Toward Interoperable and Scalable Representations of Complex Heterogeneous Digitized Historical Media

Pauline Conti[1], **Simon Clematide**[2], **Maud Ehrmann**[1]
[1]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[2]University of Zurich (UZH), Switzerland
<firstname>.<lastname>@epfl.ch, <firstname>.<lastname>@uzh.ch

Abstract

The value of digitized historical media archives for computational historical research is now well established, yet an underexplored challenge concerns data management itself: how to represent and process, at scale, complex primary sources that vary widely in digitization granularity, refinement quality, and archival organization and curation practices. This paper presents the data representation framework designed for large-scale processing and indexing of historical newspapers and radio broadcasts developed within the Impresso project. Grounded in a structured characterization of the heterogeneity found in digitized historical media collections, it identifies the distinct dimensions along which collections diverge and the challenges they pose for a unified representation and processing framework. The framework navigates the competing demands of machine learning pipelines requiring uniform and lightweight document representations, information retrieval systems requiring well-defined indexable content units, user-facing interfaces requiring fidelity to original sources, and the need to return semantically enriched data to archival holders in interoperable formats. We describe the design principles guiding the framework and discuss how it reconciles these constraints across highly heterogeneous collections into a unified and research-ready corpus.

Keywords: digitized newspapers, digitized broadcasts, data representation and processing, heterogeneity and interoperability, machine learning, information retrieval

1. Introduction

The digitization of historical media collections has accelerated considerably over the past two decades, producing vast repositories of machine-readable content from newspapers and, increasingly, radio broadcasts (Balk and Conteh, 2011; Neudecker and Antonacopoulos, 2016). Held by libraries and cultural heritage institutions across Europe and beyond, these collections have opened new avenues for historical research: from full-text search and browsing to semantic enrichment and, more recently, to semantic indexing enabling similarity-based exploration across large corpora (Neudecker, 2022; Düring et al., 2023). The historical and scholarly value of such resources is now widely recognized (Bunout et al., 2023).

Yet this value is greatly amplified when research can be conducted at scale, across institutional, linguistic, and national boundaries. Historical newspapers and broadcasts are, in practice, fragmented across institutional silos: digitized collections are typically bound to a single institution, language, or country, maintained in distinct formats, and designed primarily for preservation and consultation rather than programmatic processing and analysis. This fragmentation makes large-scale processing difficult, hinders the construction of longitudinal research datasets, and severely limits interoperability across collections (Padilla, 2019; Ehrmann et al., 2023a).

Bridging these silos is one of the main objectives of the 'Impresso - Media Monitoring of the Past' project, which focuses on processing, enriching, and enabling the exploration of large-scale historical media sources to increase their accessibility and usability for digital historical research[1]. In its second phase, the project pioneers the joint exploration of Western European newspapers and radio content across temporal, linguistic, and national boundaries, drawing on collections from more than 20 partner institutions. It develops and applies machine learning-based text and image mining approaches – including named entity recognition, topic modeling, text reuse detection, and embedding-based similarity search – to enrich and index a transnational corpus of facsimiles, OCR and ASR transcripts, and associated metadata. The resulting enriched corpus is made accessible through a graphical web application, the Impresso WebApp, and a programmatic access ecosystem, the Impresso Datalab, to support both human-centred and data-driven historical inquiry with transnational and transmedia perspectives.

Beyond well-documented challenges such as OCR noise, the difficulty of applying NLP to historical text, and digitization bias in collections (van Strien et al., 2020; Ehrmann et al., 2023b; Beelen et al., 2025; Opitz et al., 2026), a less frequently examined challenge concerns the technical dimension of data preparation and collection management. Difficulties arise at three levels: at the level of the historical source, newspapers are complex

[1] https://impresso-project.ch

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 72–81
May 16, 2026. ©2026 Association for Computational Linguistics

SOURCES	*Media*	NEWSPAPERS (INCLUDING RADIO MAGAZINES)		RADIO BROADCASTS	
	Medium	Print		Typescripts	Radio records
INPUTS	*Modality*	Image	Text (OCR)	Text (ASR)	Audio
	Metadata	Publication year, place, publisher, author, size or length, copyright, ...			
	Language	—	Dutch, English, German, French, Luxembourgish		
PROCESSING	*Semantic enrichment*	Article and images alignment, image classification	Semantic segmentation, NERC, EL, Keyphrases, Topics, Classes, Opinion, Content reuse		—
	Semantic Indexing	Dense vector representation	(Clustered) multilingual dense vector representation of text and enrichments		—
ACCESS	*Content Retrieval*	Visual search and text search (captions, related articles)	Cross-lingual faceted text search and exploration of enrichments		—
	Displayed Objects	Images	Images & Text (OCR)	Text (ASR)	Audio streams

