Towards an interoperable Hungarian historical newspaper corpus

Noémi Ligeti-Nagy, Henrietta Szabó
ELTE Research Centre for Linguistics
Benczúr u 33. Budapest Hungary
{surname.firstname}@nytud.elte.hu

Abstract

PressMint is a CLARIN initiative that aims to build multilingual, comparable and interoperable corpora of historical newspapers. For Hungarian, the main challenge is not a lack of material but fragmentation: newspapers are distributed across several portals, with heterogeneous metadata, access paths and OCR quality. This extended abstract reports the current status of the Hungarian PressMint subcorpus, focusing on the 19th century and the early 20th century (roughly 1800–1920). We describe two project artefacts already used in practice: a structured source inventory and a validation-driven repository. We summarise source scouting across Europeana, Hungaricana, OSZK–EPA, DiFMOE and related portals, including a curated 12-title Hungaricana manual-download pilot list with explicit target coverage periods. We then outline a reproducible pipeline for acquisition, OCR, layout analysis and conversion to PressMint-compatible TEI with facsimile linkage. Finally, we specify near-term deliverables for a first Hungarian release candidate and the evaluation steps planned for OCR and layout processing.

Keywords: historical newspapers, interoperability, TEI, OCR, CLARIN, PressMint, Hungarian

1. Workshop context and scope

PressMint aims to provide a common format and comparable linguistic annotation for multilingual corpora of historical newspapers that overlap, at least partly, in time. For the Hungarian contribution, we adopt a strict time window: the 19th century and the early 20th century (roughly 1800–1920). This period is highly valuable for historical research and is usually less constrained by copyright and reuse conditions than later newspaper material. It also aligns well with the turn-of-the-century perspective that often motivates cross-national press comparison.

Hungarian brings two PressMint-relevant characteristics. First, its diacritics-rich orthography and the frequent hyphenation found in narrow newspaper columns mean that OCR noise can directly affect tokenisation, sentence splitting and later annotation. Second, Hungarian-language newspapers were historically published both within and beyond the borders of present-day Hungary, so provenance and cross-border coverage are essential selection criteria.

Our main design principle is to keep *selection* separate from *processing*. Selection must remain transparent and revisable as new sources are identified and coverage gaps become visible. Processing must remain reproducible and validation-driven. This is why the project is organised around two central artefacts already in daily use: a structured inventory and a repository that turns inventory records into buildable TEI releases.

2. Related work, existing resources and standards

PressMint follows a well-established CLARIN pattern: multilingual partners converge on a shared, validation-driven interchange format while keeping national acquisition and preprocessing pipelines partly heterogeneous. ParlaMint is a strong point of reference here, demonstrating that shared TEI modelling, stable identifiers and coordinated release governance can work at scale (Erjavec et al., 2023).

In the historical newspaper domain, several projects have shown that interoperability depends on profiling and conversion rather than on a single upstream format. Europeana Newspapers promoted METS/ALTO profiling for exchange and highlighted the practical cost of harmonising provider metadata (Mühlberger, 2014; Freire et al., 2019). Neudecker's overview of searchable historical newspapers documents the diversity of OCR outputs and the need for conversion layers (Neudecker and Antonacopoulos, 2016). Impresso is a prominent example of combining facsimile-linked OCR text with curated metadata and stand-off enrichments (Ehrmann et al., 2020). OCR quality is also not a secondary technical issue: user-oriented studies show measurable effects of OCR quality on the perceived usefulness of historical newspaper collections (Kettunen et al., 2022).

For Hungarian, important digitised historical newspaper holdings already exist in Hungaricana, OSZK–EPA, Arcanum, ANNO and smaller institutional collections. These resources are highly valuable, but they are usually delivery platforms or

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 50–55

digitised holdings rather than interoperable corpora distributed in a shared TEI representation with stable identifiers and reproducible conversion steps. The Hungarian PressMint work therefore builds on existing digitisation efforts rather than duplicating them: the goal is to turn a selected subset of available material into a corpus that is consistent with PressMint-wide requirements.

