



HAL
open science

impresso Text Reuse at Scale

Marten Düring, Matteo Romanello, Maud Ehrmann, Kaspar Beelen, Daniele Guido, Brecht Deseure, Estelle Bunout, Jana Keck, Petros Apostolopoulos

► **To cite this version:**

Marten Düring, Matteo Romanello, Maud Ehrmann, Kaspar Beelen, Daniele Guido, et al.. impresso Text Reuse at Scale. 2023. hal-04151808v3

HAL Id: hal-04151808

<https://hal.science/hal-04151808v3>

Preprint submitted on 27 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

impresso Text Reuse at Scale. A Prototype Interface for the Exploration of Text Reuse Data in Semantically Enriched Historical Newspapers.

Marten Düring^{1,*}, Matteo Romanello², Maud Ehrmann³, Kaspar Beelen⁴,
Daniele Guido¹, Brecht Deseure⁵, Estelle Bunout¹, Jana Keck⁶, and Petros
Apostolopoulos¹

¹Luxembourg Centre for Contemporary and Digital History (C2DH)

²University of Lausanne

³École polytechnique fédérale de Lausanne (EPFL)

⁴School of Advanced Study, University of London

⁵Royal Library of Belgium

⁶German Historical Institute Washington

Correspondence*:

Marten Düring

marten.during@uni.lu

2 ABSTRACT

3 Text Reuse reveals meaningful reiterations of text in large corpora. Humanities researchers use
4 text reuse to study e.g. the posterior reception of influential texts, and to reveal evolving publication
5 practices of historical media. This research is often supported by interactive visualisations which
6 highlight relations and differences between text segments. In this paper, we build on earlier work
7 in this domain. We present *impresso* Text Reuse at Scale, the to our knowledge first interface
8 which integrates text reuse data with other forms of semantic enrichment to enable a versatile
9 and scalable exploration of intertextual relations in historical newspaper corpora. The Text Reuse
10 at Scale interface was developed as part of the *impresso* project and combines powerful search
11 and filter operations with close and distant reading perspectives. We integrate text reuse data with
12 enrichments derived from topic modeling, named entity recognition and classification, language
13 and document type detection as well as a rich set of newspaper metadata. We report on historical
14 research objectives and common user tasks for the analysis of historical text reuse data and
15 present the prototype interface together with the results of a user evaluation.

16 **Keywords:** text reuse, historical newspapers, user tasks, scalable reading, data visualisation, comparison, impresso

1 INTRODUCTION

17 Text reuse detection (TRD) is a powerful technique to identify “meaningful reiteration[s] of text, usually
18 beyond the simple repetition of common language” (Romanello et al., 2014). TRD identifies repeated text
19 segments (or *passages*) and groups them automatically into text reuse *clusters*. In the domain of Digital
20 Humanities research, TRD is often used to trace quotations, allusions, and paraphrases. Not only the
21 presence but also the frequency of reuse can be meaningful: the frequency with which a text is quoted
22 by later authors, for example, serves as a proxy for literary or scholarly reception. A popular example
23 of research enabled by TRD is on historical newspaper collections, as demonstrated by projects such as
24 Oceanic Exchanges (Oiva et al., 2020; Keck et al., 2022) and Viral Texts (Cordell, 2015). In both cases,
25 TRD helped to capture prevalent journalistic practices, such as the repurposing and editing of content, or
26 news phenomena, such as the viral circulation of content.

27 The project “*impresso* - Media Monitoring of the Past” (2017-2020)¹ detected text reuse within a
28 corpus of Swiss and Luxembourgish newspapers alongside other forms of semantic enrichment (e.g.
29 topic modeling, named entity recognition, image similarity detection, the detection of content type and
30 language).² Its corpus consists of 76 newspapers published in French, German, and Luxembourgish
31 between 1738 and 2018 and contains ca. 50 million content items³ detected in 5.5 million pages. The
32 *impresso* application⁴ supports historians and other humanities researchers with powerful search, filter, and
33 discovery functionalities for the exploration of the enriched data. It is generic in the sense that it supports a
34 wide variety of different use cases. This includes, for example, advanced search, the visualisation-aided
35 comparison of large user-generated article collections, and the creation of research datasets for further
36 processing outside the application. In addition, *impresso* publishes accompanying datasets in dedicated
37 data repositories⁵.

38 In this paper we present the Text Reuse at Scale interface for the visualisation-aided discovery which
39 will complement the *impresso* application. The new interface facilitates “scalable reading” of text reuse in
40 historical media for historians and other scholars in the humanities. Scalable reading combines close and
41 distant reading. In the case of newspapers, close reading corresponds to either the inspection of individual
42 text reuse clusters and the passages they contain or the study of the articles to which they belong. Distant
43 reading refers to the distributions of text reuse measures, metadata, and semantic enrichments. We describe
44 generic (media) historical research objectives and identify a list of generic tasks for the exploration of text
45 reuse data.

46 The interface design and tasks were partly informed by the outcomes of a two-day workshop organised by
47 the *impresso* team, which brought together a group of 10 researchers (from inside and outside the *impresso*
48 project), including professionals from various disciplines such as design, natural language processing, data
49 science and (media) history. The workshop produced a list of historical research objectives (also in the
50 light of previous work), associated tasks and three interface mockups.

51 The structure of this paper roughly follows our interface creation process. Section 2 positions our work in
52 relation to the state of the art: it introduces the methods and tools for TRD, discusses recent advances and
53 remaining challenges for detecting and exploring text reuse in historical newspapers, and concludes with a
54 brief presentation of the *impresso* text reuse data. Turning to historical research interests in text reuse data,
55 Section 3 focuses on five high-level research objectives in (media) history and associated 11 generic tasks
56 that we identified. Next, Section 4 presents the prototype interface in relation to case studies illustrating
57 specific tasks, and Section 5 reports on the results of an evaluation undertaken by 13 users. Finally, Section
58 6 closes the paper with an outlook on future work.

2 STATE OF THE ART: TEXT REUSE DETECTION IN HISTORICAL TEXTS

59 This section situates our work in the current state of the art in text reuse detection and usage for humanities
60 research. We begin with an overview of tools and methods, continue with current directions in the
61 visualisation-aided exploration of text reuse data, and conclude with a description of TRD in the context of
62 the *impresso* project.

63 2.1 What is text reuse and how is it detected?

64 Methods for TRD are shaped by the disciplines in which they emerged. Since text reuse in literary texts
65 is often more subtle than the mere repetition of words (e.g. in the case of paraphrase, allusion, translation,
66 or parody), researchers strive to go beyond lexical similarities in order to capture affinities in syntax,
67 content, or metrical structure (Büchler et al., 2014; Moritz and Steding, 2018; Scheirer et al., 2016). In the

¹ <https://impresso-project.ch/>

² Further information regarding the enrichment of the newspaper corpus can be found in the project blog (<https://impresso-project.ch/blog/>) and FAQ section (<https://impresso-project.ch/app/faq>).

³ Content items refer to newspaper contents below the page level, such as articles, advertisements, images, tables, weather forecasts, obituaries, etc. For further details see: <https://impresso-project.ch/news/2020/01/23/state-corpus-january2020.html>

⁴ <https://impresso-project.ch/app>

⁵ <https://zenodo.org/communities/impresso/>

68 design of TRACER⁶ Büchler et al. (2014) have addressed this subtlety of text reuse in literary texts by
69 giving users access to a wide array of Information Retrieval (IR) algorithms, as well as direct access to the
70 tool's output at each step of the processing chain. More recent studies have investigated the usefulness of
71 sentence and word embeddings, especially with respect to detecting these more allusive forms of text reuse
72 (Manjavacas et al., 2019; Liebl and Burghardt, 2020), finding that they do not bring substantial advantages
73 over traditional IR techniques.

74 On the other hand, the challenges of detecting text reuse in the newspapers domain are quite different. The
75 substantial amount of OCR noise present in digitised newspapers asks for fuzzy methods that are resilient
76 to differences between two or more copies of the same textual content. Moreover, the scale of materials —
77 with corpora that can be several orders of magnitude bigger than those in the literary domain — led to the
78 development of efficient and scalable methods. As a matter of fact, methods that were developed for TRD
79 in the newspapers domain had to deal with both challenges, namely OCR noise and scalability. Vesanto
80 et al. (2017) adapted the Basic Local Alignment Search Tool (BLAST) algorithm, originally developed
81 for the alignment of biomedical sequences, to the task of character alignment.⁷ An alternative approach
82 to TRD consists in performing alignments between documents at the level of longer sequences of words,
83 a.k.a. n-grams, instead of individual characters. This was the approach followed by Smith et al. (2015)
84 whose TRD algorithm, implemented in the tool *passim*⁸, uses n-gram-based filtering to reduce the number
85 of text passage pairs to compare — thus achieving scalability — and combines it with local and global
86 alignment algorithms to handle gaps and variants in longer sequences of aligned texts.

87 2.2 Interactive visualisations of text reuse

88 Text reuse instances can be visualised, analysed and explored at various levels:

- 89 • *Corpus-level* analysis considers all text reuse instances within a corpus; the size and composition of
90 corpora vary; user-defined collections can also be considered as corpora in their own right. Scalability
91 is a typical challenge for visualisations at this level of analysis.
- 92 • *Document-level* analysis considers all text reuse instances within a single document or across sets
93 of documents; compared to the corpus-level, this level of analysis is more meaningful for longer
94 documents such as entire books or book chapters, but it can be applied as well to shorter documents
95 such as journal articles. When applied across documents, this approach provides insights into the
96 genealogy of texts (multiple versions of the same book, different books that have borrowed from one
97 another).
- 98 • *Cluster-level* analysis considers one single instance of text reuse, with a specific focus on higher-level
99 patterns (e.g. diachronic development of a cluster as a proxy for information spreading).
- 100 • *Passage-level* analysis considers a single instance of text reuse but focuses on existing differences
101 between (pairs of) witnesses (i.e. text passages that are deemed to contain the same text despite some
102 variations). The possibility of inspecting text reuse witnesses in their original broader context (i.e.
103 the position of a reused passage within the book or newspaper page) is an important aspect of the
104 contextualisation of the reused text.

105 Generally, distant reading approaches tend to privilege analysis of corpus-level and document-level text
106 reuse, while the close reading approach is more concerned with cluster-level and passage-level reuses.
107 Existing interactive visualizations of text reuse tend to support multiple levels of analysis at once and often
108 allow users to seamlessly move between levels. Visualisation techniques for cluster- and passage-level
109 text reuse resemble those used to represent text alignment in other scenarios, e.g. translation alignment,
110 collation of sources, etc. (Yousef and Janicke, 2021).

111 The interface developed for the *Graph – Text reuse in rare books*⁹ project constitutes a compelling
112 example of interfaces supporting multiple levels of exploration. It was developed to enable the exploration
113 of text reuse passages extracted from a corpus of 1,300 OCRed rare books. Firstly, corpus-level text reuse

⁶ <https://www.etrapp.eu/research/tracer/>

⁷ The Python package `textreuse-blast` provides an implementation of this method.