Table 1: Alignment of sources with the types of input they correspond to, the processing they undergo, and the search and rendering modes supported by the interface.

objects whose value lies precisely in their material and editorial structure: a heterogeneous mix of text, images, tables, and graphical elements, organized across issues, pages, and articles in ways that are historically meaningful. Radio broadcasts share a similar complexity, albeit in less documented and more irregular ways, raising non-trivial questions about how the two media can be aligned and compared. At the level of the digitized record, this complexity is inherited and extended: a digitized newspaper content item is not simply raw text, but a layered object comprising facsimiles, OCR transcripts, layout segmentation, content organization, and bibliographic metadata. At the level of the collection, finally, decades of digitization campaigns across institutions compound the difficulty further: collections differ in file formats, processing granularity and quality, and archival organization. All of this is further heightened by the volume of data involved, the requirements of large-scale machine learning and information retrieval, and the need to remain responsive to how historians work with primary sources.

This raises the question: How can complex historical image-text objects, heterogeneous in origin and digitization practices, be represented, processed and indexed at scale without losing what makes them meaningful as historical sources? This paper presents a data representation model and conversion architecture developed within Impresso, grounded in explicit design principles and a structured characterization of collection heterogeneity. The framework reconciles documentary fidelity with machine learning and information retrieval requirements across heterogeneous historical newspapers and broadcast sources, and is validated on a large transnational multilingual corpus.

The rest of this paper is structured as follows. Section 2 outlines the design principles that shaped the framework, Section 3 details the types of heterogeneity we face with such collections, Section 4 reviews existing representation formats, Section 5 describes the framework, and Section 6 discusses and concludes.

2. Design Principles and Requirements

Impresso collects 500+ digitized newspapers and radio sources from libraries and archives, processes them through a semantic enrichment pipeline, and makes the resulting data accessible via two interfaces. Table 1 gives an overview of the media sources, their input types, and the processing and access scenarios they undergo, from visual search over newspaper images to cross-lingual faceted search over enriched transcripts.

The raw input consists, for newspapers, of page facsimiles, OCR transcripts organized as word- or region-level bounding boxes, and layout segmentation when available; for audio sources, of recordings and ASR transcripts. Both are accompanied by bibliographic metadata, which is not considered further in this paper.

The stewardship of such a corpus — that is, its representation and manipulation at scale — is a complex undertaking, shaped by application-specific requirements from the Impresso context and guided by the FAIR principles (Findable, Accessible, Interoperable and Reusable), with which our framework strives to comply. We describe these requirements in turn below.

Breaking Silos: Heterogeneous Collections at Scale The first guiding principle is to break collection silos, of which we identify three kinds. *Media*

silos: historical newspapers and radio broadcasts are inherently preserved in separate collections, reflect different production and consumption practices, and follow different archival logics — yet both are historical media sources worth studying in conjunction, and our framework must accommodate the representation of both. *Digitization silos*: each institution has conducted its own digitization campaigns over the past thirty years, resulting in collections that differ in file formats, processing granularity — from raw OCR output to fine-grained article-level segmentation — quality, and archival organization. *Language silos*: the multilingual nature of the corpus introduces tokenization differences and variable accuracy of automatic language identification, both of which affect downstream text processing. Integrating these heterogeneous collections into a unified corpus is one of the primary drivers of our framework's design, detailed in Section 3.

Fidelity to the Source A second requirement concerns fidelity to the source. Historical newspapers are not merely containers of text but also a structured arrangement of articles, advertisements, images, tables, and other content types across pages and issues, reflecting editorial choices and historical contexts. Radio broadcasts similarly carry meaning through their program structure, sequencing, and temporal organization. A user navigating the Impresso interface should be able to read an article in the context of its page or a broadcast segment within its program, access the original facsimile or audio recording, and situate it within the broader collection. This requires that our data representation maintain the logical structure of each source along with the coordinates – spatial for newspapers, temporal for radio – linking transcripts back to the original medium.

Amenability to Machine Learning and Information Retrieval Enabling large-scale semantic enrichment and indexing requires a data representation that is uniform and lightweight, properties that are in tension with the source fidelity requirement. For machine learning (ML), data must be prepared in a format that is stripped of archival and layout overhead, and in which text content is reconstructed as running text rather than the word-by-word output of raw OCR. This reconstructed text is what the enrichment pipeline operates on. A complementary requirement concerns the identification of the canonical unit of indexing: what constitutes a document for the purposes of the search engine. Ideally, this unit should follow the natural structure of the source media (an article for newspapers, a broadcast episode for radio) but such document-level segmentation is not always available and must sometimes be approximated from coarser representations.