3. Source scouting: what is available for Hungarian (and how)

The Hungarian subcorpus is *not* currently driven by a single ready-to-ingest provider. Instead, our work begins with explicit source scouting and a living inventory, and we only commit to titles once (i) the acquisition path is stable and reproducible and (ii) the time window is satisfied.

3.1. Scouted portals and current extraction status

Our scouting currently covers Europeana (Willems and Atanassova, 2015), Hungaricana (Hungaricana, n.d.), OSZK–EPA (Országos Széchényi Könyvtár (OSZK), 2026), DiFMOE (Digital Forum Central and Eastern Europe, 2023), ANNO (Müller, 2004; Austrian National Library, n.d.), Arcanum (Arcanum, n.d.), the National Archive press interface and related portals. Because these resources are not equally well suited for reproducible ingestion, we record both *availability* and *acquisition feasibility*.

Four scouting results are already directly relevant for PressMint planning. First, Europeana, a pan-European aggregation platform, yielded 69,290 Hungarian-language *text-type* items through API harvesting. This set is useful for discovery, but it still requires substantial filtering because the newspaper filter was not reliable enough for Hungarian items. Second, OSZK–EPA, the National Széchényi Library's electronic periodicals archive, contains many relevant collections; our current manual survey lists at least 142 regional, 54 cross-border and 50 newspaper collections, with overlaps still to be resolved. Third, DiFMOE, a portal that includes digitised periodicals from Central and Eastern Europe, yielded eight Hungarian-language periodicals already entered into the inventory with URLs; seven of them fall within the PressMint time window. Fourth, Hungaricana, the main Hungarian digital heritage portal, is now represented by a curated manual-download list of 12 titles with title-level collection URLs, holding institutions, full available coverage periods and explicitly selected in-window target periods.

The Hungaricana list is important because it moves that source from a vague manual option to a concrete acquisition pilot. The working sheet is already detailed enough to support title-level planning: for each candidate it records a stable internal source ID, the holding institution, the Hungaricana collection title, the full available coverage range, the selected PressMint target period and the title-level URL. At the same time, it is still an acquisition manifest rather than a full issue-level inventory: issue counts, page counts, download progress and notes on source-specific constraints are not yet populated.

This title-level detail matters for planning because the selected target periods are not uniform. Most titles are currently capped at 1915 to stay safely within the PressMint window, but some have narrower selections: *Eger – hetilap* ends in 1914, *Szatmári Értesítő* is represented by a single year (1862), *Váczi Közlöny* ends in 1895, and *Felsőmagyarországi Hírlap* is currently restricted to 1903–1907. The Hungaricana pilot list therefore provides not only a count of candidate titles, but an explicit acquisition plan for what should actually be downloaded first.

An important methodological consequence is that the inventory – not the processing scripts – is the main coordination artefact. It records what is known, what is reproducible and what still remains uncertain.

3.2. Selection criteria for the PressMint tranche

Within the time window, our planned inclusion criteria are: (i) stable and citable identifiers for title and issues/pages, (ii) at least page images (PDFs or images) that allow facsimile linkage, (iii) sufficient metadata to support cross-corpus comparison (publication place, years, language), and (iv) a realistic acquisition path at scale. We also prioritise cross-border Hungarian press and regionally diverse publication places, because these are especially relevant for comparative work on public discourse and regional vocabulary.

3.3. Inventory as coordination layer

The Hungarian inventory currently contains more than 150 candidate records with a fixed column schema covering provider, title, place of publication, available years, selected target coverage, acquisition method, persistent identifiers, scan/OCR fields and ingestion bookkeeping. This inventory is validated and aligned with controlled vocabularies in the repository (Section 4). As a result, selection decisions are transparent and reversible, and downstream processing can be driven by structured manifests rather than ad-hoc manual tracking.