⁸ <https://github.com/dasmiq/passim>

⁹ <https://graph-rare-books.ethz.ch/>

114 is represented as a graph where two nodes (books) are connected when they contain reused passages, with
 115 the additional possibility of ordering the graph by time (of publication). Secondly, a static alluvial diagram
 116 allows readers to inspect more closely document-level reuse between pairs of books; this is especially
 117 useful to understand flows of reused text across books. Lastly, a facsimile side-by-side view of pairs of
 118 books permits to focus on passage-level reuse. This viewer is not aimed at highlighting differences between
 119 reuse passages, but rather at displaying them in their original context (especially meaningful in the case of
 120 rare books).

121 Graph visualization of text reuse at corpus level has also been used in the context of the Viral Texts project,
 122 which studied virality in newspapers during the interwar period. An interactive network visualization ¹⁰
 123 developed by the project provides a bird's eye-view of millions of text reuse passages, distilled into a
 124 graph that shows how newspapers formed a network of reprints and content reuse. Node size and color are
 125 used to express node centrality and grouping into community clusters, respectively, while the thickness
 126 of edges connecting nodes indicates the number of shared reprints. In addition to network visualization,
 127 geographical maps were used to support cluster-level analysis, as they allow to visualise at a glance the
 128 geographical distribution of reprints of a given text (Cordell, 2015).

129 Finally, visualizations of text reuse for the study of reception – be it literary or scholarly – privilege the
 130 corpus-level analysis of text reuse data and tend to present them in some aggregated form. In fact, what
 131 matters for the study of reception is how repetitions (quotations) are distributed, rather than the fine-grained
 132 differences between them. Examples of text reuse visualizations geared towards the study of reception are
 133 *Cited Loci of the Aeneid* (scholarly reception of Vergil's *Aeneid*) (Romanello and Snyder, 2017) and the
 134 *Reception reader* which focusses on Victorian literature (Rosson et al., 2023).

135 2.3 Text reuse detection in the *impresso* project

136 We used the open-source software Passim (Smith et al., 2015) to detect text reuse within the *impresso*
 137 corpus. Passim outputs clusters, groups of newspaper passages (or witnesses), from different newspapers,
 138 that share a common text span—the reused passage—of varying length (see Figure 1). The reason for
 139 choosing Passim over existing alternatives¹¹ was its ability to scale up, guaranteed by the software's parallel
 140 computing architecture. Preliminary tests on the *impresso* corpus showed that Passim's fuzzy alignment
 141 algorithm was able to detect reuse despite the presence of (moderate) OCR noise.¹²

142 2.3.1 Text Reuse detection and processing

143 As a pre-processing step, we ran Passim in boilerplate detection mode; this allowed us to identify — and
 144 later filter out — boilerplate content present in our corpus, i.e. portions of text that are repeated within the
 145 same newspaper in a one-month window (as opposed to reuse across different newspapers). All content
 146 items where boilerplate text was detected were filtered out from Passim's input. This pre-processing step
 147 allowed for reducing the final number of detected text reuse clusters by removing some noise from the
 148 input data. After filtering for boilerplate text, we extracted 6,177,815 text reuse clusters, for a total of
 149 16,099,821 reused passages. The reused passages are contained in 8,111,123 content items, meaning that
 150 roughly 17% of all content items in the corpus (n = 47,798,468) are part of at least one text reuse cluster.

151 We then post-processed Passim's output to enrich the detected clusters with the following information
 152 (see also Table 3):

- 153 • *Cluster size*: the number of passages contained in a cluster;
- 154 • *Lexical overlap*: the percentage of unique tokens that all passages in a cluster have in common (all text
 155 is lowercased and punctuation is stripped);
- 156 • *Time span*: the time window covered by documents in the cluster, expressed in number of days. It is
 157 computed as the difference between the publication date of the oldest and the most recent content item
 158 in the cluster.

¹⁰ <http://networks.viraltexts.org/1836to1860/>

¹¹ See Romanello and Hengchen (2020) for a list of available TR detection software. Given the abundance of existing implementations, the lack of a systematic benchmark evaluation is clearly a major limitation in determining which tool is better suited for processing a specific type of corpus.

¹² Vesanto et al. (2017, p. 55) found that BLAST outperforms Passim in terms of recall when tested on a corpus characterised by extreme OCR noise.

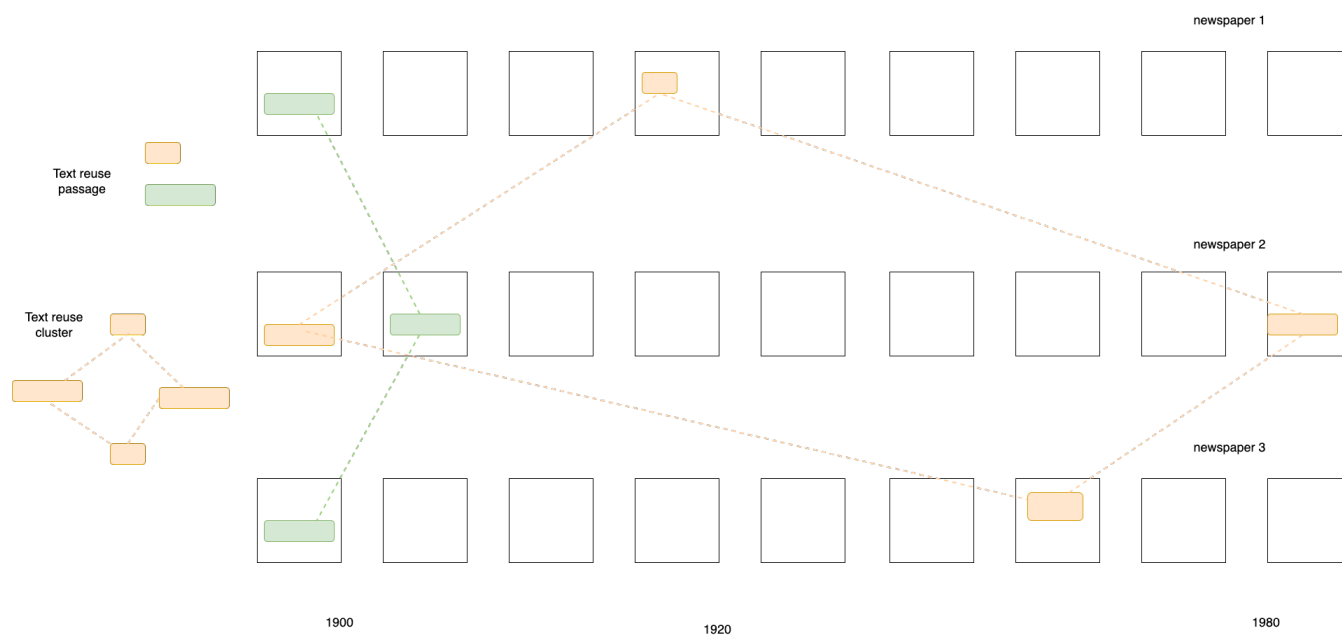


Figure 1. Schematic view of text reuse clusters and passages extracted from a newspaper corpus.

159 2.3.2 Integration of Text Reuse data in *impresso*

160 Text reuse data are already integrated and displayed in two main parts of the current *impresso* application.
 161 First, in the article reading view, coloured highlights indicate to the reader which parts of an article are
 162 reused elsewhere in the corpus (Figure 3). Second, in a text reuse explorer that precedes the prototype
 163 interface we discuss here. This first version already allows users to browse, search or filter text reuse
 164 clusters by any of the characteristics computed in the post-processing step, such as cluster size (i.e. number
 165 of passages contained), lexical overlap or time span covered. Most importantly, users can filter clusters to
 166 keep only those found in one of their collections. This functionality allows to *reveal* the presence of text
 167 reuse within a carefully selected and possibly manually curated subset of the corpus.

168 One of the main difficulties we faced in integrating text reuse into the *impresso* application was the scale
 169 of data, and more specifically how to enable an effective exploration of millions of detected clusters. Our
 170 approach to this problem consisted in providing users with as many filters as possible, as a powerful way
 171 of sifting through the large number of clusters extracted by Passim. One example is the long-term reuse of
 172 newspaper contents (Salmi et al., 2019), i.e. articles that are reprinted over and over in a relatively long
 173 period of time: Users can refine their query by setting a filter on the time span of the cluster, so that only
 174 clusters consisting of articles covering a time span of e.g. ten years are retained. This first version mainly
 175 supports cluster- and document-level research with a basic set of search options and filters without distant
 176 reading perspectives.

177 The development of the new version was motivated both by the opportunity to fully leverage the available
 178 enrichments and by the prospect of supporting additional use cases, including passage- and corpus-level
 179 research. To this end, we integrated text reuse and semantic enrichment data, i.e. named entities, topics,
 180 and content item types (where available), and aligned them with text reuse passages and clusters.

3 HISTORICAL RESEARCH AND TEXT REUSE

181 After reviewing the state of the art in TRD for historical texts, this section discusses the motivations and
 182 needs of historians interested in working with newspaper collections.

183 Past and present media have been, and still are entangled in complex communication networks,
 184 manifested in the form of interactions between different stakeholders such as journalists and press agencies.
 185 Connectivity in such networks was influenced by various factors, including geography, politics, technology,

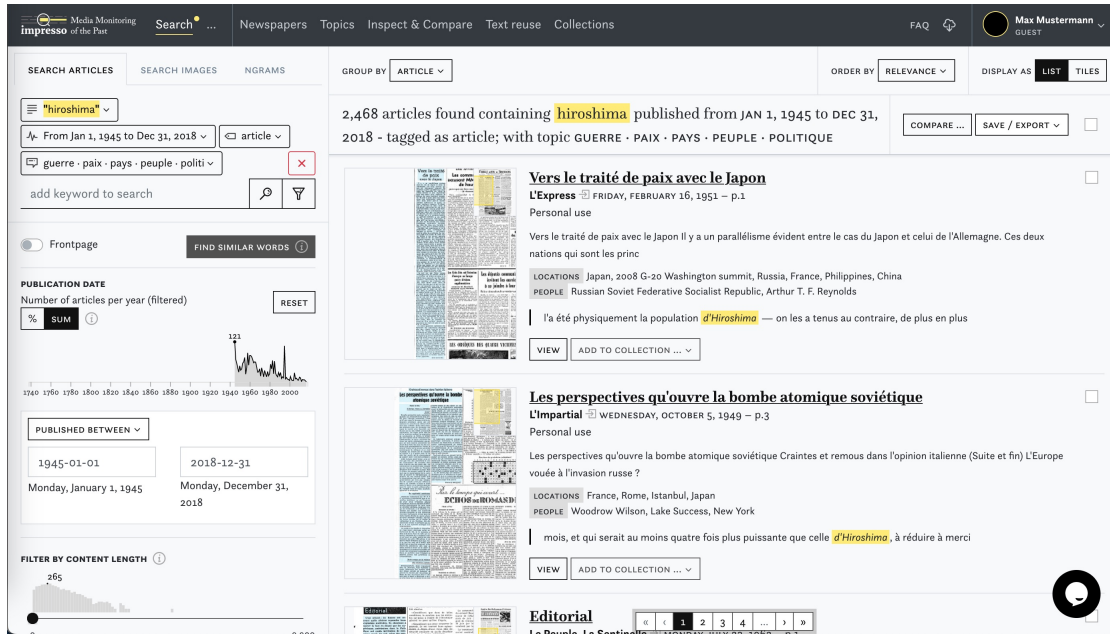


Figure 2. Screenshot of the *impresso* application for the exploration of semantically enriched historical newspapers.