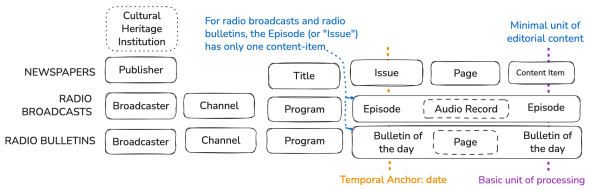

Figure 1: Newspaper and radio source alignment.

Corpus Growth and Framework Modularity Next, the Impresso corpus is not static: new collections are added regularly, and new source types may be integrated as the project evolves. This expectation of growth requires a data representation that is unified yet flexible enough to accommodate new collection variants. First, it must provide a clear and stable scheme for assigning identifiers to each element across the corpus, so that partial updates, such as the re-OCRisation of an existing collection, do not destabilize the broader corpus. Second, it must support a consistent collection organization – a unified file structure, a rigorous protocol for integrating new collections, and a principled versioning scheme covering both staging and releases – allowing new sources to be onboarded without disrupting the existing structure.

Returning Enriched Data to Archival Holders A final principle concerns the return of enriched data to partner institutions. Having provided their collections for processing, institutions expect to receive back the semantic enrichments — named entities, topics, embeddings, and other annotations — in formats compatible with their own systems. This requires that our representations maintain clear links to each institution's original identifiers, directly reinforcing the interoperability and reusability dimensions of FAIR.

Collectively, these principles define a design space shaped by three partly competing demands: fidelity to the source, interoperability, and ML and IR efficiency — respectively requiring preservation of source structure, stable and standardized representations, and lightweight uniform text optimized for large-scale processing. Yet these are requirements set against a complex material reality, which in practice must be understood before it can be managed.

3. A Typology of Heterogeneity in Digitized Historical Media

At scale, heterogeneity is the most pervasive characteristic of digitized historical media collections.

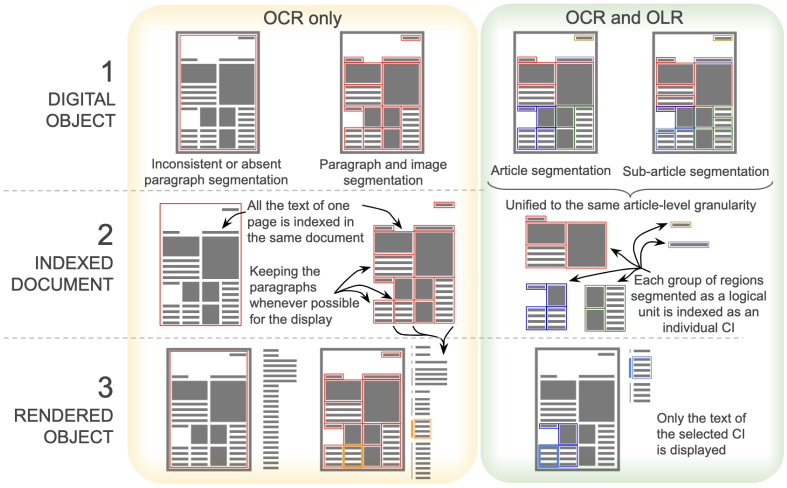

Figure 2: Illustration of the various granularity levels of the OCR and OLR structure present, and how they are indexed and rendered in the Impresso App.

Mapping these dimensions is a prerequisite for organizing the data, as it reveals what must be accounted for. We identify four such dimensions: media sources, digitized records, collection organization and identification, and language and tokenization.

3.1. Heterogeneity of Media Sources

The Impresso corpus spans two broad categories of media sources differing in medium, modality, and archival structure, requiring shared abstractions.

Sources differ in their medium and modality. Newspapers and radio bulletins[2] are print or typescript-based, and their text content is extracted via OCR, sometimes complemented by OLR for structure. Radio broadcasts, by contrast, are audio-based, and their text is produced via automatic speech recognition (ASR). These two extraction processes produce outputs in different formats, each requiring its own processing pipeline, and with characteristic error profiles that affect downstream NLP differently.

Structurally, newspapers follow a simple hierarchy: a title publishes issues at regular intervals, each composed of pages and articles. Radio sources are more irregular: broadcast episodes and bulletins group into recurring programs, with topical diversity introduced at the channel rather than the episode level. Figure 1 illustrates the resulting alignment challenges across source types.

Across this diversity, the framework strives for a common minimal unit of editorial content suitable for indexing, and a shared temporal anchor — the day — across all source types.

[2]Radio bulletins are the typescript scripts read on air by radio presenters.

3.2. Heterogeneity of Digitized Records

Beyond media sources, digitized records are layered objects whose structure and quality vary considerably across collections, along three dimensions: format and schema variants, structural refinement and segmentation, and refinement quality.