Source	Current evidence base	In-window titles	Main open issue
Europeana	69,290 Hungarian-language text items harvested via API	discovery set only	newspaper-specific filtering and metadata cleanup
OSZK–EPA	manual survey of regional, cross-border and newspaper collections	to be cleaned	overlap resolution and automated harvesting
DiFMOE	eight Hungarian-language periodicals entered in inventory with URLs	7	mostly manual acquisition, then OCR/TEI conversion
Hungaricana	curated manual-download list with title-level URLs, institutions and selected coverage periods	12	issue counting and download workflow still manual

Table 1: Current status of the main scouting tracks for Hungarian historical newspapers.

Title	Holding institution	Available years	Selected target coverage
Békésmegyei Közlöny	Békés Megyei Könyvtár	1877–1938	1877–1913
Dunántúli Protestáns Lap	Jókai Mór Városi Könyvtár (Pápa)	1890–1945	1890–1915
Eger – hetilap	Bródy Sándor Megyei és Városi Könyvtár (Heves megye)	1863–1914	1863–1914
Esztergom	Helischer József Városi Könyvtár (Esztergom)	1895–1932	1895–1915
Esztergom és Vidéke	Helischer József Városi Könyvtár (Esztergom)	1879–1944	1879–1915
Felsőmagyarországi Hírlap	MNL BAZ Megyei Levéltárának Sátoraljaújhelyi Fióklevéltára	1903–1917	1903–1907
Független Budapest (Az Erzsébetváros)	Erzsébetváros Önkormányzata	1906–1938	1906–1915
Nyírvidék	Móricz Zsigmond Megyei és Városi Könyvtár (Szabolcs-Szatmár-Bereg)	1867–1942	1867–1915
Szatmári Értesítő	Móricz Zsigmond Megyei és Városi Könyvtár (Szabolcs-Szatmár-Bereg)	1862	1862
Váci Hírlap	Katona Lajos Városi Könyvtár (Vác)	1887–1942	1887–1915
Váczi Közlöny	Katona Lajos Városi Könyvtár (Vác)	1881–1895	1881–1895
Zemplén	MNL BAZ Megyei Levéltárának Sátoraljaújhelyi Fióklevéltára	1886–1937	1886–1915

Table 2: Curated Hungaricana manual-download pilot list

4. Repository, conventions and validation

The project already has an enforceable engineering substrate. The `pressmint-hu` repository contains: (i) CI validation (`.github/validate-inventory.yml`), (ii) documentation (`docs/hu_source_inventory.ods`, `docs/conventions.md`), (iii) an inventory schema and controlled vocabularies (`inventory/schema.json`, `inventory/sources.yaml`, `inventory/vocabularies.yaml`), (iv) a TEI header template (`inventory/tei_header_template.xml`), and (v) scripts for crawling (Europeana, Hungaricana collection-name extraction) and utilities (IDs, TXT→ODS import, YAML→TEI export. The repository also stores intermediate scouting artefacts (e.g. `hu_europena_corpus.json`, `hungaricana_collections.txt`).

This structure operationalises interoperability in three ways:

- **Deterministic IDs:** title, issue and page identifiers are generated from inventory fields, which keeps them stable across rebuilds.
- **Constrained metadata:** controlled vocabularies reduce drift in source descriptions, access methods, scan formats and OCR flags.
- **TEI by construction:** TEI headers are generated from structured YAML, which reduces

manual editing and supports validation-driven release builds.

The Hungarian workflow also aligns with the broader PressMint logic. Upstream steps such as source acquisition, OCR and some normalisation details may differ by national partner, but the target representation, mandatory metadata, identifier logic and facsimile linkage are shared. This allows local heterogeneity in acquisition while keeping downstream comparison possible.

The repository structure is designed so that schema files, vocabularies and conversion code can be released publicly with minimal changes. Public release of every project artefact, however, will depend on cleaning project-internal notes and checking source-specific redistribution constraints.

5. Processing pipeline: ingestion, OCR and TEI conversion

5.1. Ingestion and normalisation

Ingestion starts with a provider-specific acquisition step that yields a local artefact (typically PDFs or page images) plus a manifest entry in the inventory. The local storage layout is organised by Source/Title/Year, mirroring the inventory keys and making partial re-ingestion possible when target coverage changes.

