impresso Text Reuse at Scale

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

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

Marten Düring [1,*], **Matteo Romanello** [2], **Maud Ehrmann** [3], **Kaspar Beelen** [4], **Daniele Guido** [1], **Brecht Deseure** [5], **Estelle Bunout** [1], **Jana Keck** [6], **and Petros Apostolopoulos** [1]

[1] *Luxembourg Centre for Contemporary and Digital History (C2DH)*
[2] *University of Lausanne*
[3] *École polytechnique fédérale de Lausanne (EPFL)*
[4] *School of Advanced Study, University of London*
[5] *Royal Library of Belgium*
[6] *German Historical Institute Washington*

Correspondence*:
Marten Düring
marten.during@uni.lu

ABSTRACT

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

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

1 INTRODUCTION

Text reuse detection (TRD), best known for its capacity to detect plagiarism, is a powerful and popular technique to identify "meaningful reiteration[s] of text, usually beyond the simple repetition of common language" (Romanello et al., 2014). TRD typically identifies text segments (or *passages*) that are reused in different units of a corpus and groups them automatically into text reuse *clusters*. In the domain of Digital Humanities research, it is commonly applied to assist the study of intertextuality in literary texts; in fact, phenomena such as quotations, allusions and paraphrases can all be considered instances of text reuse. Not only the presence but also the frequency of reuse can be meaningful: the frequency with which a text is quoted by later authors, for example, can be taken as a useful indicator to study the literary or scholarly reception of that text. Beyond the realm of literary studies, TRD has been employed to trace and better understand patterns in content production — especially historical newspapers — as demonstrated in projects such as OceanicExchanges (Oiva et al., 2020; Keck et al., 2022) and Viral Texts (Cordell, 2015). Indeed, TRD allows for capturing phenomena that are frequent in journalistic texts such as the repurposing of content (with or without variations), as well as the viral circulation of news.

- why the press - sexy example

The project "*impresso* - Media Monitoring of the Past" (2017-2020)[1] detected text reuse within a corpus of Swiss and Luxembourgish newspapers alongside other forms of semantic enrichment (e.g. based on topic modelling, named entity recognition, the detection of content type and language). The *impresso* application[2] supports historians and other humanities researchers with powerful search, filter and discovery functionalities for the exploration of the enriched data. In contrast to most applications for the exploration of historical newspapers, *impresso* is not limited to search and filtering operations alone, but offers a set of integrated tools for the exploration of semantic enrichments such as image similarity, n-grams or named entities. The *impresso* application is generic in the sense that it not focused on specific use cases or research interests, but supports a wide variety of different use cases. This includes, for example, advanced search, the visualisation-aided comparison of large user-generated article collections and the creation of research datasets for further processing outside the application. Parallel to supporting a variety of exploratory workflows, *impresso* publishes accompanying datasets in dedicated data repositories[3].

In this paper we present the Text Reuse at Scale prototype interface for the visualisation-aided discovery and scalable reading of text reuse data in historical media. Scalable reading is understood as the seamless shift between close and distant reading views. In the case of newspapers, close reading corresponds to either the inspection of individual text reuse clusters and the passages they contain or the study of the articles to which they belong. Distant reading refers to visualisations of the distributions of text reuse measures, metadata and semantic enrichments which should be configurable to take into account various filtering operations. We describe generic (media) historical research objectives and identify a list of generic tasks for the exploration of text reuse data. In its final form, the prototype will achieve a close integration of text reuse data with other forms of semantic enrichment and will be integrated with the existing application.

Prototype-design and tasks were partially informed by the outcomes of a 2-day workshop organised by the *impresso* team. This event brought together a group of 10 researchers (within and outside the *impresso* project) and included professionals from various disciplines, such as design, natural language processing, data science and (media) history. The workshop aimed at exploring several scenarios for the usage of text reuse data in historical research, and consisted of presentations on the current state of research on text reuse in the digital humanities, followed by reflections on virality as a historical concept and a report on the value of text reuse data for the detection of historical event coverage. One of the workshop's outcomes was a list of historical research objectives (also in light of earlier work), associated tasks and three design mockups of potential applications.

The structure of the paper roughly follows our process in the creation of the interface: Section 2 positions our work in the current state of the art. It introduces the methods and tools for TRD and discusses recent advancements and remaining challenges for the detection and (visualisation-aided) exploration of text reuse in historical newspapers. It concludes with a brief overview of the *impresso* text reuse data. Section 3 concentrates on historical research interests in text reuse data. We identify five high-level objectives for research in (media) history and 11 generic tasks which derive from these objectives. In Section 4 we showcase the prototype interface and map it to the specific tasks using case studies. Section 5 reports on the results of an evaluation undertaken by 13 users. Section 6 closes the paper with an outlook on future work.

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

This section situates our work in the current state of the art in TRD for humanities research. We begin with an overview of tools and methods, as well as current directions in the visualisation-aided exploration of text reuse data; we conclude with an in-depth description of TRD in the context of the *impresso* project.

2.1 What is text reuse and how is it detected?

Methods for TRD are deeply shaped by the disciplines in which they emerged. Since text reuse in literary texts is often more subtle than the mere repetition of words from the target text (e.g. in the case of

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

[2] https://impresso-project.ch/app

[3] https://zenodo.org/communities/impresso/

paraphrase, allusion, translation or parody), the main challenge tackled by research has been how to go beyond lexical similarities in order to capture similarity in syntax, content or metrical structure (Büchler et al., 2014; Moritz and Steding, 2018; Scheirer et al., 2016). In the design of TRACER[4] Büchler et al. (2014) have addressed this subtlety of text reuse in literary texts by striving to give users access to a wide array of Information Retrieval (IR) algorithms, as well as direct access to the tool's output at the each step of the processing chain. Moreover, recent studies have investigated the usefulness of sentence and word embeddings, especially with respect to detecting these more allusive forms of text reuse (Manjavacas et al., 2019; Liebl and Burghardt, 2020), finding that they do not bring substantial advantages over traditional IR techniques.

On the other hand, the challenges of detecting text reuse in the newspapers domain are quite different. The substantial amount of OCR noise present in digitised newspapers asks for fuzzy methods that are resilient to differences between two or more copies of the same textual content. Moreover, the scale of materials — with corpora that can be several orders of magnitude bigger than those in the literary domain — led to the development of efficient and scalable methods. As a matter of fact, methods that were developed for TRD in the newspapers domain had to deal with both challenges, namely OCR noise and scalability. Vesanto et al. (2017) adapted the Basic Local Alignment Search Tool (BLAST) algorithm, originally developed for the alignment of biomedical sequences, to the task of character alignment.[5] An alternative approach to TRD consists in performing alignments between documents at the level of longer sequences of words, a.k.a. n-grams, instead of individual characters. This was the approach followed by Smith et al. (2015) whose TRD algorithm, implemented in the tool passim[6], uses n-gram-based filtering to reduce the number of text passage pairs to compare — thus achieving scalability — and combines it with local and global alignment algorithms to handle gaps and variants in longer sequences of aligned texts.

Finally, in terms of existing tools for TRD, we find the textreuse package (R)[7], TextPAIR v. 2 (Python) and Tesserae (Perl/PHP), in addition to the previously mentioned TRACER (Java), BLAST (C++/Python) and Passim (Java/Python). Despite the abundance of 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.