3.2.1. Format and Schema Variants

The dominant formats for encoding digitized newspaper content are METS and ALTO[3] — METS for the logical structure of an issue, ALTO for the OCR transcription of individual pages. Both are widely adopted standards yet, in practice, each digitization campaign produces its own flavor: slightly different element names, attribute conventions, nesting structures, and levels of detail. In some collections, only ALTO files are present; page-level OCR is available but not logical grouping of content into articles. Other collections provide both, but with varying degrees of completeness. Additional formats encountered include hOCR and PDF-derived text, each with their own structural conventions. Page images, which serve as the visual anchor for all text coordinates, are similarly provided in varying formats and resolutions. This diversity means that each collection — sometimes each sub-collection within a single institution — requires its own preprocessing and ingestion pipeline, and that any unified representation must abstract away from these format variations.

3.2.2. Structural Refinement and Segmentation

Digitization campaigns differ in how much source structure they recover, distinguishing physical lay-

[3]loc.gov/standards/mets/, loc.gov/standards/alto/

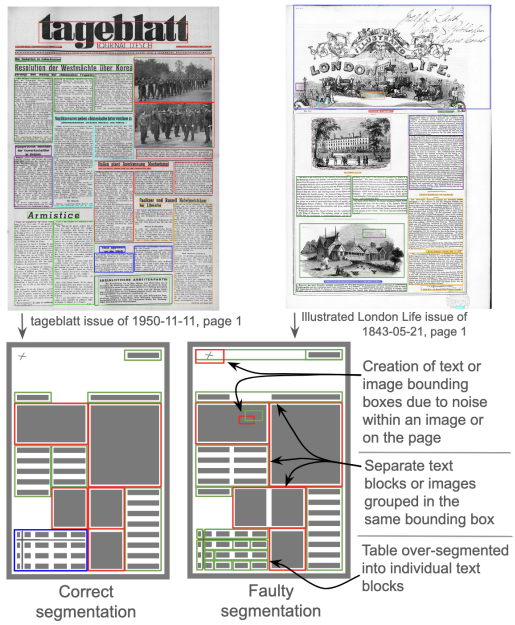

Figure 3: Real and schematic examples of typical OCR/OLR segmentation errors. In the facsimiles, the boxes of one article share their color. In the schemas, colored boxes indicate content type labels: images (red), text (green), tables (blue).

out recognition — identifying page regions such as text blocks, images, and tables — from logical layout recognition, which groups these into meaningful units such as articles or advertisements. We distinguish four levels of increasing structural recovery, illustrated in Figure 2:

No physical layout: OCR embedded in PDFs or older collections where paragraph segmentation is absent or inconsistent, including radio bulletins. *Physical layout only*: text blocks and region boundaries are identified, but not grouped into semantic units; images and tables may be detected, though their relation to surrounding text is not captured. This is the most common level in older campaigns. *Physical and logical layout*: OLR is applied alongside OCR, grouping text blocks into content items — articles, advertisements, obituaries — linking image regions to text, and enabling reconstruction of articles spanning multiple pages. This is the target granularity our framework aims to match. *Extended logical layout*: At the most refined level, content items are further decomposed into labeled sub-elements at paragraph level. Present in only a small subset of collections, the resulting nested structures can be difficult to parse reliably.

Radio sources are generally sparser: logical layout does not apply to audio recordings, and segmentation, when present, reflects program or segment boundaries rather than article-level structure.

The semantic labeling of identified units also varies across collections. Most OLR-processed data features a set of content type labels — article, advertisement, image, table, obituary — but the exact vocabulary, granularity of classes, and consistency of application differ from one campaign to the next. Label recall is often low, particularly when coupled with segmentation errors, and image-caption linking is frequently incomplete.

3.2.3. Refinement Quality

Independently of segmentation level, the quality of digitization outputs varies considerably across and within collections. OCR and ASR quality — affected by image resolution, font style, conservation state, and the age of the material — constitutes a persistent challenge for downstream NLP, though it falls outside the scope of the representation framework itself. OLR quality, by contrast, directly shapes what structural information is available to represent.

OLR errors manifest at the level of region segmentation and classification: columns merged into single blocks, tables hyper-segmented into individual cells, advertisements grouped together, or bounding boxes hallucinated over images or handwritten annotations. Misclassification of content types further reduces the reliability of article-level groupings. Figure 3 illustrates typical examples: the two paragraphs with purple boxes should be green, and the linking between the green-box article and its images and captions was partially lost.

Taken together, these variations mean that a robust framework must provide a unified abstraction over these document element types and granularities, while retaining enough flexibility to account for each collection's specific segmentation characteristics.

3.3. Heterogeneity of Collection Organization and Identification

Each institution organizes its collections according to its own archival logic, resulting in heterogeneous identifier schemes across the corpus. Identifiers exist at each level of the hierarchy — title, issue, page, content item — but vary in scope, permanence, and semantics. Some are persistent and institution-wide, such as ARK IDs resolving to stable URLs[4]; others are internal to a digitization campaign and may be reassigned upon reprocessing. Some encode explicit information — title, date, page number — while others are opaque UUID-like strings requiring standoff metadata.