186 communication infrastructures, languages, and commercial interests. By studying how texts circulated,
 187 historians can reconstruct the emergence and dissolution of links between stakeholders across time and
 188 space. Studies of copy-paste journalism, plagiarism, paraphrasing, literary and scholarly citation, the
 189 dissemination of specific discourses, and similar phenomena therefore all stand to benefit from text reuse
 190 detection.

191 TRD reveals many different types of text reuse such as jokes, adverts, boilerplates, speeches, or religious
 192 texts, but also short stories and reprints of book segments. Each of them is tied to a different logic and
 193 motivation and enables researchers to study different aspects of past media. We identify five high-level
 194 historical research objectives in the study of historical media and subsequently derive 11 common tasks
 195 from these objectives.

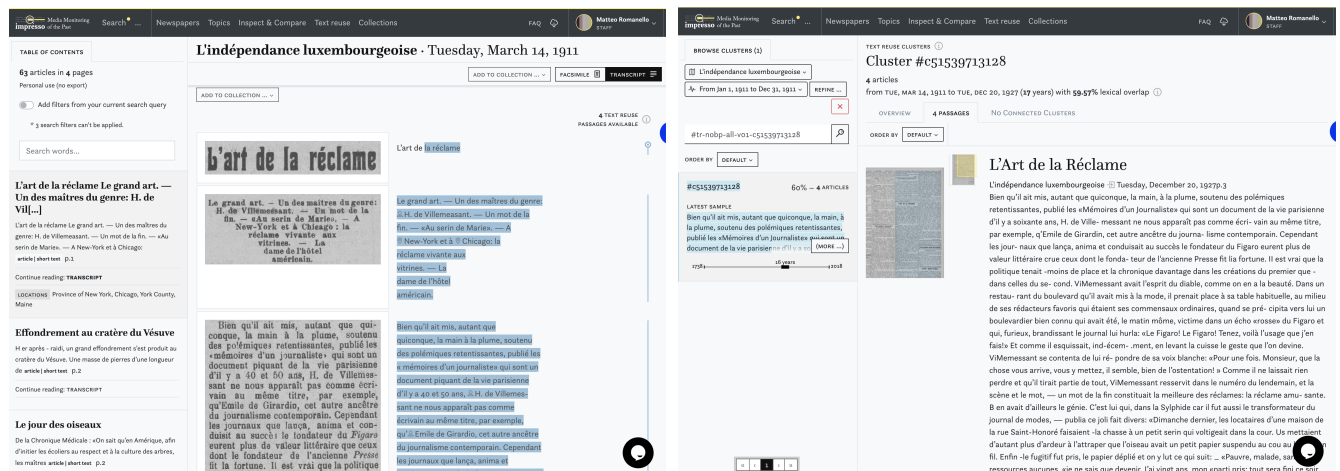


Figure 3. Display of text reuse in the *impresso* application's article reading view (left) and first version of the text reuse explorer (right).

Table 1. Types of temporalities in text reuse in historical newspapers.

<i>Type</i>	<i>Description</i>	<i>Measures</i>	<i>Examples</i>
Duration	The time period which is covered by a cluster ranging from the earliest to the latest publication date of individual passages.	Publication date	Paju's et al.'s notions of fast and slow text reuse fall into this category.
Virality	The speed (measured in days) and breadth of text reuse passages spreading within a corpus. Speed corresponds to time passed (e.g. days) whereas breadth corresponds to the number of publications which contain a passage at a given point in time.	Publication date, number of publications	News of the sinking of the Titanic or the destruction of the Hindenburg Zeppelin travelled around the world within days or weeks.
Rhythm	Pattern with which text reuse passages appear over time.	Distance between publication dates	Reprints of articles on the occasion of their anniversary, e.g. on the occasion of the bombing of Hiroshima.

196 3.1 High-level historical research objectives

197 Within historical research objectives, we distinguish between media-centric and content-centric
 198 perspectives. Media-centric perspectives seek to understand the functioning and evolution of the press as a
 199 system of information production and dissemination. Content-centric perspectives use historical media to
 200 reconstruct public attitudes and focus on the representation of past discourses.

201 3.1.1 (Trans-) National Media Ecosystems

202 With the increasing availability of digitised newspaper collections, media historians have begun to
 203 broaden the scope of their analyses: Attention has shifted from the in-depth reconstruction of the history of
 204 individual titles to a broader approach that embeds them in a transnational media ecosystem and emphasises
 205 the critical role of such connections in the creation and dissemination of information. Current research
 206 seeks to understand the functioning of this ecosystem and the agents which shaped it. This includes
 207 questions about the underlying ideological, commercial, and financial structures on which historical media
 208 ecosystems were based. Previous research, for example, has shown the importance of telegraph lines
 209 and railways in the spread of information within the United States (Smith et al., 2013) and pointed to
 210 individual cities as information dissemination hubs (Cordell, 2015; Salmi et al., 2020). Other work has
 211 studied multilingual information flows from a transnational perspective to examine the connections, gaps
 212 and silences in the system, and the press as a site of manipulation (Keck et al., 2022; Paju et al., 2023;
 213 Paasikivi et al., 2022).

214 The increasing availability of text reuse data for different countries will further advance systematic
 215 analyses of transnational (re-) printing dynamics. Of particular interest are e.g. internationally
 216 operating press agencies with their ability to disseminate content across borders and languages nearly
 217 simultaneously.¹³ A transnational perspective reveals the ways in which news is altered and contextualised
 218 as it travels. Scrutinising the reproductions of text by examining additions and deletions as traces of
 219 adaption helps us understand what was considered common knowledge in one (national) context but not in
 220 another. It foregrounds how perceptions and descriptions are adapted to new audiences.

221 3.1.2 Newspaper Content as Bricolage

222 From a linguistic perspective, newspaper discourse is not necessarily original or innovative. Many genres
 223 such as weather forecasts or sports reporting operate within constraints and happen to be almost formulaic.
 224 Thérenty and Venayre (2021) and Walma (2015) paid special attention to the relations between these genres:
 225 how did content travel between them? In general, articles emerge through a process of creative re-use and

¹³ Discussions at the workshop also motivated work on a Master's thesis project which led to the successful detection of press agencies in the *impresso* corpus (Marxen, 2023) and the ability to filter for content from individual press agencies will be available in future versions of the interface.

226 re-appropriation. Whole fragments, sentences and quotations are often transferred to novel contexts. In this
227 sense, newspaper content emerges through a process of what could be called *bricolage*, in which content is
228 soldered together from existing fragments and textual patterns. In other words, newspapers content is often
229 harvested from a wide range of available textual material.

230 This research objective investigates text reuse through the angle of compilation and content production.
231 Text reuse measures can be used to encode textual relations and connections, and thereby enable researchers
232 to critically disentangle the genesis of newspaper content. Moreover, the concept of *bricolage* opens up
233 a graded, more nuanced approach, to the study of text reuse: it foregrounds how the creation of news
234 content emerges in a complex process of transformation, compilation and innovation. Newspaper titles
235 operate in a media ecosystem compiling and recreating content harvested from the “grid” (press agencies,
236 or newspapers) and merging it with self-generated content (ads, journalistic work, external contributors
237 etc.).

238 3.1.3 Historicising Virality

239 Virality is more commonly understood as a phenomenon of the internet era and is associated with three
240 characteristics: High speed, high volume and the ability to adapt and spread quickly. Paju et al. (2022)
241 have used text reuse data in an attempt to measure and compare different degrees of virality for content
242 that was republished within days or weeks. They define a virality score based on the number of titles
243 within a cluster, the number of unique printing locations, and the distance in days between the first and last
244 passage publication date. They show that different types content qualify for different types of repetition:
245 An advertisement for Finnish cigarettes in 1916 constituted the most viral content in their corpus while
246 institutional announcements, literary, and religious texts often fall into the mid-range virality category.

247 Such measures may yield additional insights into the functioning of historical media ecosystems, e.g.
248 by revealing which types of texts circulated more efficiently than others within and beyond national
249 boundaries, and who was responsible for their creation and dissemination. Virality also offers insights into
250 how transport infrastructures and geography shaped information dissemination, or how other factors such
251 as religious and political affiliations influenced the reception or rejection of content.

252 3.1.4 Tracing Historical Events

253 The press is a system of knowledge production and representation that not only presents events to the
254 public, but also situates them within a specific political, economic, social, and cultural context. This
255 influences the perception of historical events by the public. A comparative approach to event coverage
256 therefore allows us to reconstruct the political, social, and cultural identities of individual newspaper titles
257 and how they evolved over time. We identify two strategies to trace the coverage of historical events
258 via text reuse. The first is bottom-up and concentrates on individual, known events and the question of
259 whether or not they were picked up by the press and if so, how. For example, Oiva et al. (2020) studied
260 how news of the assassination of Nikolay Bobrikov, the Governor-General of Finland in 1904, traveled
261 in waves across historical communication infrastructures. The second strategy is top-down and focuses
262 on the types of events that successfully spread across media ecosystems. For example, Keck et al. (2022)
263 used TRD across newspaper collections from the United States, Britain, Germany, Austria and Finland
264 to identify global media events. Through this approach, they discovered a substantial number of articles
265 that circulated during Hungarian revolutionary Lajos Kossuth’s tour of America to seek US financial
266 support for another revolution in Europe. His arrival in New York in December 1851 and his subsequent
267 travels to Washington, DC triggered a proliferation of coverage and reprinted texts. Comparing text reuse
268 across national and linguistic borders highlights the specific patterns and complexities of transatlantic
269 news circulation, including pathways, reach, temporality, vagaries, and gaps. While this work illustrates
270 the usefulness of TRD, it also highlights the benefits of international cooperation when working with
271 multilingual datasets.