For near-term work, manually tractable sources are the most realistic starting point. DiFMOE provides a small set of periodicals with item-level URLs already recorded in the inventory. Hungaricana now provides a second concrete pilot track through the curated 12-title manual-download list. In parallel, Europeana remains a metadata-first discovery track: it helps us locate candidate titles and estimate coverage, but acquisition feasibility still has to be checked provider by provider.

5.2. OCR and layout analysis: from prototype to evaluation

Our OCR strategy is driven by the fact that newspaper layout is not a marginal issue but a central source of downstream error. An initial prototype log, based on a single test page image, showed that faint column separators can break automatic segmentation and thereby damage reading order. Although this first test page lies outside the final PressMint time window, the lesson is still relevant: for historical newspapers, OCR must be evaluated as a *layout-aware pipeline*, not only as plain text recognition.

At the current stage we do not commit to a single OCR engine. The material is too heterogeneous for premature tool lock-in, and the key question is not only character recognition accuracy but also preservation of columns, reading order and article boundaries. We therefore plan a small but systematic evaluation on manually corrected pilot pages. The evaluation will compare at least a page-level OCR baseline with a layout-aware pipeline and will measure: (i) character and word error rates on sampled pages, (ii) preservation of column structure and reading order, (iii) systematic error classes particularly relevant for Hungarian newspaper print, such as diacritics and line-break hyphenation, and (iv) robustness across different scan qualities and page layouts.

The pipeline itself is organised in four steps: (i) page image extraction and normalisation, (ii) layout analysis to detect columns and reading order, (iii) OCR on layout-aware regions, and (iv) conversion into PressMint-compatible TEI with explicit facsimile linkage. This architecture keeps open the possibility of article- or section-level segmentation where the layout evidence is strong enough, without making it a hard requirement for the first release.

5.3. Quality assurance and comparability

To ensure that the Hungarian tranche remains comparable across PressMint partners, we plan quality monitoring at three levels:

- **Metadata QA:** validation against controlled vocabularies and mandatory fields such as title, years, place and identifiers.
- **OCR QA:** periodic sampling with character error rate estimation and tracking of systematic error classes such as hyphenation, diacritics and ligatures.
- **Structural QA:** validation that each TEI document preserves facsimile linkage and that identifiers remain stable across rebuilds.

Where feasible, a small manually corrected gold set will be prepared to calibrate OCR and layout components, following broader recommendations for multilingual and historical OCR research agendas (Smith and Cordell, 2018).

5.4. TEI conversion and optional stand-off layers

Following PressMint interoperability goals, TEI is the hub representation for text, metadata and provenance. We implement TEI generation in two layers:

1. **Guaranteed page-level TEI** (issue $\rightarrow$ pages) with stable identifiers and facsimile pointers.
2. **Optional article or section segmentation** when layout cues allow reliable inference.

To support comparative NLP while preserving reprocessability, linguistic annotations (tokenisation, sentence boundaries and, optionally, POS or lemma information) will be kept as stand-off layers whenever feasible, so that OCR and normalisation updates do not invalidate previously released base TEI.

6. DH use cases enabled by the Hungarian tranche

The Hungarian PressMint tranche is meant to support both close and distant reading. Concrete use cases include: (i) tracing the vocabulary of industrialisation, public health or administration across the long 19th century; (ii) comparing how the same events were reported in Budapest and in cross-border or regional newspapers; (iii) studying lexical and stylistic differences across publication places such as Esztergom, Kassa, Nyíregyháza, Sátoraljaújhely or Vác; and (iv) tracking named entities, institutions and quoted actors in a way that remains anchored in facsimile-linked evidence.

These use cases motivate our emphasis on provenance-rich TEI for citation and on an OCR/layout pipeline that preserves reading order and, where possible, section boundaries.