2.2 Interactive visualisations of text reuse

Text reuse instances within a corpus can be analysed and visualised at various levels of depth, which are directly linked to the purpose of the exploratory analysis:

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

[4] https://www.etrap.eu/research/tracer/

[5] The Python package `textreuse-blast` provides an implementation of this method.

[6] `https://github.com/dasmiq/passim`

[7] `https://docs.ropensci.org/textreuse`

Depending on the research focus and questions at hand, one or more of these levels will be considered. Generally speaking, distant reading approaches tend to privilege analysis of corpus-level and document-level text reuse, while the close reading approach is more concerned with cluster-level and passage-level reuses. Existing interactive visualizations of text reuse tend to support multiple levels of analysis at once and often allow users to seamlessly move between levels. The techniques used to visualize text reuse overlap substantially with those used to represent text alignment in other scenarios, e.g. translation alignment, collation of sources, etc. (Yousef and Janicke, 2021).

The interface developed for the *Graph – Text reuse in rare books*[8] project constitutes a compelling example of interfaces supporting multiple levels of exploration. It was developed to enable the exploration of text reuse passages extracted from a corpus of 1,300 OCRed rare books. Firstly, corpus-level text reuse is represented as a graph where two nodes (books) are connected when they contain reused passages, with the additional possibility of ordering the graph by time (of publication). Secondly, a static alluvial diagram allows readers to inspect more closely document-level reuse between pairs of books; this is especially useful to understand flows of reused text across books. Lastly, a facsmile side-by-side view of pairs of books permits to focus on passage-level reuse; such a viewer is not aimed at highlighting differences between reuse passages, but rather at displaying them in their original context (especially meaningful in the case of rare books).

Graph visualization of text reuse at corpus-level was used also in the context of the Viral Texts project, which studied virality in newspapers during the interwar period. One of the interactive networks visualization[9] developed in the project provides a bird's-eye-view of millions of text reuse passages, distilled into a graph showing how newspapers formed a network of reprints and content reuse. Node size and color are used respectively to express node centrality and grouping into community clusters, while the thickness of lines connecting nodes indicates the number of shared reprints. Beside network visualization, geographical maps were employed to support cluster-level analysis, as they allow for visualizing at a glance the geographical distribution of reprints of a certain text (Cordell, 2015).

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

2.3 Text reuse detection in the *impresso* project

We used the open source software Passim (Smith et al., 2015) to automatically detect text reuse within the *impresso* corpus, consisting of 47.8 million content items[10]. The output of Passim are clusters, namely groups of passages (or witnesses), from different newspapers, that share a common text span—the reused passage—of varying length (see Figure 1). The reason for choosing Passim over existing alternatives[11] was its ability to scale up, guaranteed by the software's parallel computing architecture. Preliminary tests on the *impresso* corpus showed that Passim's fuzzy alignment algorithm was able to detect reuse despite the presence of (moderate) OCR noise.[12]

2.3.1 Text Reuse detection and processing

As a pre-processing step, we ran Passim in boilerplate detection mode; this allowed us to identify — and, later on, filter out — boilerplate content present in our corpus, i.e. portions of text that get repeated within the same newspaper in a time window of a month (as opposed to reuse across different newspapers). All content items where boilerplate text was detected were filtered out from Passim's input. This pre-processing

[8] `https://graph-rare-books.ethz.ch/`

[9] `http://networks.viraltexts.org/1836to1860/`

[10] Content item is the term we use to refer to newspaper contents below the page level. Typically, pages image are segmented and classified into finer-grained content units such as articles, advertisements, images, tables, weather forecasts, obituaries, etc. – this is precisely what is referred to by content items. See also `https://impresso-project.ch/news/2020/01/23/state-corpus-january2020.html`.

[11] See Romanello and Hengchen (2020) for a list of available TR detection software.

[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.

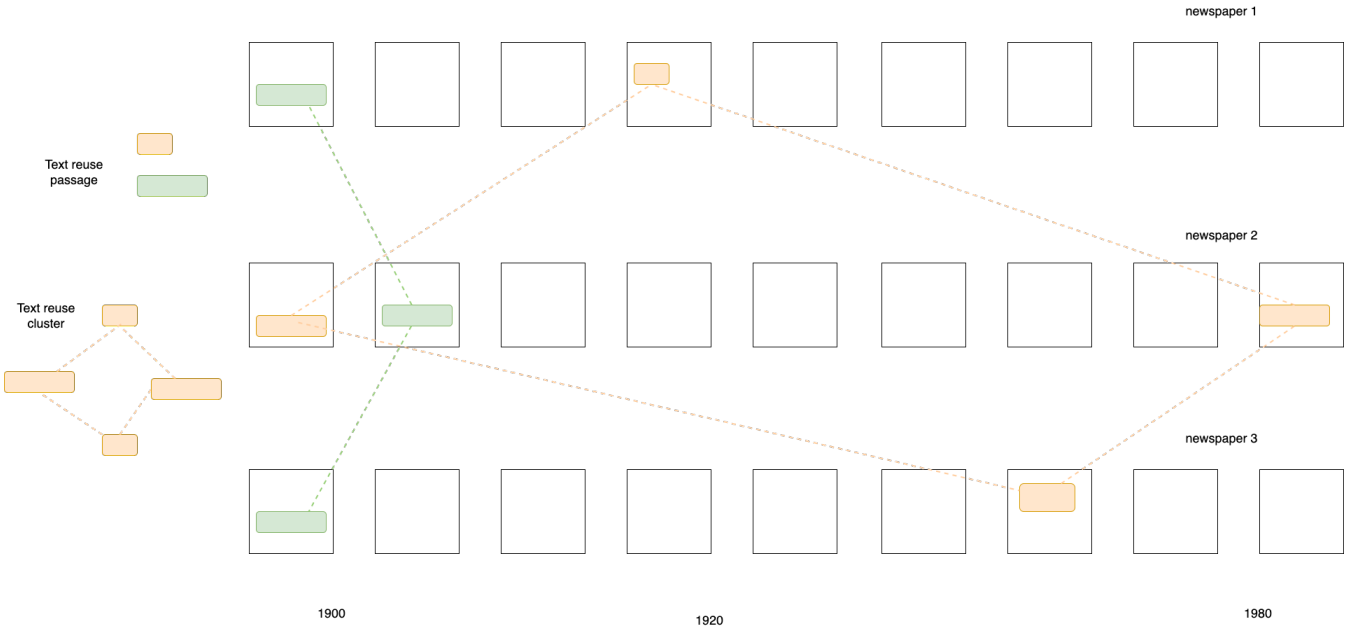

Figure 1. Schematic view of text reuse clusters and passages extracted from a newspaper corpus.

step allowed for reducing the final number of detected text reuse clusters by removing some noise from the input data. After boilerplate filtering, we extracted 6,177,815 text reuse clusters, for a total of 16,099,821 reused passages, meaning that roughly 17% of all content items in the corpus are part of at least one text reuse cluster.

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

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

2.3.2 Integration of Text Reuse data in *impresso*

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

One of the main difficulties we faced in the integration of text reuse into the *impresso* application is the scale of data, and more specifically how to enable an effective exploration of millions of detected clusters. Our approach to this problem consisted in providing users with as many filters as possible, as a powerful way of sifting through the large number of clusters extracted by Passim. For example, users interested in long-term reuse (Salmi et al., 2019) of newspaper contents—i.e. articles that get reprinted over and over again, within a relatively long period of time—can refine their query by setting a filter on the cluster's time span, thus keeping only clusters consisting of articles that cover a time span of e.g. ten years.