272 3.1.5 Capturing Historical Zeitgeist

273 To some degree, historical media record the attitudes, norms, beliefs, moods and feelings of past
274 generations, and can thus serve as a proxy for the study of “Zeitgeist”. This concept alludes to notions
275 of similarity and parallel evolution, and explains how texts which were produced independently come to
276 share certain characteristics. Because of its fuzziness, Zeitgeist marks the borderline of what text reuse can
277 capture. Zeitgeist manifest itself in various forms, such as mental maps that informed the editing of texts,

Table 2. List of tasks and current degree of support by the Text Reuse at Scale interface.

Task	Title	Level	Support
1	Obtain an overview of text reuse in a corpus, collection or query	Corpus	yes
2	Obtain an overview of a single cluster	Cluster	yes
3	Compare passages	Passage	yes
4	Compare clusters	Cluster	yes
5	Identify different types of text reuse	Corpus	yes
6	Generate research corpora based on text reuse clusters	Corpus	yes
7	Identify connections	Corpus	partial
8	Detect and trace virality	Corpus	no
9	Search for passages	Passage	no
10	De-duplicate content	Corpus	no
11	Export of text reuse data	All	planned

278 advertises for (cultural) products, or the mere existence of coverage of cultural practices. These manifestations
 279 are created using persistent and implicit templates that change their content over time - an example would
 280 be dance fads such as Polka or Macarena, which are dominant at one point but then slowly fade away.
 281 Related work looks at conceptual change over time (Verheul et al., 2022) or the cultural impact of Cholera
 282 epidemics (Paasikivi et al., 2022).

283 3.2 Tasks for the Exploration of Text Reuse in Historical Newspapers

284 This section operationalises the high-level objectives by formulating concrete user tasks. The tasks are
 285 not directly linked to objectives, but provide building blocks that compose workflows for the exploration of
 286 text reuse data. Table 2 relates these tasks to analytical levels described in Section 2 and also mentions the
 287 degree of support provided by the interface.

288 **Task 1: Obtain an overview of text reuse at the corpus, collection or query level.** Before analysis,
 289 users need to determine whether or not a given dataset, such as a corpus, a corpus subset or a query result,
 290 contains instances of text reuse. Metadata plays an essential role in contextualising the presence of text
 291 reuse data. In the case of the *impresso* data, this includes publication dates, newspaper titles, country of
 292 publication, content types, languages, topics, and named entities. It also consists of text reuse specific
 293 measures such as lexical overlap, time span between publication dates, cluster size and number of passages.

294 Computing measures of spread offers additional insights into the distribution of text reuse data at different
 295 levels of granularity. This includes finding the largest/smallest clusters, clusters with the highest/lowest
 296 lexical overlap, the earliest/latest cluster, or clusters with the longest time span between publication dates.

297 **Task 2: Obtain an overview of a single cluster.** This task is similar to Task 1 but focuses on the
 298 properties of a single cluster: the number of passages, their content, the lexical overlap between them, the
 299 time span and rhythm between publication dates, and the distribution of metadata.

300 **Task 3: Compare passages.** This task compares two or more passages and reveals textual differences and
 301 similarities through parallel reading. A common use case is the study of editorial changes in news agency
 302 dispatches. Comparisons can reveal adaptations to suit the political preferences of a newspaper's audience,
 303 additional explanations and clarifications to address varying knowledge horizons, but also unintended
 304 differences such as text degeneration caused by OCR errors.

305 **Task 4: Compare clusters.** Comparison is a powerful analytical instrument. This task compares sets of
 306 clusters based on the distribution of a) text reuse measures and b) metadata and semantic enrichments such
 307 as topics or named entities.

308 **Task 5: Identify different types of text reuse.** TRD captures many different types of reused texts that
 309 need to be distinguished from each other. In combination, semantic enrichments, TRD data and their

310 temporal distribution (see Table 1) provide powerful means to identify different types of text reuse. For
 311 example, older articles that are reprinted after years, or advertisements that were widely published in
 312 parallel, tend to contain named entities and topics that are typically associated with particular types of
 313 media content.

314 **Task 6: Generate research corpora based on text reuse clusters.** This task supports the fine-grained
 315 selection of content and the creation of meaningful subsets of text reuse clusters and their associated
 316 passages based on the aforementioned measures, enrichments and metadata.

317 **Task 7: Identify connections.** This group of tasks concentrates on the relational structure of media
 318 ecosystems. Examples include the (unwanted) reprinting of articles by different newspapers, the exchange
 319 of content between news agencies, or regular co-publication agreements between newspaper titles.

320 **Task 8: Detect and trace virality.** This task corresponds to the pioneering work of Paju et al. (2022) and
 321 measures the efficiency and speed by which content spreads across newspapers.

322 **Task 9: Search for passages.** This task describes a search scenario in which a seed text is used as a
 323 query and compared to known text reuse passages. An example usage would be the upload of a speech to
 324 determine whether parts of it were ever published in a given corpus.

325 **Task 10: De-duplicate content.** This processing step removes duplicate texts, e.g. to avoid over-
 326 representations of highly circulated texts or recurrent elements such as adverts or boilerplate.

327 **Task 11: Data export.** Data export allows further processing outside the constraints of an application.
 328 This would allow network or geo-spatial analyses or further processing.

4 THE IMPRESSO TEXT REUSE AT SCALE INTERFACE

329 4.1 Main interface components

330 The *impresso* Text Reuse at Scale interface consists of a search and filter pane on the left and three tabs
 331 in the centre (see Figure 4). This section introduces these components and provides examples to illustrate
 332 their usage.

333 4.1.1 Search and Filter Pane

334 Figure 4 shows the search and filter pane. Users can compile queries using the versatile *impresso* Search
 335 component together with a variety of filters. These filters include newspaper metadata, user-generated
 336 article collections, text reuse clusters, and semantic enrichments (topics, language, content type, and named
 337 entities). In addition, the component allows filtering based on the number of passages over time, lexical
 338 overlap within clusters, cluster size and time span (for details see Table 3).

339 Complementary modal dialogues as shown in Figure 5 (centre) serve as a bridge between distant and
 340 close reading. They reveal, for example, notable peaks in the distribution of lexical overlaps, cluster sizes
 341 and time spans between publication dates and allow users to quickly browse corresponding passages.

342 Taken together, these search and filtering capabilities offer a versatile framework for querying text reuse
 343 data. As a first illustration, different types of text reuse can be retrieved by filtering based on the time
 344 span between passage publication dates as described in Task 5 - Types. Paju et al. (2023) point to the
 345 different speeds at which text reuse occurs, distinguishing between rapid (within 1 year) and mid-range (up
 346 to 50 years) cycles. Anecdotal evidence suggests that slow text reuse (up to 140 years) is typically tied
 347 to conscious re-prints of archived materials. Following Paju et al.'s classification of text reuse speeds, we
 348 find 5,546,859 clusters that qualify as rapid (0-365 days), 557,555 clusters that qualify as mid-range (1-50
 349 years), and 21,799 clusters that qualify as slow (50-200 years).

350 As another illustration, let's consider the press coverage of the US attack on the Japanese cities of
 351 Hiroshima and Nagasaki on 6 August 1945. We begin with a basic keyword query for *hiroshima* which
 352 yields 2897 passages in 1465 clusters. We note clusters with very short time spans (0-2) concentrated in
 353 1945 and 1995. Upon closer inspection, cluster c466008 stands out: it includes 9 passages from articles
 354 that were republished surrounding the anniversaries of the attacks, making it an example of cyclical text
 355 reuse. The Swiss newspaper *L'Impartial* published them with minor changes irregularly between 2007 and
 356 2015 in commemoration; in 2009 and 2012 the article was also republished by *L'Express*.

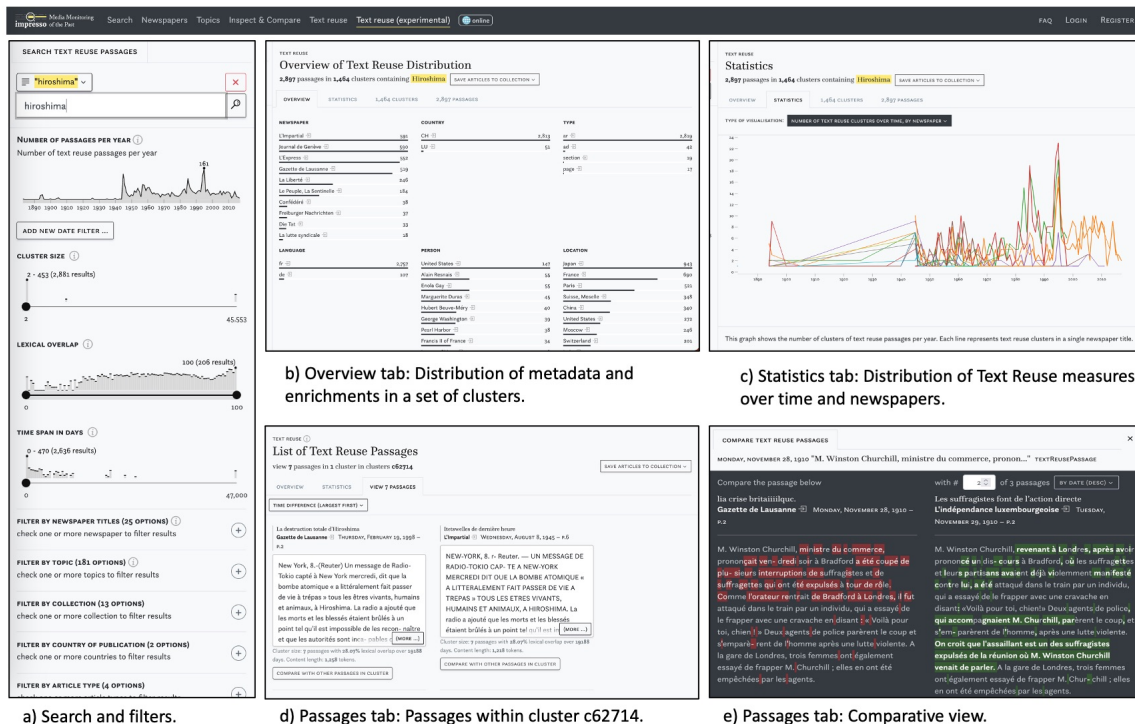


Figure 4. Main components of the text reuse interface: Search and Filter Pane (left) and the Overview, Statistics and Passages tabs (centre).

357 Finally, Figure 7 shows a query for text reuse in adverts that are part of a large collection of articles about
 358 nuclear power and linked to the topic *eau · énergie · gaz · électricité · air*. The peak in 1977 points to cluster
 359 c276252 which comprises 21 adverts in favour of nuclear power, published in parallel in several Swiss
 360 newspapers.

361 4.1.2 Overview Tab

362 Figure 4b) displays the *Overview* tab which was inspired by Task 1 - Overview. It shows the distribution
 363 of semantic enrichments and metadata relative to a search or filtering operation. In this case, it displays the
 364 results for the preceding keyword query for the string *hiroshima*. Enrichments are grouped by type and
 365 represented using small multiples of bar charts.