A more structural dimension concerns the presence of identifiers at the content item level, which is directly tied to segmentation: in the absence of OLR, articles carry no identifiers, creating compatibility and sustainability challenges when segmenta-

[4] arks.org/about/ark-overview/

tion is added later. Such challenges are not merely theoretical: we encountered cases where collections were re-OCRed or re-segmented by partner institutions, causing identifiers to be reassigned and breaking the correspondence with our enrichments.

These variations required careful documentation of each collection's identifier logic before any processing could take place, and informed our own identifier scheme's design, described in Section 5.

3.4. Heterogeneity of Language and Tokenization

A final dimension concerns the textual units produced by OCR and ASR, and the way text is tokenized within them. Each digitization campaign has used different OCR engines and post-processing pipelines, resulting in inconsistent tokenization across collections: word boundaries, hyphenation handling, and the treatment of punctuation — whether encoded as a separate token or not — vary from one collection to the next, and sometimes within the same collection across digitization campaigns.

This variability is further compounded by the multilinguality of the corpus. Each language carries its own orthographic conventions, and the accuracy of language identification — a prerequisite for applying language-specific processing — is itself variable across collections and historical periods. Older sources in particular may use archaic spelling, ligatures, or non-standard character encodings that challenge both language identification and tokenization.

Reconstructing running text from raw OCR output — a requirement for downstream NLP, as discussed in Section 2 — therefore requires language-dependent normalization rules that account for these variations: resolving hyphenation across line breaks, reattaching punctuation, standardizing encoding, and handling script-specific conventions. These rules must be defined and maintained per collection and per language.

Before introducing our data representation model, we briefly survey existing frameworks.

4. Related Work

The *Text Encoding Initiative* (TEI) P5 is a widely adopted XML standard for richly structured and semantically expressive text encoding in the humanities. TEI schemas define an extensive set of elements for capturing detailed structural, editorial, and semantic markup, and projects typically customize the base TEI schema via ODD (One Document Does It All) to match corpus-specific needs (Consortium, 2025; Cummings, 2019). The PressMint initiative, a flagship effort of the CLARIN infrastructure, applies a customized TEI P5 schema to compile interoperable corpora of newspapers, from which multiple downstream formats (e.g., CoNLL-U, JSON) are derived for analysis and tooling (PressMint, 2026).

Models based on TEI P5, such as PressMint, prioritize semantic richness and hierarchical text representation that supports detailed editorial tasks and scholarly interoperability across languages and national contexts. However, because TEI encapsulates a large, optional element space, fully utilizing TEI in computational pipelines often requires extensive schema profiles and transformations to extract *fixed-granularity*, machine-ready units suitable for indexing or machine learning tasks. Customization of TEI vocabularies further implies that distinct corpora may diverge in practice unless governed by shared profiles prior to conversion.

In contrast, our JSONL-based representation defines a flat, uniform, processing-ready document abstraction in which heterogeneous canonical metadata and text segments are systematically consolidated into consistent records optimized for large-scale indexing and downstream machine learning workflows. This design aligns with engineering requirements for scalable processing, trading off some of the expressivity and embedded semantic nuance of TEI's richly structured models in favor of simplicity, consistency, and integration with modern data science tooling.

5. Proposed Framework

The design principles and heterogeneity typology outlined above guided the development of our data representation framework. While it does not resolve every challenge, it provides a robust foundation that accommodates evolving data and user needs.

5.1. Conceptual Model: Two Complementary Representations

Impresso's data representation is based on two complementary data structures which, together, hold the physical and logical content of the source material, each addressing a specific design constraint mentioned in Section 2, namely format uniformity and fidelity to the source.

Canonical Format The canonical format is the first and most source-faithful of the two representations. Its primary objective is to bring all incoming data — regardless of origin, format, or source type — into a single unified representation, retaining only information relevant to the source's logical structure and layout. It is composed of two conceptual data

objects: *Issues* and *Physical Supports*, the latter instantiated as pages or audio records.

Issue objects aggregate all relevant information about their structure: the physical supports they rely on, the list of content items they comprise — editorial units below the page level such as articles, advertisements, images, tables, obituaries, and other content types — and technical metadata. They are uniquely identified by their media title, publication or airing date, and edition, the latter distinguishing multiple editions published on the same day.

Initially based on the newspaper archival object, the issue representation now accommodates other source types — radio broadcasts and bulletins — as illustrated in Figure 1. Mapping radio content to this structure raised the question of what constitutes a radio equivalent to a newspaper issue. Two options were considered: grouping all broadcasts of a given day and channel into a single issue (akin to a daily newspaper edition), or treating each broadcast episode as its own issue. The first option echoes the topical variety and publication regularity of a newspaper issue, while the second reflects the grouping of broadcasts into thematic programs, which become equivalent to newspaper titles. The latter was favored as it better reflects radio's archival and thematic organization.