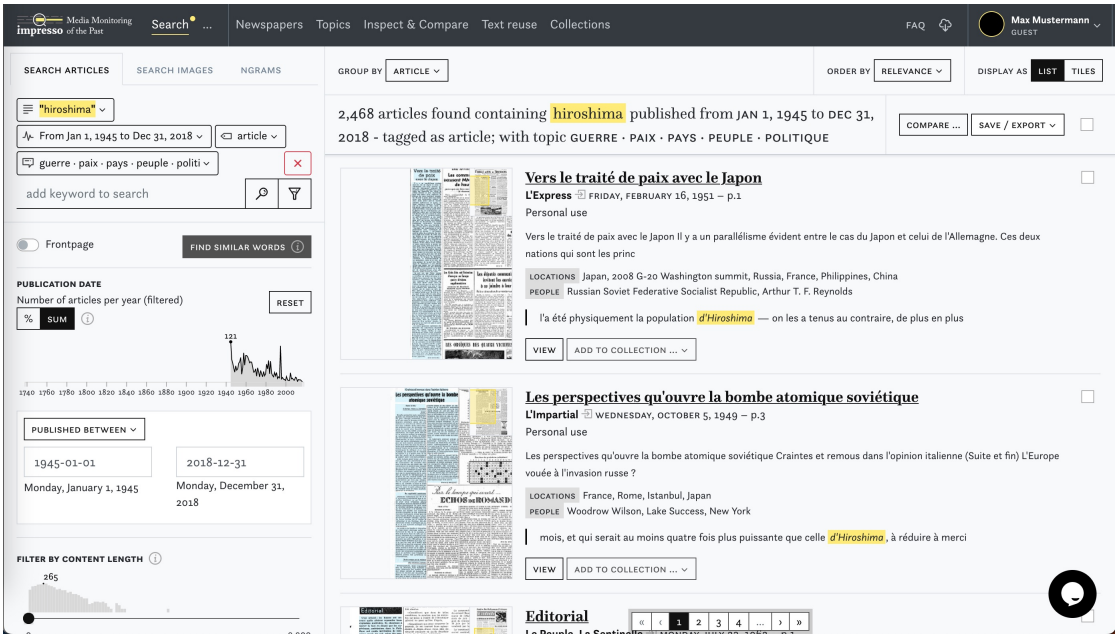

Figure 2. Screenshot of the *impresso* application for the exploration of semantically enriched historical newspapers.

This previous explorer mostly supports cluster- and document-level research with a basic set of search capabilities and filters with no distant reading perspectives. Development of the new version was motivated both by the opportunity to fully utilise the available enrichments as well as the prospect to support additional use cases including passage- and corpus-level research. To this end, text reuse and semantic enrichment data have been integrated, i.e. named entities, topics, and content item types (when available) were aligned with text reuse passages and clusters.

3 HISTORICAL RESEARCH AND TEXT REUSE

Following the overview of the state of the art in TRD for historical text, this Section discusses a second important prerequisite for the design of the text reuse interface: the motivations and needs from the perspectives of historians interested in newspaper collections. As we have seen, TRD has a variety of

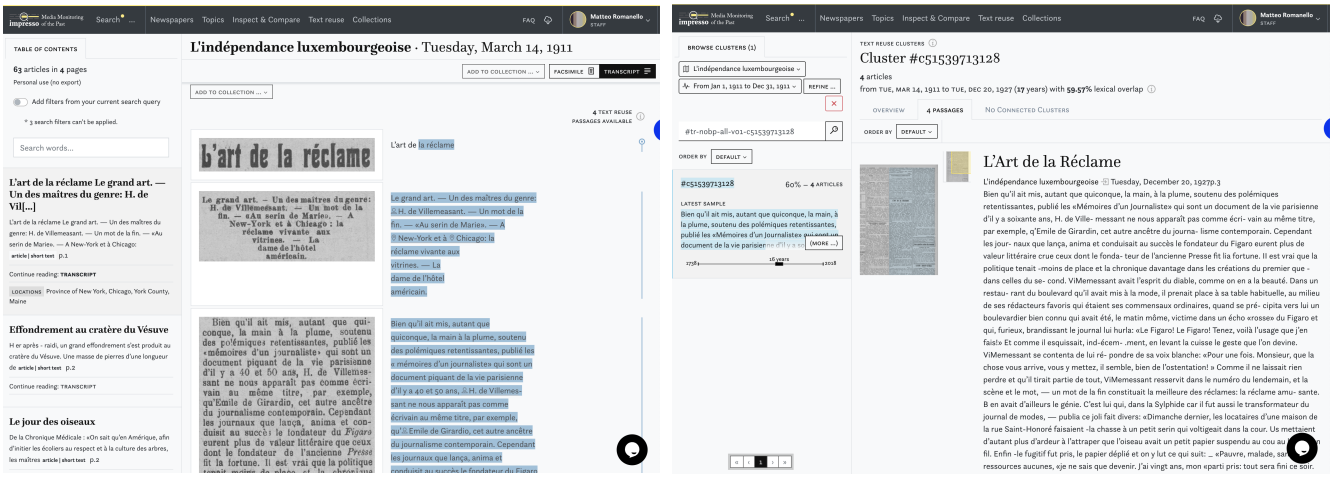

Figure 3. Display of text reuse in the *impresso* application's article reading view (left) and first version of the text reuse explorer (right).

applications for the analysis of large collections of historical text. From a historical perspective, past and present media can be thought of as complex communication networks which manifested themselves in form of interactions between different stakeholders such as individual journalists and press agencies. Connectivity in such networks was determined by many different factors, including geography, politics, technology, communication infrastructures, languages, commercial interests and not least contemporary tastes. Different types of text circulated within these networks and the detection of these flows allows us to reconstruct the emergence and dissolution of links between stakeholders across time and space. Studies into copy-paste journalism, plagiarism, paraphrasing, literary and scholarly citation, the dissemination of specific discourses and similar phenomena in historical text corpora all stand to benefit from text reuse data.

Apart from the notion of mere connectivity through information exchange within large scale document collections, other work focused on the different types of texts and content genres which circulated: Jokes, adverts, boilerplates, speeches, religious texts but also short stories and re-prints of book segments are prominent examples of different logics and motivations behind text reuse. Paju et al. (2023) point to the different speeds in which text reuse occurs. They propose to distinguish between rapid (within 1 year) and mid-range (up to 50 years). Slow text reuse (up to 140 years), anecdotal evidence suggests, is typically tied to conscious re-prints of archived materials. In Table 1 we classify three types of temporality we encounter in the study of text reuse in historical media: duration (which includes rapid and slow text reuse), virality, and rhythm. Another perspective on text reuse considers the breadth of content circulation within the media and raises the question, to which extent once popular ideas and social practices re-emerge as measurable instances of text reuse.

We distinguish between five high-level historical research objectives in the study of historical media and 11 common tasks which derive from these objectives.

3.1 High-level historical research objectives

Within these research objectives, we distinguish between media-centric, content-centric, and data-centric perspectives. Media-centric perspectives seek to understand the functioning and evolution of the press as an information production and dissemination system. Content-centric perspectives use historical media coverage to approximate the reconstruction of historical public mindsets. They concentrate on the reflection and representation of past discourses. Data-centric perspectives finally regard text reuse as a means to clean and evaluate textual data for statistical analysis. As we will show below, the prominence of each of these perspectives varies among the objectives.