366 In this instance, we learn that the vast majority of text reuse passages that contain *hiroshima* are linked
 367 to content published in French in Switzerland. Content published in German in Luxembourg remains the
 368 exception. A closer look at the newspaper titles suggests that roughly 80% of these passages appear in just
 369 four newspapers. Unsurprisingly, the most prominent topics are associated with war, nuclear technologies,
 370 and aviation.

371 4.1.3 Statistics Tab

372 The second tab, *Statistics*, visualises the distribution of text reuse measures in relation to queries and
 373 provides a distant reading perspective. A drop-down menu offers access to five views: four-line charts and
 374 a matrix visualisation, which are shown in Figure 8 and discussed in more detail below.

375 Figure 8a) displays the **passage count over time by newspaper title**. This view reveals periods of
 376 heightened or reduced text reuse activity for one or more newspaper titles. A complementary view represents
 377 the number of clusters over time (not shown).

378 As an example, we will compare the distribution of text reuse in Swiss and Luxembourgish newspapers.
 379 In the search and filter panel, we set the time delta filter to a range between 0 and 100 days. Lexical overlap
 380 is set to a moderately high range of 20-99%, which should also retrieve reused text segments of smaller
 381 size embedded in a larger text. Finally, we use the cluster size filter to exclude a disproportionately large

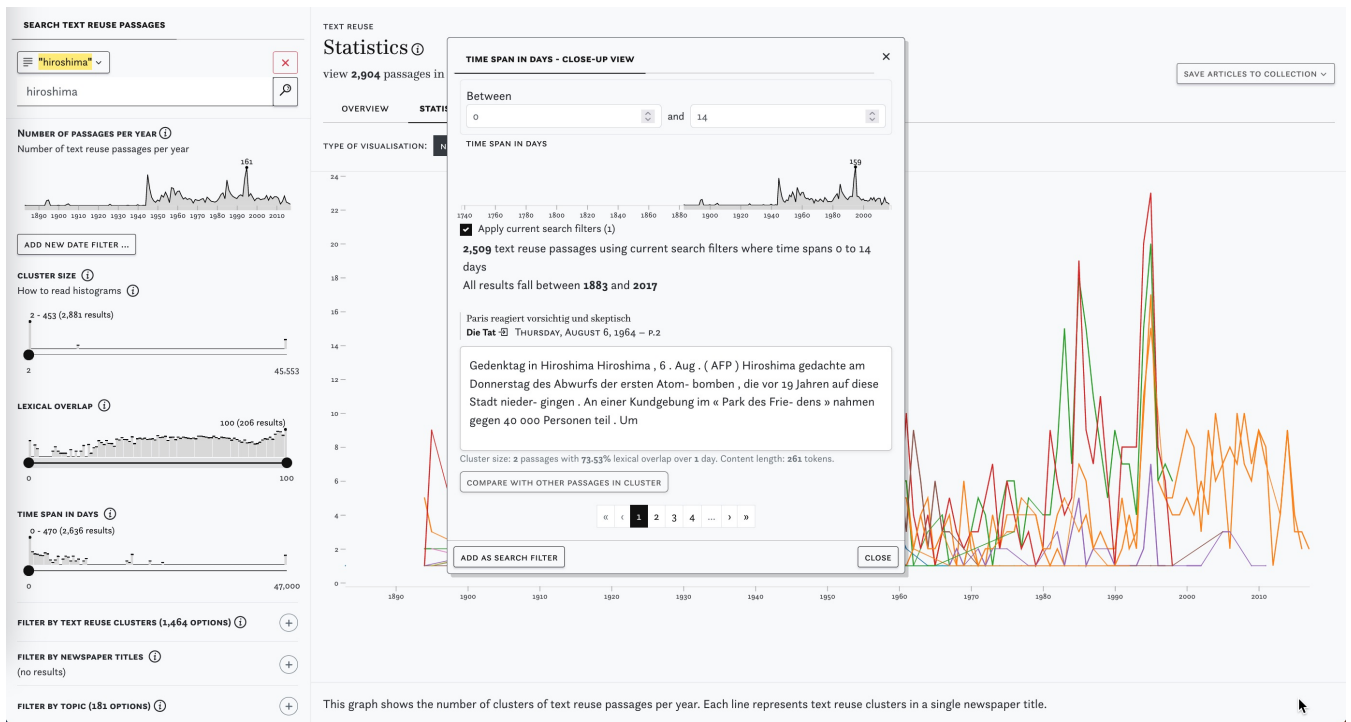


Figure 5. Screenshot of the search and filter pane with a keyword search for *hiroshima* (left) and the close-up view with a time span filter for 0 to 14 days (centre).

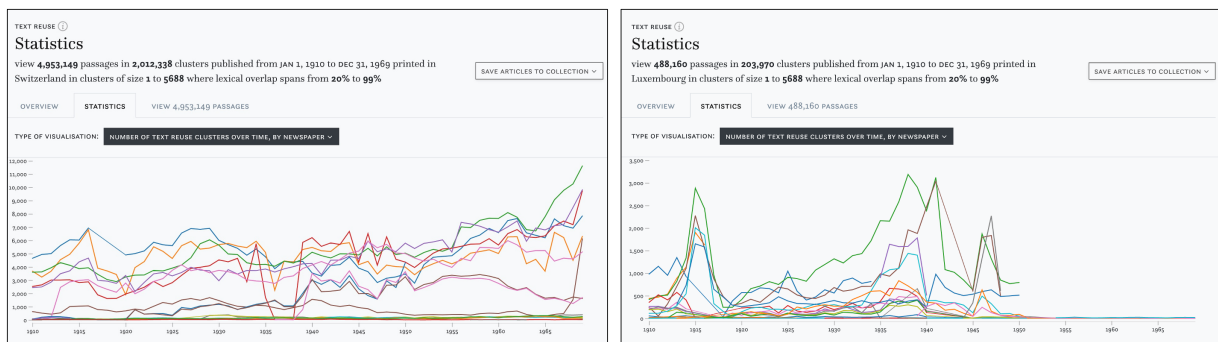


Figure 6. Number of clusters detected in Swiss (left) and Luxembourgish (right) newspapers 1910 - 1970.

382 cluster of 45.000 passages. This results in approximately 5.5 million passages for our time period. Looking
 383 at the distribution of cluster sizes in Switzerland, Figure 6 (left) suggests some variation between titles, but
 384 otherwise no change between the pre- and post-war period. In contrast, the Luxembourgish press (right)
 385 exhibits a growing number of passages since the 1930s and clear peaks in 1915 and during the Second
 386 World War, followed by a stark decline after 1950 which can be explained by the limited availability of
 387 Luxembourgish content for this time period in our corpus.

388 The **minimum, mean, and maximum cluster sizes over time** are shown in Figure 8b). Overall, the
 389 number of text reuse clusters and passages rises constantly over time, parallel to the number of available
 390 content in the *impresso* corpus. For another example, we make use of the *impresso* Collections feature
 391 which stores sets of articles based on either manual selection or queries (see Task 6 - Research corpora).
 392 Collections can also be created based on text reuse clusters - albeit with the caveat that it saves both
 393 passages and entire articles. The above-mentioned Figure 7 shows the component.

394 Figure 8c) captures the **minimum, mean, and maximum cluster sizes per newspaper**. This view
 395 depicts the distribution of cluster sizes across titles and shows which newspaper produced the smallest (or

Table 3. Text reuse measures and their representation in the interface.

<i>Measure</i>	<i>Description</i>	<i>Implementation in interface prototype</i>
Passages per year	Number of passages counted in a given year.	Line chart which displays the count of passages per year for a given query or filter operation. This gives a first indication, during which years text reuse occurred more commonly. Time sliders and precise date entry allow users to filter for exact date ranges to inspect.
Cluster size	The number of passages contained in a cluster.	Histogram which shows the distribution of text reuse cluster sizes and indicates the highest score. The histogram groups clusters of size n and displays their sum. This gives a first indication of averages as well as outliers. Sliders can be used to specify a cluster size range of interest. Filtering by cluster size allows to exclude or explicitly focus on outliers but different cluster sizes may also correspond to different types of content.
Lexical overlap	The percentage of unique tokens that all passages in a cluster have in common. All text was lowercased and punctuation was stripped.	Histogram which shows the distribution of lexical overlap in percent and indicates the largest number of clusters for a given score. Extremely low lexical overlap decreases the chance to discover meaningful text reuse whilst extremely high overlap will only reveal near-copies of content and may be too restrictive for some purposes.
Time span	The time window covered by documents in the cluster, measured in number of days.	Histogram which shows the gap between the earliest publication date of an article in a text reuse cluster and the latest measured in days and indicates the largest number of passages for a given score. This is an efficient approach to discover or filter for instances of slow, mid-range and rapid text reuse. The histogram groups clusters by the number of days in between publication dates and displays their sum.
Text reuse clusters	Clusters store text segments (or passages) that are reused in different units of a corpus.	List of text reuse clusters which match a given query, sorted by number of passages. Each cluster is characterised with basic information (passages count, lexical overlap, time periods and years covered) as well as a snippet preview of the passage. Clusters are sorted by the number of matching passages. Clusters can be selected manually for further inspection in the Text reuse app or in other <i>impresso</i> components such as Search.

396 largest) clusters. In this case, both the newspapers *Le Peuple*, *La Sentinelle* and *Die Tat* stand out with
 397 above average maximum cluster sizes (orange).

398 **Lexical overlap between newspaper titles** is shown in Figure 8d) while Figure 8f) uses a matrix view to
 399 highlight **co-occurring text reuse clusters between newspaper titles**. Both views reveal particularly high
 400 lexical overlaps and a large number of shared passages for *Journal de Geneve* and *Gazette de Lausanne*.
 401 These findings confirm our prior understanding of the frequent co-publication patterns among these titles.

402 Finally, the distributions of **lexical overlap including minimum, maximum and mean across all**
 403 **clusters over time** are shown in Figure 8e) and offer corpus-level insights. For example, the maximum and
 404 mean lexical overlap increases from the 1970s onwards, which may be a result of the improvement of OCR
 405 quality over time. On the basis of individual titles, it also shows that *Confédéré* defies this trend, with mean
 406 and maximum overlap constantly decreasing since the 1970s.