Physical support objects — pages or audio records — establish the link between the abstract representation and the digitized source (page image or audio record file). Page objects contain basic metadata along with bounding-box coordinates, textual content, and content-item affiliation at region, paragraph, line, and token level. Audio record objects follow the same logic, with timestamps for each speech segment and, where available, speaker turns.

Together, these two object types reduce the storage overhead of layout information while maintaining a close connection to the source and its archival structure, providing a consistent representation across source types for subsequent pipeline steps and interface display.

Rebuilt Format The rebuilt format assembles, for each content item, a self-contained and ML-ready document representation (a content item). Metadata is drawn from the issue object, while full-text content is reconstructed as running text from the token-level bounding boxes or timestamps present in the physical support objects, along with text offsets marking line, paragraph, and region boundaries. The result is a uniform, lightweight media document that abstracts away from source-specific formatting and is ready for downstream processing.

Like the canonical format, the rebuilt format accommodates both page-based and audio-based content items. It constitutes the basic unit of processing for all pipeline steps — semantic enrichment, embedding, and indexing — and provides the fixed granularity level at which content is indexed for information retrieval.

5.2. Design Choices: Addressing the Requirements

Systematically Qualifying the Media Sources To characterize the sources beyond the simple newspaper/radio binary distinction – which does not fully capture their diversity – we define two properties for all data objects: source type and source medium.

Source *type* refers to the specific media of the source (labels on the left of Figure 1), and allows to distinguish between the different types of radio sources. Source *medium* refers to the format in which the source was originally produced — print, typescript, audio recording — as listed in Table 1.

Together, these two properties classify all sources into a fixed set of scenarios that inform how our processing and interface display should adapt to each source. Based on source type and medium values, the set of required and expected attributes in each data object shifts slightly, allowing the processing pipeline to adapt dynamically to each format's specific requirement.

Overall, the attributes of each data object fall into two categories: those required to uniquely identify a document — such as media title and date — and categorical attributes that drive dynamic pipeline adaptation, including source type, source medium, and whether OLR was performed. This defines a shared information structure across all sources, making the framework robust to data heterogeneity.

OCR-Only Data: Pages as Content Items The left part of Figure 2 illustrates the case where OLR was not applied to a collection. As discussed in Section 3, no systematic means exists to determine which paragraphs and images belong together in such cases. The entire text of a given page is therefore treated as a single content item, assigned the type "page" — as opposed to "article" or other semantic types — to indicate that it does not represent a semantically coherent unit of content. These content items are processed in the same way as individually segmented items, and any available paragraphs are rendered in the interface accordingly.

An analogous situation arises for radio sources. ASR transcription does not contain audio chapters or segment boundaries equivalent to newspaper article segmentation: the full transcript of a given audio recording constitutes a single content item for its issue. Similarly, a radio bulletin issue contains

Object	Impresso ID
Issues	`[alias]-[YYYY]-[MM]-[DD]-[ed]`
Pages	`[alias]-[YYYY]-[MM]-[DD]-[ed]-p[page #]`
Audio Records	`[alias]-[YYYY]-[MM]-[DD]-[ed]-r[record #]`
Content Items	`[alias]-[YYYY]-[MM]-[DD]-[ed]-i[ci #]`

Table 2: Impresso identifiers for each format.

a single content item, though it may span multiple page supports.

These cases demonstrate the flexibility of the framework in accommodating sources where segmentation is absent or incomplete. Instead of imposing a single model, it treats segmentation as an issue-level variable while preserving a consistent representation.

Constructing Parsable and Deductible Identifiers Another design choice concerns the identifier scheme assigned to each object in the framework. Identifiers must be unique, parsable, and deductible from basic metadata, establishing a stable mapping between the main representation backbone (issues, physical supports, content items) and other data management components, such as bibliographic metadata.

Each media title is assigned a unique alias, sometimes inherited from the original collection. As shown in Table 2, issue identifiers are composed of the media alias, publication or airing date in year-month-day format, and edition (ed) number. Page, audio record, and content item identifiers are derived by appending a type-specific suffix to the issue identifier, followed by a zero-padded four-digit index indicating the object's position within the issue.

5.3. Implementation

Data Represented Through JSON Schemas All object representations are defined and validated through JSON schemas[5], ensuring that generated data complies with the framework's requirements and specifications. Individual documents are aggregated into JSON-line file archives by title and year. The schemas are publicly available in a GitHub repository[6] and help ensure that the framework remains stable throughout each step of the pipeline.

Combined with the identifier scheme, this supports flexible updates and partial re-runs, fine-grained monitoring of collection statistics, as well as the identification of data inconsistencies or leaks within the pipeline.