3.1.1 (Trans-) National Media Ecosystems

With the increasing availability of digitised newspaper collections, media historians have begun to broaden the scope of their analyses: Attention shifts from the in-depth reconstruction of the history of individual titles to a view that sees them as part of a transnational media ecosystem which facilitates the creation and dissemination of information. Current research seeks to understand the functioning of this ecosystem and the agents which shaped it through facilitation and control. This includes questions regarding the underlying ideological, commercial and financial structures that have shaped historical media ecosystems. Previous research has, for example shown the relevance of telegraph lines and railways in the spread of information within the United States (Smith et al. (2013)) and pointed to individual cities as information dissemination hubs (Cordell (2015); Salmi et al. (2020)). Other work has studied multilingual information flows in transnational perspective to examine the connections, gaps and silences in the system, and the press as a site for manipulation (Keck et al., 2022; Paju et al., 2023; Paasikivi et al., 2022).

The increasing availability of text reuse data for different countries will also allow more systematic comparisons of (re-) printing cultures in transnational perspectives. One agent of particular interest in this regard are internationally operating press agencies with their ability to disseminate content across borders and languages nearly simultaneously. Such a transnational perspective reveals the ways news is altered and contextualised as it travels. Scrutinising the reproductions of text——investigating additions and deletions as traces of adaption——help us understand what was considered common knowledge in one (national) context but not in the other. It foregrounds how perceptions and descriptions are adapted to novel audiences.

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

Type	*Description*	*Measures*	*Examples*
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 Hindeburg 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.

However interesting, multilingualism itself remains a major hurdle for studying text reuse in transnational context. Text reuse tools still mainly operate on the "surface" level of language, i.e. detect repeating patterns at the character and/or token level, but not at the semantic level (i.e. they do not recognize translation as reuse). Luckily, recent advances in machine translation as well as in multilingual language modelling and semantic indexing may provide solutions in this direction. If the technical bottlenecks eventually are removed, multilingual text reuse would enable novel computational approaches to translation studies.

3.1.2 Newspaper Content as Bricolage

Newspaper discourse is not necessarily very original or innovative (linguistically). Besides containing repeated tropes or cliches, many genres such as weather forecasts or sport reporting, often operate within strict constraints and happen to be almost formulaic at times. **Therenty et al. and Walma paid special attention to the relations between these** genres: how did content travel between them? Thérenty and Venayre (2021); Walma (2015) In general, articles emerge through a process of creative re-use and re-appropriation. Whole fragments, sentences and quotes are often fitted within novel contexts. In this sense, newspaper content emerges through a process of what could be called *bricolage*, in which texts are soldered together from existing fragments and textual patterns. Put differently, articles are always a construction or creation harvested from a diverse range of available things.

This objective investigates text reuse through the angle of compilation and the evolving forms of content production. We can employ reuse measure to encode textual relations and connections, and thereby enable researchers to critically disentangle the genesis of newspaper content. What type of reuse is meaningful depends on the research question. This implies that data entry points such as applications/API should be agnostic in this respect. Moreover, the concept of *bricolage* opens up a graded, more nuanced approach, to the study of text reuse: it foregrounds how the creation of news content emerges in a complex process of multiple text transformations, compilations and innovations. Newspaper titles operate within a media ecosystem compiling and recreating content harvested from the "grid" (press agencies, or newspapers) and merging it self-generated content (ads, journalistic work, external contributors etc).

3.1.3 Historicising Virality

Virality is more commonly understood as a phenomenon of the internet era and often associated with three characteristics: High speed, high volume and the ability to adapt or to be "contagious" in the sense of rapid spreading. Paju et al. (2022) have used text reuse data in an attempt to measure and compare different degrees of virality for content that was republished within days or weeks. They define a virality score based on the number of titles within a cluster, the number of unique printing locations and the distance between the first and last passage publication date in days. They show that different types of genres and content qualify for different types of repetition: An advertising for Finnish cigarettes in 1916 constituted the most

viral content in their corpus while institutional announcement, literary and religious texts often fall in the mid-range category.

Such measures may yield additional insights into the functioning and the comparison of historical media ecosystems, e.g. by revealing which types of text circulated more efficiently than others within and beyond national boundaries and which institutions and individuals were responsible for their creation and dissemination. Virality may also offer complementary insights into historical information dissemination shaped by information and transportation infrastructures and geography as well as the reception or rejection of content, e.g. on the grounds of religious and political ideology or censorship.

3.1.4 Tracing Historical Events

The press is a system of knowledge production and representation that not only presents events to the public, but also places them within a specific political, economic, social and cultural framework. This framework determines the way the public perceives and understands a historical event to an important extent. At the same time, such frameworks help to position the political, social, and cultural identities of individual newspaper titles. However, such identities are not stable, but may differ across time and space.

Text reuse helps to observe such frameworks in action. It not only allows scholars to reconstruct the spread of historical event coverage across newspapers and over time, but allows to explore how they perceived and represented them.

There are two ways to start in order to trace the coverage of historical events using text reuse. The first approach is bottom-up: One already knows what topic to examine and looks to reconstruct its coverage develops across time and space (Oiva et al., 2020). The second is top-down: TRD is applied to a given corpus as a means to identify media events. As an example of the latter, Keck et al. (2022) used newspaper collections from the United States, Britain, Germany, Austria, and Finland and TRD to identify global media events. Through this approach, they discovered a staggering number of articles that circulated during Hungarian Revolutionary Lajos Kossuth's tour of America to seek financial support from the U.S. for another revolution in Europe. His arrival in New York in December 1851 and his subsequent travels to Washington, DC sparked a proliferation of coverage and reprinted texts. Comparing text reuse across national and linguistic borders highlights the specific patterns and complexities of transatlantic news circulation, including pathways, reach, temporality, vagaries, and gaps. While this work illustrates the usefulness of TRD paired with data exploration by means of interactive visualization, it also emphasizes the benefits of international cooperation when working with multilingual datasets.

3.1.5 Capturing Historical Zeitgeist

Historical media can also be seen to capture attitudes, norms, beliefs, moods, and feelings of humanity at a given point in time and to thereby serve as a proxy for the study of a more general phenomenon: Zeitgeist. This entails the idea of similarity and parallel evolution: Texts which share characteristics, were produced independently under the influence of a prevailing Zeitgeist. This constitutes the border zone of what text reuse can capture. This Zeitgeist can manifest itself in different forms ranging from mental maps which informed the creation of editing of texts, adverts for (cultural) products and the mere existence of coverage of cultural practices. These manifestations are created using persistent and implicit templates which change their content over time - an example would be dance fads such as Polka or Macarena which are dominant at one point, but then slowly fade away. Related work looks at conceptual change over time (Verheul et al., 2022) or the cultural impact of Cholera epidemics (Paasikivi et al., 2022).

3.2 Tasks for the Exploration of Text Reuse in Historical Newspapers

We move now from high-level objectives towards their operationalisation in form of user tasks. Tasks are not directly linked to any of the objectives. They are rather building blocks which can be used to create individual workflows for the exploration of text reuse data. Table 2 gives an overview of how these relate to the different levels of analysis we describe in Section 2 and the current degree of support by the interface prototype.

Task 1: Obtain an overview of text reuse at the level of the corpus, collection or query. Before analysis, users need to determine whether or not a given corpus, corpus subset, query result or collection contains instances of text reuse. This task therefore provides an overview of the presence of text reuse in a selected dataset and describes its general distribution.