407 4.1.4 Passages Tab

408 The third tab titled *Passages* supports close reading of a given text reuse cluster (Task 2 - Cluster
 409 overview). The list of passages can be sorted by date, lexical overlap, cluster size, time span and passage

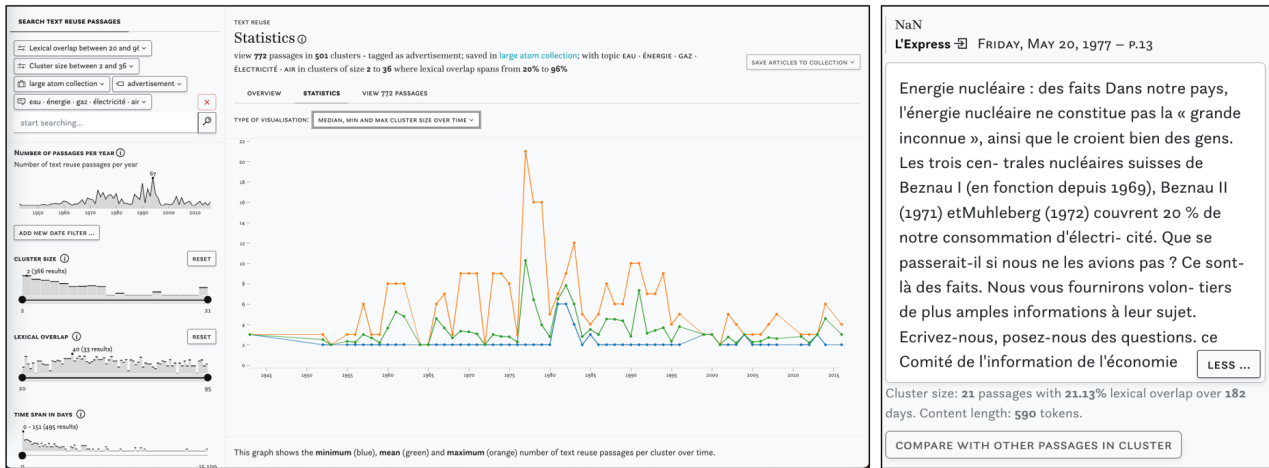


Figure 7. Example of a complex query using multiple semantic enrichments and *impresso*'s collections. Distribution of clusters on the left and passage from the largest cluster in 1977 on the left.

410 size. Figure 4d) displays cluster c62714, which has a large time span of 19,188 days (ca. 52 years). A
 411 closer look reveals that it contains an article published in 1945 about the attack on Hiroshima, which was
 412 republished by multiple newspapers at the time and then rediscovered in 1998, when it was republished
 413 again, this time by *Gazette de Lausanne* and *Journal de Geneve*.

414 Within the same tab, the *Compare* button below the snippet preview opens a comparison view (Task 3 -
 415 Compare passages). Figure 4e) highlights differences between two passages. Such differences can result
 416 from editorial interventions, including additions and omissions, but also from OCR variation. Users select
 417 a “start passage” of interest, which appears on the left side, and can then cycle through all other passages
 418 in a cluster using arrow buttons on the right side. Characters that are only present in the start passage are
 419 highlighted in red, those that are only present in the compared passage are highlighted in green. Here we
 420 see a side-by-side view of two passages that cover protests by suffragette activists and an ensuing attack on
 421 Winston Churchill in 1910. On closer inspection, it also reveals interesting nuances in the coverage of the
 422 event: whereas the *Gazette de Lausanne* (left) does not make an explicit link between the assailant and the
 423 suffragettes, *L'indépendance luxembourgeoise* (right) asserts that the attacker was believed to be part of the
 424 movement.

5 EVALUATION

5.1 Evaluation setting

425 To evaluate the prototype, we invited various scholars to review the interface independently and remotely.
 426 We received 13 responses: 5 from (digital) historians with research experience in historical newspapers,
 427 4 from computational linguists with experience in TRD, 3 from humanities scholars with experience in
 428 text reuse and virality, and 1 from a software developer with experience in text reuse visualisation. Five of
 429 these evaluators had also participated in the workshop.

431 The evaluation was carried out by means of a form¹⁴ which contained five evaluation tasks that highlighted
 432 the different functionalities of the interface and were illustrated with instructive examples. They correspond
 433 to Task 1 - Overview, Task 2 - Cluster overview, Task 3 - Compare passages, Task 5 - Types, and Task 6 -
 434 Research corpora.

435 Since the evaluation took place remotely, reviewers had to first familiarise themselves with the interface
 436 and the selected tasks. Therefore, for each evaluation task, the evaluators were first presented with the task
 437 definition (with an opportunity to comment) before being given instructions on specific operations to be
 438 tested in the interface, along with illustrative examples. For each evaluation task, the evaluators were asked

¹⁴ <https://zenodo.org/record/8009613/>

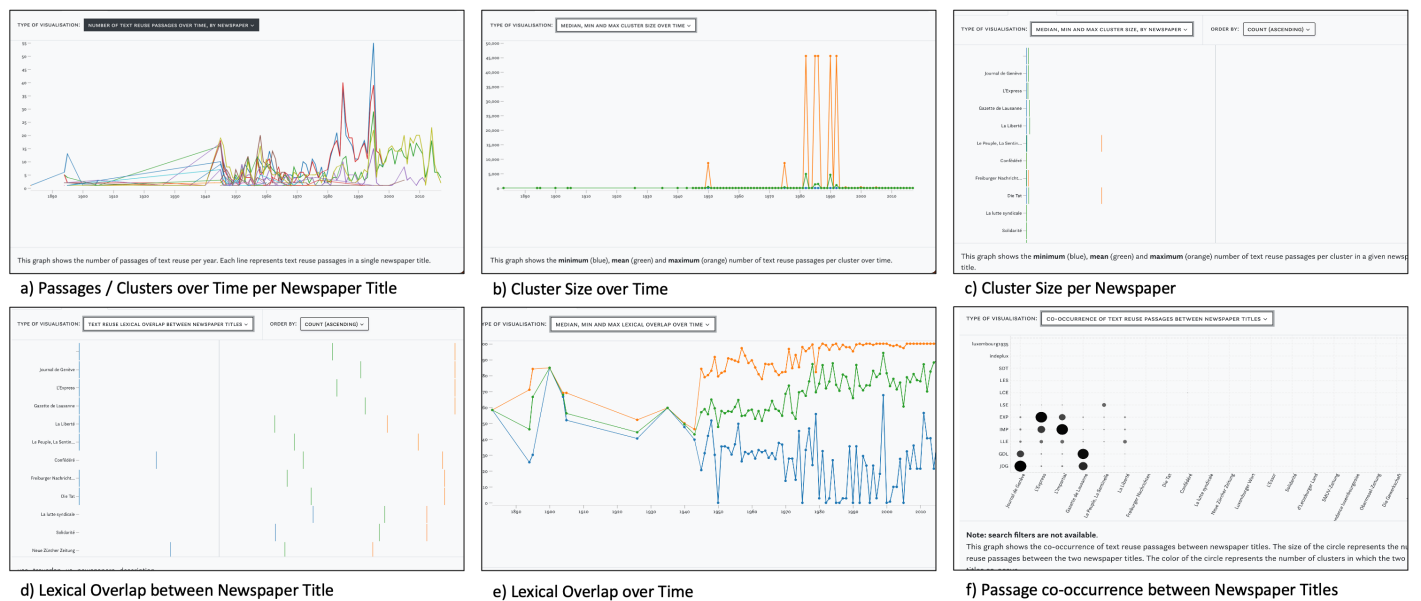


Figure 8. Views in the Statistics tab with distributions of text reuse measures across time and newspaper titles.

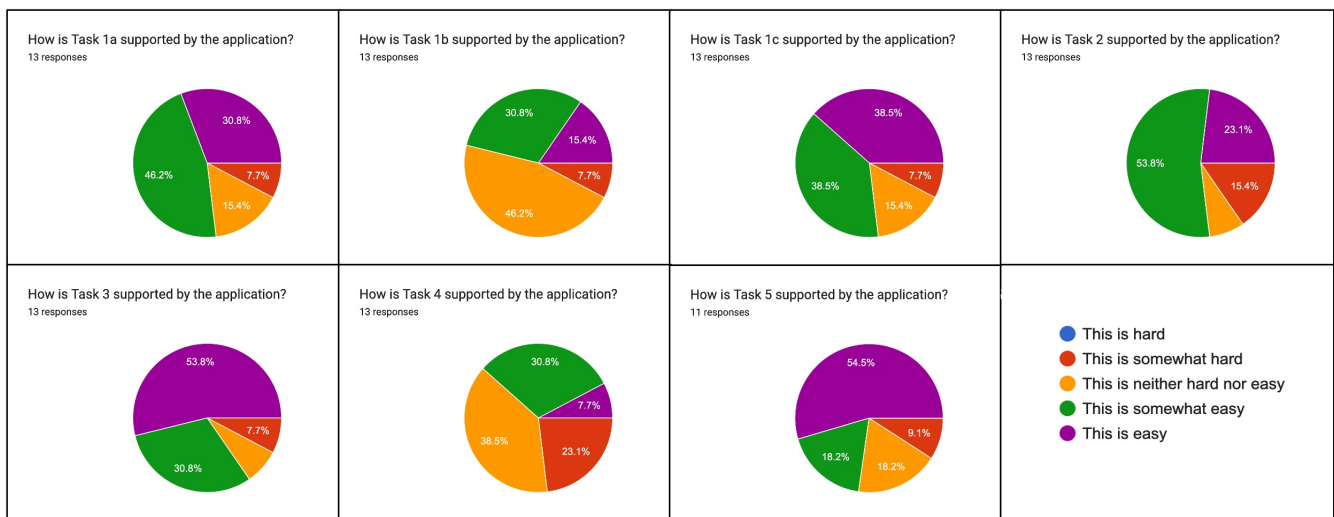


Figure 9. Rating of evaluation tasks regarding their perceived difficulty.

439 to rate the difficulty of the task on a five-point scale (see Figure 9). A concluding evaluation section allowed
 440 for an overall assessment of the interface, including its ability to effectively support the tasks presented,
 441 the quality of accompanying information, the ease of navigation, and the responsiveness of the system.
 442 Finally, evaluators were asked to indicate any irritations (“*Is there anything in the application that does*
 443 *not make sense? Does anything feel out of place?*”) and recommendations for its improvement (“*Future*
 444 *development of the application should focus on these tasks / features / overall improvements*”).

445 **5.2 Discussion of evaluation results**

446 **Evaluation tasks 1a-c: Obtain an overview of text reuse in a corpus, collection, or query.** This
 447 first segment familiarised evaluators with different aspects of the interface, notably search, filters, the
 448 tab views and the close-up view and was divided into three sub-tasks.¹⁵ Evaluators found the interface

¹⁵ For the sake of simplicity, we combine feedback on task definitions and their implementation in the following paragraphs.

449 intuitive overall, but also confirmed that familiarity with text reuse concepts such as clusters or passages is
450 an important prerequisite. Suggested improvements included the ability to compare the presence of text
451 reuse in the entire corpus with the presence of text reuse discovered as a result of specific query in order to
452 better contextualise the findings.

453 Especially for this first task, the feedback also reflected the learning experience of those evaluators who
454 were using the *impresso* application for the first time. Critiques of individual interface components will be
455 discussed below.