Format Converters and Module Architecture In terms of implementation, the publicly available code is separated into two Python submodules[7] : one producing the canonical data from each input format – the *importers*, and one extracting the content items from the canonical format to create our document representation – the *rebuilder*.

The main complexity of the importer module lies in adapting to each input format. It is built around abstract classes for issue and physical support objects, which are then extended as format-specific implementations (classes). The abstract classes ensure that all functions required by the main conversion script are consistently defined, while format-specific classes adapt to each collection format's particularities – for instance, handling the METS/ALTO files of the Berlin State Library differently from the PDF-embedded OCR of the SwissInfo collection. All generated issue and page objects are validated against the corresponding JSON schema.

The rebuilder module operates on the unified canonical representation and focuses on the reconstruction of content items, in a similar way across all collections. Constructing content items in these two steps enables to apply the same processing logic to all data, since the input to the second step is already a unified representation produced by our own processing. This promotes uniformity among the content items – the core unit of information retrieval – while keeping the representation lightweight and independent of each collection's specificities.

Object	2025	2026 (in progress)
Media titles	134	567
Issues	780 186	1 021 869
Pages	7 483 588	9 094 381
Content Items	52 358 158	73 475 049
Images	4 002 089	5 589 521
Tokens	15 652 402 700	$>24 \times 10^9$

Table 3: Impresso Corpus statistics: available in the Interfaces (2025) and in preparation (2026).

Large-scale Processing and Downstream Use The large-scale processing and downstream use of the Impresso corpus provides a concrete illustration of the framework in practice.

Through the canonical format, heterogeneous collections from multiple institutions have been converted into a unified representation, enabling the tenfold growth and diversification of the corpus since the framework's deployment in the first iteration of the project. As shown in Table 3, the 2025 release comprised 134 newspaper titles from 9 institutions, amounting to 52 million content items[8]. Several large collections are currently being prepared for iterative release throughout 2026 and the next release will span collections from 15 institu-

[5] json-schema.org

[6] github.com/impresso/impresso-schemas

[7] github.com/impresso/impresso-text-acquisition

[8] github.com/impresso/impresso-data-release

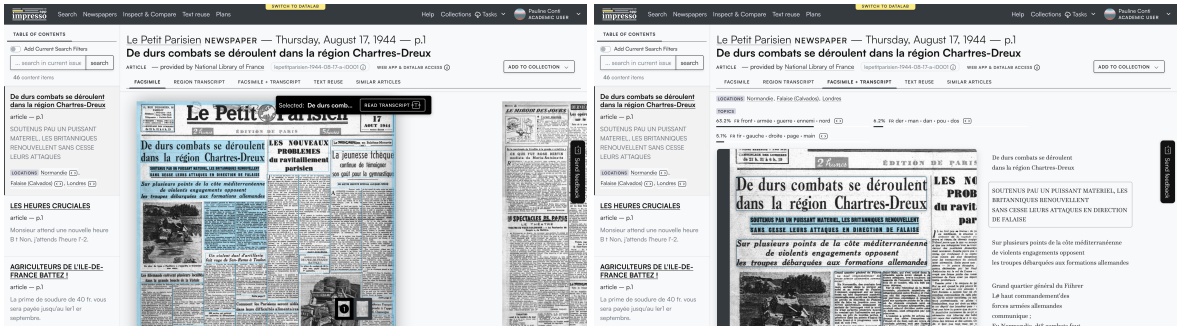

Figure 4: The Facsimile and Transcript views for the same article (`lepetitparisien-1944-08-17-a-i0001`) in the Impresso App, in the case of OCR and OLR data as described in Figure 2.

tions.

The rebuilt format serves as the basic unit of processing and as input to all downstream enrichment steps. This lightweight and uniform representation of content items enables parallelized computation and the application of ML models to an ever-growing collection. Enrichments currently produced include language identification, OCR quality assessment, key phrase extraction, named entity recognition and linking, news agency recognition, topic modeling, text-reuse detection, word and text embeddings, multimodal image embeddings, and image classification.

The resulting enriched transnational, transmedia and multilingual corpus is indexed and made accessible through two interfaces, whose full description lies beyond the scope of this paper. The WebApp offers keyword search, semantic faceted filtering, embedding-based retrieval, comparative and corpus views, and much more. Figure 4 illustrates two source views: the Facsimile view allows navigation between the pages of an issue and selection of an article of interest, while the Transcript view displays the corresponding text alongside bounding boxes on the facsimile. The Datalab offers programmatic access to the corpus, semantic enrichments and models via the Impresso Public API, and hosts notebooks guiding users in the use of these resources. Together, these interfaces represent the foremost output of the project, through which the public and scholars can access data across media, language, and institutional borders.