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

Occurrences of text reuse should be understood in relation to their properties. This can be facilitated by overviews of the distribution of newspaper metadata and semantic enrichments. These are, for example, time, newspapers, countries, content types, languages and named entities but also text reuse-specific measures such as lexical overlap, time span between publication dates, cluster size and number of passages.

Computing measures of spread (and inspecting outliers) offers additional insights in the distribution of text reuse data at different levels of granularity. This includes the inspection of largest/smallest clusters, clusters with the highest/lowest lexical overlap, the ability to filter for earliest/latest cluster in the corpus or the longest time span between publication dates and constellations of any of these measures.

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

For example, text reuse clusters detected in two co-publishing newspapers like *Gazette de Lausanne* and *Journal de Geneve* typically comprise two passages with a high lexical overlap and a time span between publication dates of 1 to 3 days. On the other side of the spectrum, clusters of slightly modified job adverts usually feature a large number of passages that overlap only partially and time deltas which can span years.

Task 3: Compare passages. This task concerns the comparison of two or more passages to reveal differences and similarities of the text they contain. A common motivation for such a comparison are editorial edits of texts under circulation, such as press agency dispatches. Comparisons can e.g. reveal adaptations to suit the political preferences of a newspaper's audience, clarifications - what is obvious to one set of readers may need additional explanation to others, but also unintended differences such as degeneration, for example caused by OCR errors.

Task 4: Compare clusters. Comparison is a powerful means to obtain insights. This task focuses on comparing sets of clusters based on the distribution of a) text reuse measures and b) metadata and semantic enrichments such as topics or named entities.

Task 5: Identify different types of text reuse. Text reuse encompasses different forms of reiterated text, including e.g. co-publication, template-based content such as adverts or TV programmes, and press agency reports. But we can also distinguish the date range between passage publication dates within a cluster, its virality (see Task 8), or the re-publication of content over time (see Table 1). Each of these phenomena maps to text reuse data characteristics and semantic enrichments offer a highly versatile approach to segment text reuse data into meaningful categories. For instance, cinema adverts should be characterised by large number of named entities (persons), topics associated with media content, and a high lexical overlap, a large cluster size and a smaller time spans. An example of a re-appearance would be a public service announcements which were reprinted regularly. Clusters like these are characterised by a high degree of lexical overlap and long time spans.

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

Task 7: Identify connections. This group of tasks concentrates on the relational aspect of text reuse data and information flows in the media. Previously discussed work on the flow of content between countries or titles fits into this task. Examples include the re-print of a newspaper article by a different newspaper after its original publication, an advertising campaign which relies on multiple newspapers simultaneously to gain visibility or regular co-publication agreements between newspaper titles.

Task 8: Detect and trace virality. This task corresponds to the pioneering work of Paju et al. (2022) and adds a measure of the efficiency and speed of content spreading within a newspaper corpus.

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

Task 10: De-duplicate content. This processing steps removes duplicate text for a corpus, e.g. to avoid over-representations due to highly circulated texts or recurrent elements such as adverts of boilerplate phrases.

Task 11: Data export. Data export allows further processing outside the constraints of an application. This, e.g., for network- or geo-spatial analyses or for further processing.

4 THE *IMPRESSO* TEXT REUSE AT SCALE INTERFACE

This section describes the prototype of the Text Reuse at Scale interface for the exploration of text reuse. It was developed by the *impresso* team and inspired by the aforementioned research objectives and associated tasks. Its title signals our ambition to offer scalable and versatile perspectives on text reuse data and to enable a variety of close and distant reading activities in conjunction with semantic enrichments. Development prioritised tasks (see Table 2) which fit the overall scope and design of the *impresso* application. This translates to a focus on discovery and exploration through search, filtering and comparisons. A more refined version of the interface will be integrated in the *impresso* application while the latest implementation is already available for testing.[13]

4.1 Main interface components

The interface consists of a search and filter pane on the left and three tabs in the centre (see Figure 4). This Section introduces these components and uses examples to illustrate their usage.

4.1.1 Search and Filter Pane

Figure 4a shows the search and filter pane. Users can compile queries using the full versatility *impresso*'s Search component together with a variety of filters. These include newspaper metadata, user generated article collections, text reuse clusters, and semantic enrichments (topics, language, content type and named entities). In addition, the component displays the distribution of passages, lexical overlap, cluster size over time as well as the time span between earliest and latest publication date (for details see Table 3).

Complementary modal dialogues as shown in Figure 5 (centre) allow users to select specific data ranges and display passages which match the selection. These close-up views serve as a bridge between distant and close reading perspectives. They allow users to quickly inspect passages associated, for example, with notable peaks in the distribution of lexical overlaps, cluster sizes, and time spans between publication dates.

Taken together, these search and filtering capabilities enable a highly versatile querying of the text reuse data. For example, filtering for time span between passage publication dates reveals different types of text reuse as described in Task 5 - Types. Following Paju et al.'s classification of time periods we find 13,980,938 passages which qualify as rapid (0-365 days), 1,516,190 passages which qualify as mid-range (1-50 years), and 59,265 passages which qualify as slow (50-200 years).

[13] `https://impresso-project.netlify.app/text-reuse/`

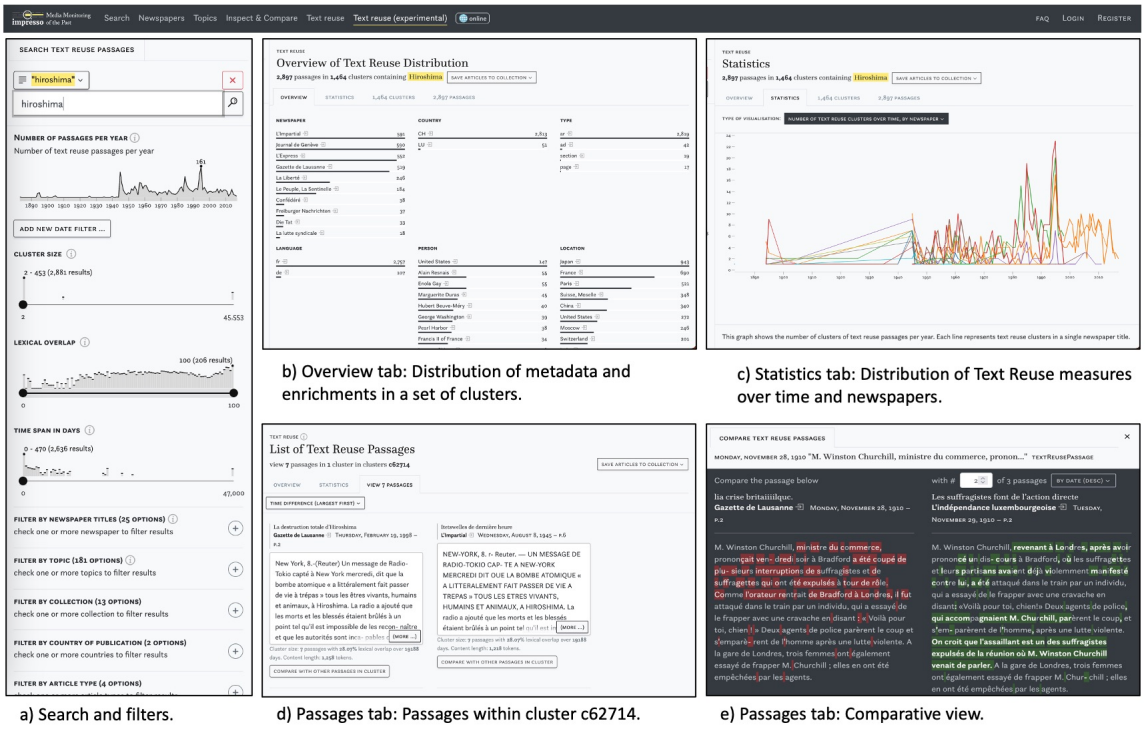

Figure 4. Main components of the text reuse interface: Search and Filter Pane (left) and the Overview, Statistics and Passages tabs (centre).