456 **Evaluation task 2: Obtain an overview of a single cluster.** The task was identified as part of an
457 exploratory workflow: “*It seems a typical task again, like drilling down into a specific set of documents*
458 *after first gathering a larger scale view in task 1*”. Regarding task implementation, evaluators found the
459 interface to be “*convenient and intuitive*” and suggested high-level fingerprint views for (sets of) clusters
460 to help with the assessment of cluster content.

461 **Evaluation task 3: Compare differences between passages within a cluster.** Evaluators pointed out
462 that this task helps scholars to identify different ideological lines in newspapers, but also enables tool
463 criticism. Overall, it complements distant reading operations: “*I feel this task foregrounds the complexities*
464 *of text reuse that remain hidden to the viewer who only gazes at the high-level statistics.*” Another evaluator
465 noted: “*Useful on how newspapers frame and present an event based on their ideological and political*
466 *preference.*” In terms of implementation, evaluators appreciated the ease of use of the comparative view
467 and suggested more abstract exploration of editorial practices and the ability to compare multiple passages
468 simultaneously.

469 **Evaluation task 4: Identify different types of text reuse.**¹⁶ The addition of semantic levels to the
470 exploration of text reuse data was overall welcomed. Task 4 was considered to be of particular interest to
471 scholars. Evaluators noted that the gap between filtering operations and empirically observable types of
472 text reuse had/has not yet been satisfactorily closed: “*It would be helpful to have some introduction to 1) a*
473 *taxonomy of reuse types, and 2) the different kinds of phenomena and how each maps to various (meta)data*
474 *variables.*” Feedback on task implementation was mixed, with a majority of evaluators finding the task as
475 either “hard” or “somewhat hard” (Figure 9). Several evaluators suggested the creation of specific filters
476 for empirically observed types of text reuse. This includes, for example, the reuse of older content by a
477 newspaper title and explicit support to filter for cyclical reuse. Others, and this may echo the previous
478 feedback, felt overwhelmed by the options to filter and visualise and did not know where to start. Still,
479 others were content: “*Takes some getting used to the filters and functionalities, but nothing problematic.*”

480 **Evaluation task 5: Generate research corpora based on text reuse clusters.**¹⁷ This task received
481 comparably little feedback since not all evaluators registered an *impresso* account in time to be able to test
482 it. One evaluator called it useful for historians “*though there is a bit of a conceptual gap between reused*
483 *passages and reused articles.*” Feedback on the implementation was generally positive, with some critiques
484 of the slow speed of collection processing and of the difficulty of finding the data export function.

485 We move to the discussion of individual components within the interface:

486 **Search.** With one exception, all evaluators either “somewhat” or “fully agree” that the Search component
487 facilitates effective exploration of text reuse data (Figure 10). We note, however, that previous experience
488 with the *impresso* application was an advantage and that some evaluators who were new to it struggled at
489 times, for example with searching for entities or the logic of removing filters.

490 **Filter pane.** Feedback on the filter pane was even more positive, but evaluators identified opportunities
491 for improvement. These included the addition of units to histogram mouse-overs, a better indication that
492 they are interactive, and pointers to a bug that prevented newspaper titles from being displayed as filter
493 options.

494 **Overview tab.** Again, feedback was overwhelmingly positive (see Evaluation task 1, above). Critical
495 remarks addressed its limited utility for exploring individual clusters, and the leap between text reuse
496 passages and the display of article-level enrichments such as named entities or topics.

¹⁶ Note that this evaluation task corresponds to Task 5 - Types.

¹⁷ Note that this evaluation task corresponds to Task 6 - Research corpora.

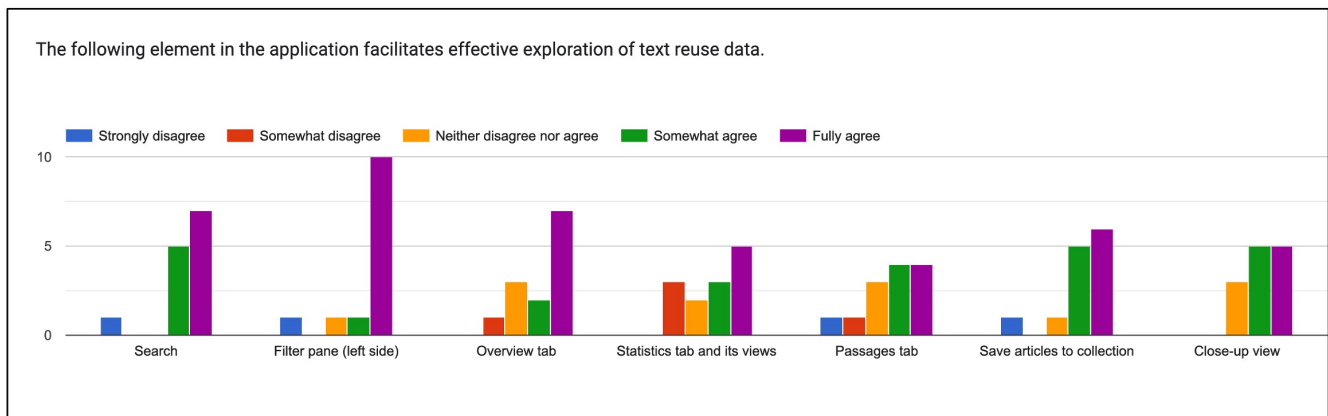


Figure 10. Rating of different components and views in the interface.

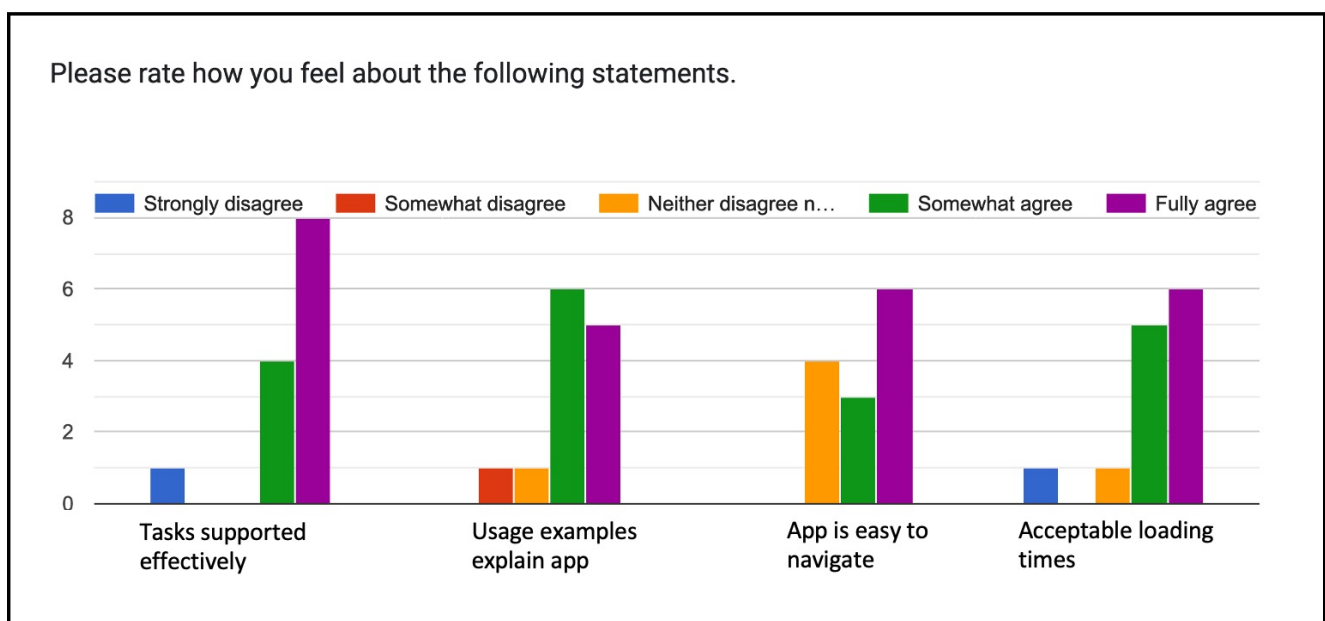


Figure 11. Overall rating of the interface.

497 **Statistics tab.** Feedback on the statistics tab revealed a need for more documentation and design
 498 improvements. Some evaluators struggled to read and interpret the charts, missed the option to zoom in on
 499 timelines, as well as more detailed information about their computation.

500 **Passages tab and passage comparison.** This segment divided evaluators. Some found it “again easy
 501 and intuitive” and “Very user friendly, no remarks.” Others missed a grouping of passages by cluster and
 502 struggled to find and operate the comparative view. Regarding the contrastive view, some found it difficult
 503 to cycle through different passages; they suggested changing the colour scheme and eliminating some of
 504 the mismatches such as white space or OCR mistakes for easier viewing.

505 **Close-up view.** The close-up view was again rated positively, with the only criticism being the difficulty
 506 offinding it without direct instructions and a bug that prevented the display of passage previews.

507 In the overall rating of the interface (Figure 11), the vast majority of the 13 evaluators either “somewhat”
 508 or “fully” agreed that the interface supports the evaluation tasks (12), that the usage examples explained the
 509 interface (11), that it was easy to navigate (9), and that the loading times were acceptable (11). The interface
 510 clearly has a learning curve, as described by one evaluator: “The functionality of the filters available here
 511 is impressive and of reasonable simplicity. I wouldn’t describe it as ‘easy’, mostly because there’s a lot

512 going on and a researcher not familiar with the dynamics of text reuse might be a bit lost, but I'm not sure
513 I would trade the current depth of filtering for easier use."

514 The answers to our questions regarding irritations and future improvements confirm the critiques of
515 the statistics tab and passages tab discussed above. At this stage of development, six evaluators found
516 them "either difficult to read or [they] did not provide useful insights." In addition, recommendations
517 for future development addressed the already foreseen integration of *impresso*'s Inspect & Compare
518 component (Düring et al., 2021) for side-by-side comparisons of article sets, higher speed for the creation
519 of collections, API access to the data, and new filters based on a yet to be created taxonomy of text reuse
520 types.

6 CONCLUSION AND FUTURE WORK

521 In this paper, we have presented the prototype of the Text Reuse at Scale interface, to our knowledge the
522 first interface to integrate text reuse data with other forms of semantic enrichment. We argue that it enables
523 a versatile and scalable exploration of intertextual relations in historical newspapers. The interface was
524 developed as part of the *impresso* project and combines powerful search and filter operations with close
525 and distant reading perspectives. We reported on high-level research objectives and common user tasks
526 for the analysis of historical text reuse data, and presented the interface together with the results of a user
527 evaluation.