7. Deliverables and roadmap

The immediate goal is a Hungarian release candidate that (i) fits the 19th/early-20th century time window and (ii) can be rebuilt reproducibly from inventory manifests. Near-term deliverables for 2026 are:

- a cleaned, deduplicated and newspaper-focused Europeana-derived candidate list starting from the harvested 69,290 Hungarian-language text items;
- a first TEI pilot release for the seven in-window DiFMOE titles and a pilot subset of the 12 curated Hungaricana titles, all with facsimile-linked pages;
- a documented OCR and layout benchmark on manually corrected sample pages, including a sampling protocol for error analysis and ground-truth creation;
- a harmonised ingestion and export toolchain (YAML→TEI) integrated into repository validation.

8. Conclusion

The Hungarian PressMint subcorpus is being built around explicit engineering artefacts – a structured inventory and a validation-driven repository – and around the assumption that OCR and layout are central interoperability constraints, not secondary implementation details. Current scouting shows a fragmented but usable landscape. Europeana is effective for large-scale discovery, while DiFMOE and the curated Hungaricana pilot list provide concrete, manually tractable starting points for acquisition. By treating selection, ingestion, OCR and TEI conversion as reproducible and testable processes, we aim to deliver a Hungarian tranche that is aligned with PressMint-wide requirements and useful for both close and distant reading research.

9. Bibliographical References

Arcanum. n.d. Arcanum newspapers. https://adt.arcanum.com/. Accessed: 2026-03-30.

Austrian National Library. n.d. Historical newspapers and periodicals of the austrian national library. https://www.onb.ac.at/en/departments/department-of-manuscripts-and-rare-books/holdings/rare-books/historic-newspapers-and-magazines. Accessed: 2026-03-30.

Digital Forum Central and Eastern Europe. 2023. Digital library of the digital forum central and eastern europe. https://www.copernico.eu/en/online-resources/digital-library-digital-forum-central-and-eastern-europe. Accessed: 2026-03-30.

Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströbel, and Raphaël Barman. 2020. Language Resources for Historical Newspapers: the Impresso Collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 958–968, Marseille, France. European Language Resources Association.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Çağrı Çöltekin, Tommaso Agnoloni, Orsolya Ring, et al. 2023. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57:415–448. Published online 2022.

Nuno Freire, Antoine Isaac, Twan Goosen, Daan Broeder, Hugo Manguinhas, and Valentine Charles. 2019. Opening Digitized Newspapers Corpora: Europeana's Full-Text Data Interoperability Case. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *Open Access Series in Informatics (OASIcs)*,

pages 22:1–22:14, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Hungaricana. n.d. Library | hungaricana. https://library.hungaricana.hu/en/. Accessed: 2026-03-30.

Kimmo Kettunen, Heikki Keskustalo, Sanna Kumpulainen, Tuula Pääkkönen, and Juha Rautiainen. 2022. OCR quality affects perceived usefulness of historical newspaper clippings – a user study. arXiv:2203.03557.

Günter Mühlberger. 2014. METS/ALTO Profile (ENMAP) – Deliverable D5.2 (Draft). Technical report, Europeana Newspapers Project. Accessed 2026-03-03.

Christa Müller. 2004. A N N O – Austrian Newspapers Online: Historische österreichische Zeitungen und Zeitschriften online. Eine Digitalisierungsinitiative der Österreichischen Nationalbibliothek. In Hartmut Walravens, editor, *Newspapers in Central and Eastern Europe / Zeitungen in Mittel- und Osteuropa: Papers presented at an IFLA conference held in Berlin, August 2003*, pages 141–148. K. G. Saur, Berlin and Boston.

Clemens Neudecker and Apostolos Antonacopoulos. 2016. Making Europe's Historical Newspapers Searchable. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 405–410, Santorini, Greece. IEEE.

Országos Széchényi Könyvtár (OSZK). 2026. EPA – Elektronikus Periodika Archívum. https://epa.oszk.hu/. Accessed 2026-03-03.

David A. Smith and Ryan Cordell. 2018. A Research Agenda for Historical and Multilingual Optical Character Recognition. Technical report, Northeastern University, NULab for Texts, Maps, and Networks. Accessed 2026-03-03.

Marieke Willems and Rossitza Atanassova. 2015. Europeana Newspapers: searching digitized historical newspapers from 23 European countries. *Insights*, 28(1):34–39.