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

4.1.2 Overview Tab

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

In this instance, we learn that the vast majority of text reuse passages which contain *hiroshima* are linked to French-speaking content and were published in Switzerland. German-language content and content published in Luxembourg remain the exception. A closer look at the newspaper titles suggests that roughly 80% of these passages appear in just four newspapers. Unsurprisingly, the most prominent topics are associated with war, nuclear technologies and aviation. We also note 363 articles linked to media-related content.

4.1.3 Statistics Tab

The second tab titled *Statistics* visualises the distribution of text reuse measures relative to queries, intended to offer distant reading perspectives on text reuse data. A dropdown menu offers access to five views based on line charts and a matrix visualisation which are displayed in Figure 8 and are discussed in greater detail below.

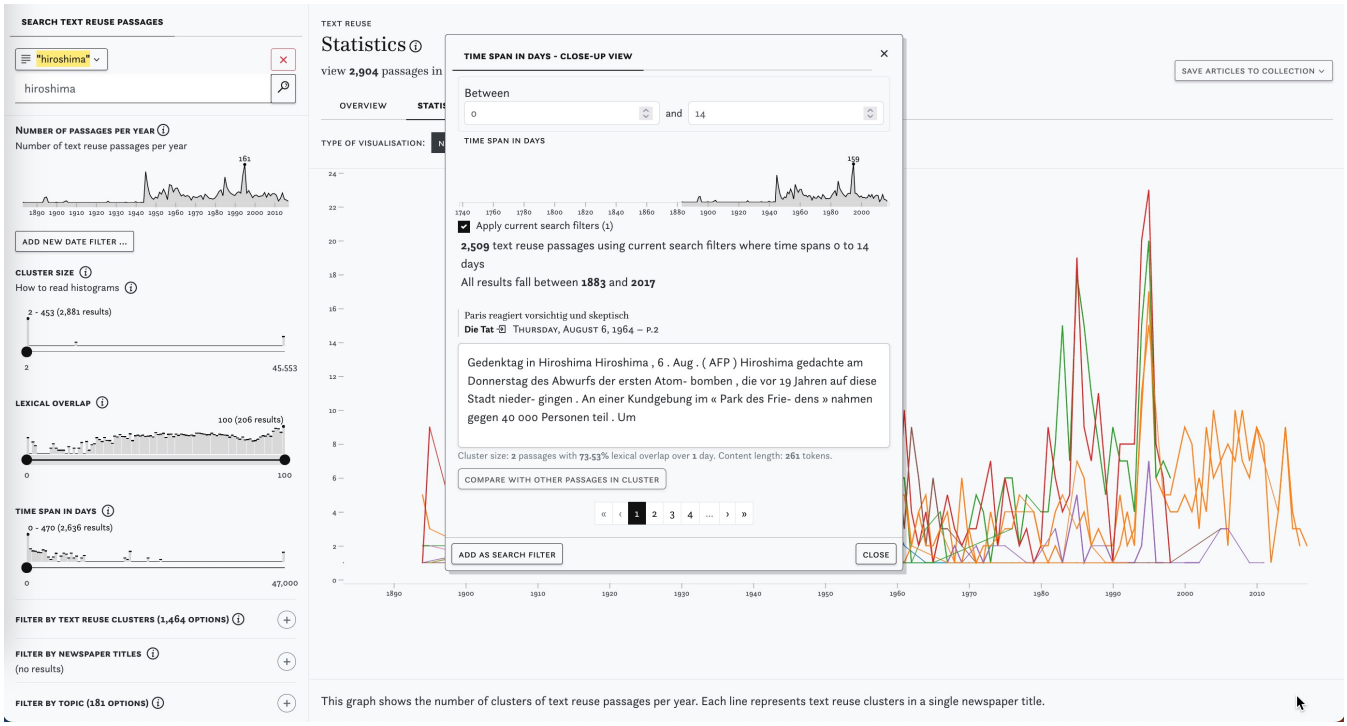

Figure 5. Screenshot of the search and filter pane with a keyword search for *hiroshima* (left) and the close-up view with a time span filter for 0 to 14 days (centre).

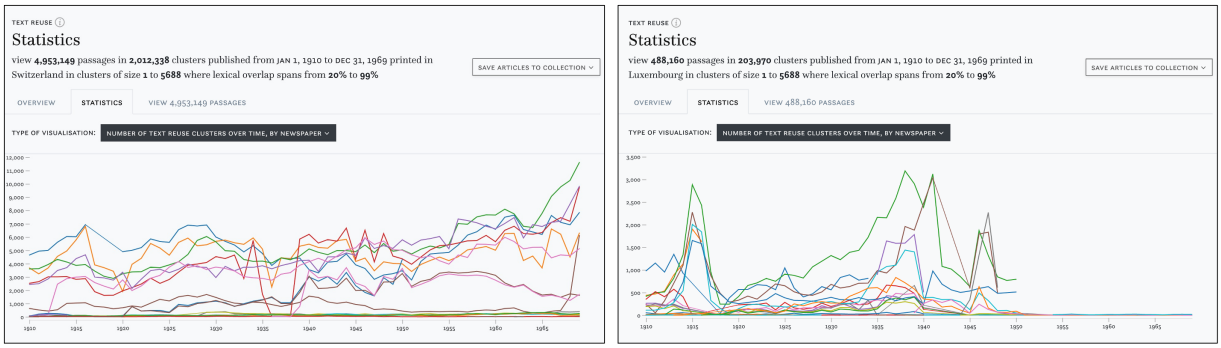

Figure 6. Number of clusters detected in Swiss (left) and Luxembourgish (right) newspapers 1910 - 1970.

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

As an example, we will compare the distribution of text reuse in Swiss and Luxembourgish newspapers. In the search and filter pane we set the time delta for 0 to 100 days. Lexical overlap is set to a moderately high range of 20-99% which should retrieve also reused text segments of smaller size embedded in a larger text. Finally, we exclude a dis-proportionally large cluster with 45.000 passages from the selection using the cluster size filter. This yields ca. 5.5 million passages which were detected during the period of observation. Looking at the distribution of cluster sizes in Switzerland in Figure 6 (left) suggests some variation between titles but otherwise no changes between the pre- and postwar period. In contrast, the Luxembourgish press (right) exhibits a growing numbers of passages since the 1930s and clear peaks in 1915 and during the Second World War followed by a stark decline after 1950 which can be explained with the composition of our newspaper corpus.