528 We have shown how the integration of text reuse data with semantic enrichment (content type, language,
529 topics, named entities) has proved advantageous: firstly, as a means of effectively filtering for relevant sets
530 of text reuse data; secondly, to help identify different types of text reuse; and thirdly, to provide overviews
531 of the content of text reuse data. We have also demonstrated the interface's ability to retrieve text reuse
532 based on temporal patterns, for example, distinguishing between content that spreads rapidly and news
533 that is rediscovered after long time periods. Examples include the coverage of the bombing of Hiroshima,
534 the reprinting of the same article on the event anniversary, and the reprinting of the 1945 article in 1998.
535 Further, the interface reveals systematic co-publication patterns, independent of content, with the *Journal*
536 *de Geneve* and *Gazette de Lausanne* as the main examples. We have also shown its ability to give insights
537 into text reuse cluster(s) based on topics, and to shift between distant and close reading operations. Finally,
538 we have shown its usage for a critical assessment of corpora and variations in the performance of TRD
539 based on the distributions of passages, cluster sizes, and lexical overlap over time.

540 At this stage of development, the Text Reuse at Scale interface supports many but not all of the previously
541 discussed tasks for the exploration of historical text reuse data (see Table 2). Future development will
542 address the integration with the *impresso* Inspect & Compare component to enable side-by-side comparisons
543 of article sets which contain text reuse passages (in support of Task 4 - Compare clusters and Task 7 -
544 Connections), and better support for temporal dimensions of text reuse data 1. We will also take into
545 account the need to improve the readability and documentation of the statistics tab and work to resolve the
546 difficulties observed in the passages tab, together with minor bugs discovered during the evaluation.

547 Finally, text reuse detection tools to date still mostly operate at the "surface" level of language, i.e. they
548 detect repeating patterns at the character and/or token level, but not at the semantic level. This means
549 that they do not recognize translation as reuse. Recent advances in machine translation, as well as in
550 multilingual language modelling and semantic indexing, may provide solutions in this direction and give
551 an additional boost to research in transnational perspectives.

CONFLICT OF INTEREST STATEMENT

552 The authors declare that the research was conducted in the absence of any commercial or financial
553 relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

554 DG was responsible for UI and UX development and integration into the *impresso* interface. MD, MR,
555 and DG contributed to the conception and design of the user workshop. ME and MR revised the *impresso*
556 technical infrastructure to suit the needs of the interface. MD, MR and DG developed the evaluation

557 procedure. MD and MR wrote the first draft of the manuscript. PA, KB and BD wrote sections of the
558 manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

559 The workshop was funded by the Luxembourg Centre for Contemporary and Digital History (C²DH). This
560 work is building on the research project “*impresso*. Media Monitoring of the Past” funded by the Swiss
561 National Science Foundation (SNSF) under grant ID CR- SII5.173719.

ACKNOWLEDGMENTS

562 We wish to express our gratitude to Fred Pailler, Matteo Romanello and Jana Keck for their contributions
563 to the workshop and to the Luxembourg Centre for Contemporary and Digital History (C²DH) for agreeing
564 to host it.

SUPPLEMENTAL DATA

565 Supplementary Material should be uploaded separately on submission, if there are Supplementary Figures,
566 please include the caption in the same file as the figure. LaTeX Supplementary Material templates can be
567 found in the Frontiers LaTeX folder.

DATA AVAILABILITY STATEMENT

568 The interface is accessible here: <https://impresso-project.netlify.app/text-reuse/>.

REFERENCES

- 569 Büchler, M., Burns, P. R., Müller, M., Franzini, E., and Franzini, G. (2014). Towards a Historical Text
570 Re-use Detection. In *Text Mining*, eds. C. Biemann and A. Mehler (Springer International Publishing),
571 Theory and Applications of Natural Language Processing. 221–238
- 572 Cordell, R. (2015). Reprinting, Circulation, and the Network Author in Antebellum Newspapers. *American*
573 *Literary History* 27, 417–445. doi:10.1093/alh/ajv028
- 574 Düring, M., Kalyakin, R., Bunout, E., and Guido, D. (2021). *Impresso* Inspect and Compare. Visual
575 Comparison of Semantically Enriched Historical Newspaper Articles. *Information* 12, 348. doi:10.3390/
576 info12090348
- 577 Keck, J., Oiva, M., and Fyfe, P. (2022). Lajos Kossuth and the Transnational News: A Computational
578 and Multilingual Approach to Digitized Newspaper Collections. *Media History* 0, 1–18. doi:10.1080/
579 13688804.2022.2146905
- 580 Liebl, B. and Burghardt, M. (2020). “Shakespeare in the Vectorian Age” – An evaluation of different
581 word embeddings and NLP parameters for the detection of Shakespeare quotes. In *Proceedings of the*
582 *The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences,*
583 *Humanities and Literature* (Online: International Committee on Computational Linguistics), 58–68
- 584 Manjavacas, E., Long, B., and Kestemont, M. (2019). On the Feasibility of Automated Detection of
585 Allusive Text Reuse. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics*
586 *for Cultural Heritage, Social Sciences, Humanities and Literature* (Minneapolis, USA: Association for
587 Computational Linguistics), 104–114. doi:10.18653/v1/W19-2514
- 588 Marxen, L. (2023). *Where Did the News Come From? Detection of News Agency Releases in Historical*
589 *Newspapers*. Master’s thesis, École Polytechnique Fédérale de Lausanne, Lausanne
- 590 Moritz, M. and Steding, D. (2018). Lexical and Semantic Features for Cross-lingual Text Reuse
591 Classification: An Experiment in English and Latin Paraphrases. In *Proceedings of the Eleventh*
592 *International Conference on Language Resources and Evaluation (LREC 2018)* (Miyazaki, Japan:
593 European Language Resources Association (ELRA)), 1976–1980
- 594 Oiva, M., Nivala, A., Salmi, H., Latva, O., Jalava, M., Keck, J., et al. (2020). Spreading News in 1904.
595 *Media History* 26, 391–407. doi:10.1080/13688804.2019.1652090

- 596 Paasikivi, S., Salmi, H., Vesanto, A., and Ginter, F. (2022). Infectious media: Cholera and the circulation
597 of texts in the finnish press, 1860–1920. *Media History* 0, 1–22. doi:10.1080/13688804.2022.2054408.
598 Publisher: Routledge _eprint: <https://doi.org/10.1080/13688804.2022.2054408>
- 599 Paju, P., Rantala, H., and Salmi, H. (2023). Towards an ontology and epistemology of text reuse. In
600 *Reflections on tools, methods and epistemology*, eds. E. Bunout, M. Ehrmann, and F. Clavert (De Gruyter
601 Oldenbourg). 253–274. doi:doi:10.1515/9783110729214-012
- 602 Paju, P., Salmi, H., Rantala, H., Lundell, P., Marjanen, and Vesanto, A. (2022). Textual migration across
603 the baltic sea : Creating a database of text reuse between finland and sweden. In *Proceedings of the 6th*
604 *Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)*, eds. K. Berglund,
605 M. La Mela, and I. Zwart (The 6th Digital Humanities in the Nordic and Baltic Countries Conference
606 (DHNB 2022)), CEUR Workshop Proceedings. 361–369. Publisher: CEUR-WS.org
- 607 Romanello, M., Berra, A., and Trachsel, A. (2014). Rethinking Text Reuse as Digital Classicists.
608 In *9th Annual International Conference of the Alliance of Digital Humanities Organizations, DH*
609 *2014, Lausanne, Switzerland, 8-12 July 2014, Conference Abstracts* (Alliance of Digital Humanities
610 Organizations (ADHO))
- 611 Romanello, M. and Hengchen, S. (2020). Detecting Text Reuse with Passim. *The Programming Historian* ,
612 /doi:10.46430/phen0092
- 613 [Dataset] Romanello, M. and Snyder, R. (2017). Cited Loci of the Aeneid : Searching through JSTOR’s
614 content the classicists’ way. (Blog post)
- 615 [Dataset] Rosson, D., Mäkelä, E., Vaara, V., Mahadevan, A., Ryan, Y., and Tolonen, M. (2023). Reception
616 Reader: Exploring Text Reuse in Early Modern British Publications. (Pre-print). doi:arXiv:2302.04084
- 617 Salmi, H., Paju, P., Rantala, H., Nivala, A., Vesanto, A., and Ginter, F. (2020). The reuse of texts in Finnish
618 newspapers and journals, 1771–1920: A digital humanities perspective. *Historical Methods: A Journal*
619 *of Quantitative and Interdisciplinary History* 0, 1–15. doi:10.1080/01615440.2020.1803166
- 620 Salmi, H., Rantala, H., Vesanto, A., and Ginter, F. (2019). The Long-Term Reuse of Text in the Finnish
621 Press, 1771-1920. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, eds.
622 C. Navarretta, M. Agirrezabal, and B. Maegaard (Copenhagen, Denmark: CEUR Workshop Proceedings),
623 vol. 2364, 253–273. doi:https://ceur-ws.org/Vol-2364/36_paper.pdf
- 624 Scheirer, W., Forstall, C., and Coffee, N. (2016). The sense of a connection: Automatic tracing of
625 intertextuality by meaning. *Digital Scholarship in the Humanities* 31, 204–217. doi:10.1093/llc/fqu058
- 626 Smith, D. A., Cordell, R., and Dillon, E. M. (2013). Infectious texts: Modeling text reuse in nineteenth-
627 century newspapers. In *2013 IEEE International Conference on Big Data*. 86–94. doi:10.1109/BigData.
628 2013.6691675
- 629 Smith, D. A., Cordell, R., and Mullen, A. (2015). Computational Methods for Uncovering Reprinted Texts
630 in Antebellum Newspapers. *American Literary History* 27, E1–E15. doi:10.1093/alh/ajv029
- 631 Thérenty, M.-E. and Venayre, S. (2021). *Le monde à la une. Une histoire de la presse par ses rubriques*
632 (Anamosa), illustrated édition edn.
- 633 Verheul, J., Salmi, H., Riedl, M., Nivala, A., Viola, L., Keck, J., et al. (2022). Using word vector models to
634 trace conceptual change over time and space in historical newspapers, 1840–1914. *Digital Humanities*
635 *Quarterly* 016. doi:<https://www.digitalhumanities.org/dhq/vol/16/2/000550/000550.html>
- 636 Vesanto, A., Nivala, A., Rantala, H., Salakoski, T., Salmi, H., and Ginter, F. (2017). Applying BLAST to
637 Text Reuse Detection in Finnish Newspapers and Journals, 1771–1910. In *Proceedings of the NoDaLiDa*
638 *2017 Workshop on Processing Historical Language*. 54–58. doi:<https://aclanthology.org/W17-0510>
- 639 Walma, L. W. B. (2015). Filtering the “news” : Uncovering morphine’s multiple meanings on delpher’s
640 dutch newspapers and the need to distinguish more article types. *Tijdschrift voor Tijdschriftstudies*
641 doi:<http://dSPACE.library.uu.nl/handle/1874/324205>
- 642 Yousef, T. and Janicke, S. (2021). A Survey of Text Alignment Visualization. *IEEE Transactions on*
643 *Visualization and Computer Graphics* 27, 1149–1159. doi:10.1109/TVCG.2020.3028975