Crucially, these developments would not have been possible without the foundational work of the data representation framework described in this paper. It is what makes the corpus unified, indexable, and processable at scale: its modular design supports the progressive integration of new collections and source types into a single coherent representation, its lightweight and uniform content items enable large-scale processing and indexing, and its stable identifier scheme accommodates partial updates and re-processing.

6. Discussion and Conclusion

Working with large-scale, heterogeneous historical media collections raises challenges beyond standard data engineering: sources differ in medium, structure, digitization quality, and archival organization, yet must be brought together into a unified, research-ready corpus. The framework presented here navigates these challenges, reconciling documentary fidelity with machine learning usability and archival interoperability — not as a perfect solution, but as a principled and practical one.

The framework presented in this paper is the outcome of an iterative and multidisciplinary process, carried out at the intersection of digital humanities, natural language processing, information retrieval, and archival science. Its design evolved through continuous dialogue with partner institutions, historians, and engineers, reflecting the complexity that such collaboration entails: priorities shift, new source types emerge, and representational choices made early must be revisited as the corpus grows.

Several limitations remain. OCR and ASR quality set a ceiling on downstream ML performance that no framework can fully overcome — the most it can do is accommodate improved transcripts as better models become available. Document size variation, a consequence of heterogeneous segmentation granularity, affects enrichment consistency. Uneven semantic labeling of segmented content further complicates cross-collection comparison.

This work does not propose a standard, but aims to advance the practical possibility of aggregating and processing historical media data at scale, making explicit the design choices that such an endeavor entails. While archival standards are well established, representation frameworks oriented toward large-scale indexing and ML remain comparatively nascent. As collections grow and computational humanities research increasingly demands large-scale, cross-collection data, principled data representation will remain a prerequisite for historical research at scale.

Acknowledgments

The authors warmly thank Matteo Romanello, who greatly contributed to laying the foundations of this data representation framework during the first edition of the Impresso project. This work has been supported by the Swiss National Science Foundation (grant No. CRSII5_213585) and by the Luxembourg National Research Fund (No. 17498891).

7. Bibliographical References

Hildelies Balk and Aly Conteh. 2011. IMPACT: Centre of Competence in Text Digitisation. In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, HIP '11, pages 155–160, Beijing, China, USA. ACM.

Kaspar Beelen, Jon Lawrence, Katherine McDonough, and Daniel C. S. Wilson. 2025. Whose news? Critical methods for assessing bias in large historical datasets. *Computational Humanities Research*, 1:e8.

Estelle Bunout, Maud Ehrmann, and Frédéric Clavert, editors. 2023. *Digitized Newspapers - A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology*. Studies in Digital History and Hermeneutics. De Gruyter Oldenbourg, Berlin, Germany.

T. E. I. Consortium. 2025. TEI P5: Guidelines for Electronic Text Encoding and Interchange.

James Cummings. 2019. A world of difference: Myths and misconceptions about the tei. *Digital Scholarship in the Humanities*, 34(Supplement_1):i58–i79.

Marten Düring, Matteo Romanello, Maud Ehrmann, Kaspar Beelen, Daniele Guido, Brecht Deseure, Estelle Bunout, Jana Keck, and Petros Apostolopoulos. 2023. Impresso Text Reuse at Scale. An interface for the exploration of text reuse data in semantically enriched historical newspapers. *Frontiers in Big Data*, 6.

Maud Ehrmann, Marten Düring, Clemens Neudecker, and Antoine Doucet. 2023a. Computational Approaches to Digitised Historical Newspapers (Dagstuhl Seminar 22292). *Dagstuhl Reports*, 12(7):112–179.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023b. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, 56(2):27:1–27:47.

Clemens Neudecker. 2022. Cultural Heritage as Data: Digital Curation and Artificial Intelligence in Libraries. In *Proceedings of the Third Conference on Digital Curation Technologies (Qurator 2022)*, volume 3234 of *CEUR Workshop Proceedings*, Berlin, Germany. CEUR.

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

Juri Opitz, Corina Raclé, Emanuela Boros, Andrianos Michail, Matteo Romanello, Maud Ehrmann, and Simon Clematide. 2026. CLEF HIPE-2026: Evaluating Accurate and Efficient Person–Place Relation Extraction from Multilingual Historical Texts. In *Advances in Information Retrieval*, pages 354–363, Cham. Springer Nature Switzerland.

Thomas Padilla. 2019. *Responsible Operations: Data Science, Machine Learning, and AI in Libraries. OCLC Research Position Paper.* ERIC.

PressMint. 2026. PressMint Project — CLARIN Flagship for Multilingual Newspaper Corpora. https://www.clarin.eu/pressmint. Accessed 2026-03.

Daniel van Strien, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the Impact of OCR Quality on Downstream NLP Tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, pages 484–496, Valletta, Malta. SCITEPRESS - Science and Technology Publications.