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

Measure	*Description*	*Implementation in interface prototype*
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.	Histograpmn 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 *impresso components such as Search.*

The **minimum, mean and maximum cluster sizes over time** are shown in Figure 8b. Overall, the number of text reuse clusters and passages rises constantly over time, parallel to the number of available content in the *impresso* corpus. For another example we make use of *impresso* Collections which store sets of articles based on either manual selection or querying, see Task 6 - Research corpora. Collections can also be used to store text reuse clusters - albeit with the caveat that not only passages but the entire articles in which they occur will be saved. Figure 7 shows a query for text reuse in adverts which are part of a large collection of articles surrounding nuclear power and linked to the topic *eau · énergie · gaz · électricité · air*. The peak in the year 1977 points to cluster c276252 which has captured 21 adverts in favour of nuclear power which where published in parallel in several Swiss newspapers.

The **minimum, mean and maximum cluster sizes per newspaper** are captured in Figure 8c. This view depicts the overall distribution of cluster sizes across titles and shows which newspaper published the smallest (or largest) clusters. In this case, both the newspapers Le Peuple, La Sentinelle and Die Tat stand out with above average maximum cluster sizes (orange).

Lexical overlap between newspaper titles is shown in Figure 8d while Figure 8f uses a matrix view to highlight **co-occurring text reuse clusters between newspaper titles**. Both views reveal particularly high

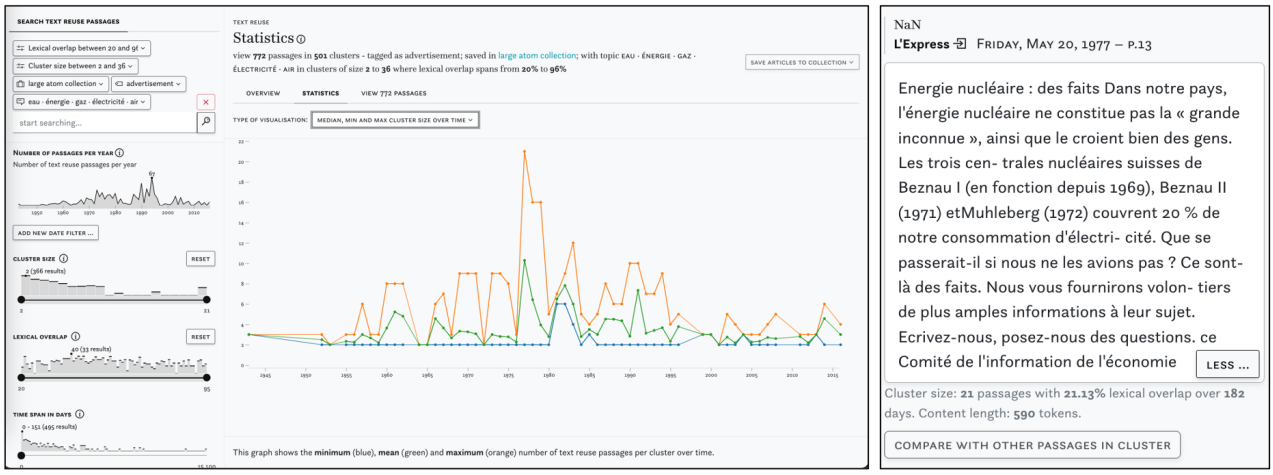

Figure 7. Example of a complex query using multiple semantic enrichments and *impresso*'s collections. Distribution of clusters on the left and passage from the largest cluster in 1977 on the left.

lexical overlaps and a large number of shared passages for example for the newspapers Journal de Geneve and Gazette de Lausanne which confirm our preceding knowledge of frequent co-publication of content.

Finally, the distributions of **lexical overlap including minimum, maximum and mean across all clusters over time** is shown in Figure 8e and offer corpus-level insights. For example, the maximum and mean lexical overlap rises from the 1970s onwards which may be a result of OCR quality improving over time. On the basis of individual titles, it also shows that Confédéré defies this trend as the mean and maximum overlap constantly decreases since the 1970s.

4.1.4 Passages Tab

The third tab titled *Passages* supports close reading of a given text reuse cluster (Task 2 - Cluster overview). The list of passages can be sorted by date, lexical overlap, cluster size, time span, and passage size. In Figure 4d we see cluster c62714 which has a large time span of 19188 days. Closer inspection reveals that it contains an article published in 1945 which described the attack on Hiroshima. The article was republished by multiple newspapers at the time and rediscovered in 1998, when it was again republished, this time by Gazette de Lausanne and Journal de Geneve.

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

5 EVALUATION

5.1 Evaluation setting

The interface was reviewed remotely by 13 evaluators: 5 (digital) historians with research experience in historical newspapers, 4 computational linguists with experience in TRD, 3 humanities scholars with experiences in text reuse and virality, 1 software developer with experience in text reuse visualisation. Five of the evaluators also participated to the workshop.

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

The evaluators worked with a form[14] which contained five evaluation tasks and gave instructive examples of their representation in the interface. These evaluation tasks were selected in light of the prototype interface's capabilities and correspond to Task 1 - Overview, Task 2 - Cluster overview, Task 3 - Compare passages, Task 5 - Types, and Task 6 - Research corpora.

Since the evaluation took place remotely, we faced the challenge to familiarise our reviewers with task definitions, their implementation in the interface and the interface components themselves as a prerequisite for their critical assessment of its capabilities. Therefore, for each evaluation task, evaluators were first presented with the task definition (based on 3.2) with the option to comment on it. Next they received instructions and usage examples which illustrated their implementation. Evaluators rated the difficulty of each task on a five-point scale (see Figure 9). A concluding segment gave the opportunity for an overall assessment of the interface. This covered its ability to effectively support the tasks presented, the quality of accompanying information, ease of navigation, and the responsiveness of the system. Finally, evaluators were asked to indicate any irritations (*"Is there anything in the application that doesn't make sense? Does anything feel out of place?"*) and recommendations for its improvement (*"Future development of the application should focus on these tasks / features / overall improvements"*).

5.2 Discussion of evaluation results

Evaluation tasks 1a-c: Obtain an overview of text reuse in a corpus, collection or query. Reviewers generally recognised the task as essential for scholars to assess the opportunities and limitations inherent in any data set.[15] They also stressed its importance as preparation for further analyses outside the interface. One evaluator noted: *"This task is absolutely critical: getting a sense whether there **is** data to interrogate helps to shape the parameters of a research questions and assess its feasibility."* The task implementation segment familiarised evaluators with different aspects of the interface, notably search, filters, the tab views and the close-up view and was split in three sub-tasks. Evaluators found the interface overall intuitive but also confirmed that familiarity with text reuse concepts such as clusters or passages is an important prerequisite. Suggested improvements for this task included the ability to compare the presence of text reuse in the entire corpus with text reuse discovered as the result of specific queries for better contextualisation of findings.

[14] `https://zenodo.org/record/8009613/`

[15] For the sake of simplicity, we merge feedback on task definitions and their implementation in the following segment.

