interoperable Corpora of Historical Newspapers</i>	
Tomaž Erjavec, Matyáš Kopp, Maciej Ogrodniczuk, Petya Osenova and German Rigau .	1
<i>The Polish PressMint Corpus</i>	
Maciej Ogrodniczuk, Dariusz Czerski and Adam Pawłowski	6
<i>PressMint QuickCheck: Operationalising Readiness Diagnostics for Interoperable Historical Newspaper Corpora</i>	
Elena Battaner Moro, Almudena Caballos Villar, María Cuevas Riaño, Marina Miguez Lamanuzzi and Dolores Romero López	11
<i>PressMint-PT - Compiling a Portuguese Historical Newspaper Corpus</i>	
Jose Aires and Amália Mendes	16
<i>Historical Newspapers in the General Regionally Annotated Corpus of Ukrainian (GRAC): Current State and PressMint Integration Prospects</i>	
Maria Shvedova and Arsenii Lukashevskyi	21
<i>CLARIAH-ES PressMint: Building Interoperable Corpora of Historical Press in Spain</i>	
Ainara Estarrona, Aritz Farwell, German Rigau and Xabier Goenaga	27
<i>Towards an Interoperable Corpus of Austrian Historical Newspapers: The case of PressMint-AT</i>	
Tanja Wissik, Jona Hassenbach, Hannes Pirker, Claudia Resch and Stefan Resch	34
<i>A Growing Literature of the Public Sphere: Fiction in Danish Newspapers (1666–1850)</i>	
Pascale Feldkamp, Alie Lassche, Rie Eriksen, Kit Morgenstjerne, Kristoffer Nielbo, Johan Heinsen and Yuri Bizzoni	40
<i>Towards an interoperable Hungarian historical newspaper corpus</i>	
Noémi Ligeti-Nagy and Henrietta Szabó	50
<i>Towards a Bulgarian Historical Newspaper Corpus – Construction of Reading Order over the Text in Searchable PDFs</i>	
Nikolay Paev, Stefan Marinov, Ivan Kratchanov, Petya Osenova and Kiril Simov	56
<i>A Survey of the Digitisation of German Newspapers in interwar Lithuania (1918–1940)</i>	
Lina Plaušinaityė and Heike Zinsmeister	65
<i>Toward Interoperable and Scalable Representations of Complex Heterogeneous Digitized Historical Media</i>	
Pauline Conti, Simon Clematide and Maud Ehrmann	72
<i>Data Matters: Looking for High-Quality Corpora to Build Robust and Reliable Models for Humanists</i>	
Jaione Macicior-Mitxelena and Ana García-Serrano	82
<i>Thematic Landscapes of the Past: Analysing Slovene Historical Periodicals With Topic Modeling</i>	
Filip Dobranić, Uroš Šmajdek, Oliver Pejić, Ciril Bohak, Vojko Gorjanc, Tina Munda and Darja Fiser	92

Workshop Program

Saturday, May 16, 2026

Opening Remarks

09:00–09:05 *Opening and Welcome*
Maciej Ogrodniczuk, Petya Osenova and Tanja Wissik

Invited Talk

09:05–10:00 *Beyond Borders: Connecting Historical Media at Scale with Impresso*
Maud Ehrmann

Project Overview

10:00–10:30 *PressMint: Towards Interoperable Corpora of Historical Newspapers*
Tomaž Erjavec, Matyáš Kopp, Maciej Ogrodniczuk, Petya Osenova and German Rigau

10:30–11:00 *Coffee Break*

11:00–12:00 **Poster Session**

The Polish PressMint Corpus
Maciej Ogrodniczuk, Dariusz Czerski and Adam Pawłowski

PressMint QuickCheck: Operationalising Readiness Diagnostics for Interoperable Historical Newspaper Corpora
Elena Battaner Moro, Almudena Caballos Villar, María Cuevas Riaño, Marina Miguez Lamanuzzi and Dolores Romero López

PressMint-PT — Compiling a Portuguese Historical Newspaper Corpus
Jose Aires and Amalia Mendes

Historical Newspapers in the General Regionally Annotated Corpus of Ukrainian (GRAC): Current State and PressMint Integration Prospects
Maria Shvedova and Arsenii Lukashevskyi

CLARIAH-ES PressMint: Building Interoperable Corpora of Historical Press in Spain
Ainara Estarrona, Aritz Farwell, German Rigau and Xabier Goenaga

Towards an Interoperable Corpus of Austrian Historical Newspapers: The case of PressMint-AT
Tanja Wissik, Jona Hassenbach, Hannes Pirker, Claudia Resch and Stefan Resch

11:00–12:00	Poster Session <i>A Growing Literature of the Public Sphere: Fiction in Danish Newspapers (1666–1850)</i> Pascale Feldkamp, Alie Lassche, Kit Morgenstjerne, Kristoffer Nielbo, Johan Heinsen and Yuri Bizzoni <i>Towards an Interoperable Hungarian Historical Newspaper Corpus</i> Noémi Ligeti-Nagy and Henrietta Szabó <i>Towards a Bulgarian Historical Newspaper Corpus – Construction of Reading Order over the Text in Searchable PDFs</i> Nikolay Paev, Stefan Marinov, Ivan Kratchanov, Petya Osenova and Kiril Simov <i>A Survey of the Digitisation of German Newspapers in interwar Lithuania (1918–1940)</i> Lina Plaušinaitytė and Heike Zinsmeister <i>Toward Interoperable and Scalable Representations of Complex Heterogeneous Digitized Historical Media</i> Pauline Conti, Simon Clematide and Maud Ehrmann
12:00–12:30	Oral Presentation: Data Modelling <i>Data Matters: Looking for High-Quality Corpora to Build Robust and Reliable Models for Humanists</i> Jaione Macicior and Ana García-Serrano
12:30–13:00	Oral Presentation: Data Processing <i>Thematic Landscapes of the Past: Analysing Slovene Historical Periodicals With Topic Modeling</i> Filip Dobranić, Uroš Šmajdek, Oliver Pejić, Ciril Bohak, Vojko Gorjanc, Tina Munda and Darja Fišer Closing Remarks
13:00–13:05	<i>Closing the workshop</i> Maciej Ogrodniczuk, Petya Osenova and Tanja Wissik