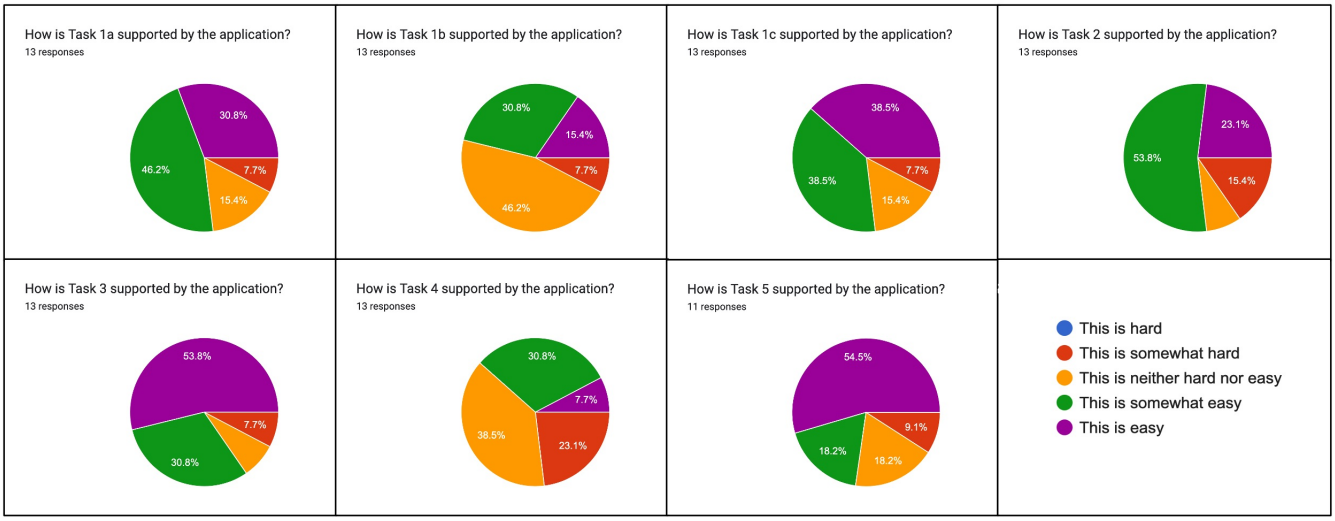

Figure 9. Rating of evaluation tasks regarding their perceived difficulty.

Feedback especially for this first task also reflected the learning experience of those evaluators who were first-time users of the *impresso* application. Critiques of individual interface components will be discussed below.

Evaluation task 2: Obtain an overview of a single cluster. Evaluators rated this task highly useful for historical research not least to assess the quality of TRD. The task was identified as part of an exploratory workflow: *"It seems a typical task again, like drilling down into a specific set of documents after first gathering a larger scale view in task 1"*. Regarding task implementation, evaluators found the interface to be *"convenient and intuitive"* and suggested high-level fingerprint views for (sets of) clusters to help with the assessment of cluster content.

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

Evaluation task 4: Identify different types of text reuse.[16] The addition of deeper semantic levels to the exploration of text reuse data was overall welcomed. Task 4 was deemed of particular interest to scholars - "*without doubt one of the most interesting aspects of the app*". Evaluators highlighted the as yet not satisfyingly closed gap between filtering operations and empirically observable types of text reuse: *"It would be helpful to have some introduction to 1) a taxonomy of reuse types, and 2) the different kinds of phenomena and how each maps to various (meta)data variables."* Feedback regarding task implementation was mixed and a majority of evaluators perceived the task as either "hard" or "somewhat hard" (Figure 9). Several evaluators suggested to create dedicated filters for empirically observed types of text reuse. This includes e.g. reuse of older content by a newspaper title and explicit support to filter for cyclical reuse. Others, and this may echo the previous feedback, felt overwhelmed by the options to filter and visualise, not knowing where to begin with. Still others were content: *"Takes some getting used to the filters and functionalities, but nothing problematic."*

Evaluation task 5: Generate research corpora based on text reuse clusters.[17] This task received comparably little feedback since not all evaluators registered an *impresso* account in time to be able to test

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

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

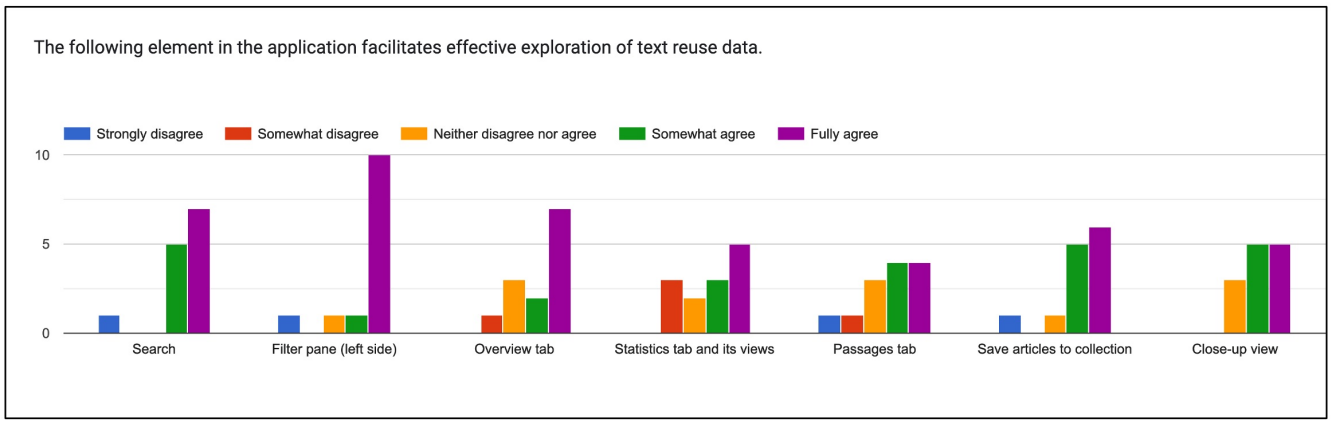

Figure 10. Rating of different components and views in the interface.

it. One evaluator called it useful for historians *"though there's a bit of a conceptual gap between reused passages and reused articles."* Implementation feedback was overall positive, *"everything looks easy on this task"*, critiques addressed the slow speed of collection processing and difficulty to find the data export function.

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

Search. With one exception, all evaluators either "somewhat" or "fully agree" that the Search component facilitates effective exploration of text reuse data (Figure 10). We note, however, that preceding experience with the *impresso* application provided an advantage and that some evaluators new to it at times struggled, this includes, e.g. search for entities or the logic of removing filters.

Filter pane. Feedback on the filter pane was even more positive but evaluators identified opportunities for improvement. This included adding units to histogram mouse-overs, better indication that they are interactive and pointers to a bug which prevented the display of newspaper titles as filter options.

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

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

Passages tab and passage comparison. This segment split evaluators. Some found it *"again easy and intuitive"* and *"Very user friendly, no remarks."* Others missed a grouping of passages by cluster and struggled to find and operate the comparative view. Regarding the contrastive view, some struggled to cycle through different passages and suggested to change the colour scheme and to eliminate some of the mismatches such as white space or OCR mistakes for easier viewing.

Close-up view. The close-up view was again rated positively, the only critical remarks pointed to the difficulty to find it without direct instructions and a bug which prevented the display of passage previews.

In the overall rating of the interface (Figure 11), the vast majority of the 13 evaluators either *"somewhat"* or *"fully"* agreed that the interface supports the evaluation tasks (12), that the usage examples to explain the interface (11), that it was easy to navigate (9) and that loading times were acceptable (11). The interface clearly has a learning curve, which was described by one evaluator: *"The functionality of the filters available here is impressive and of reasonable simplicity. I wouldn't describe it as 'easy', mostly because there's a lot going on and a researcher not familiar with the dynamics of text reuse might be a bit lost, but I'm not sure I would trade the current depth of filtering for easier use."*

The replies to our questions regarding irritations and future improvements confirm the critiques of the statistics tab and passages tab we discuss above. At this stage of development, six evaluators found them *"either difficult to read or [they] did not provide useful insights."* In addition, recommendations for future

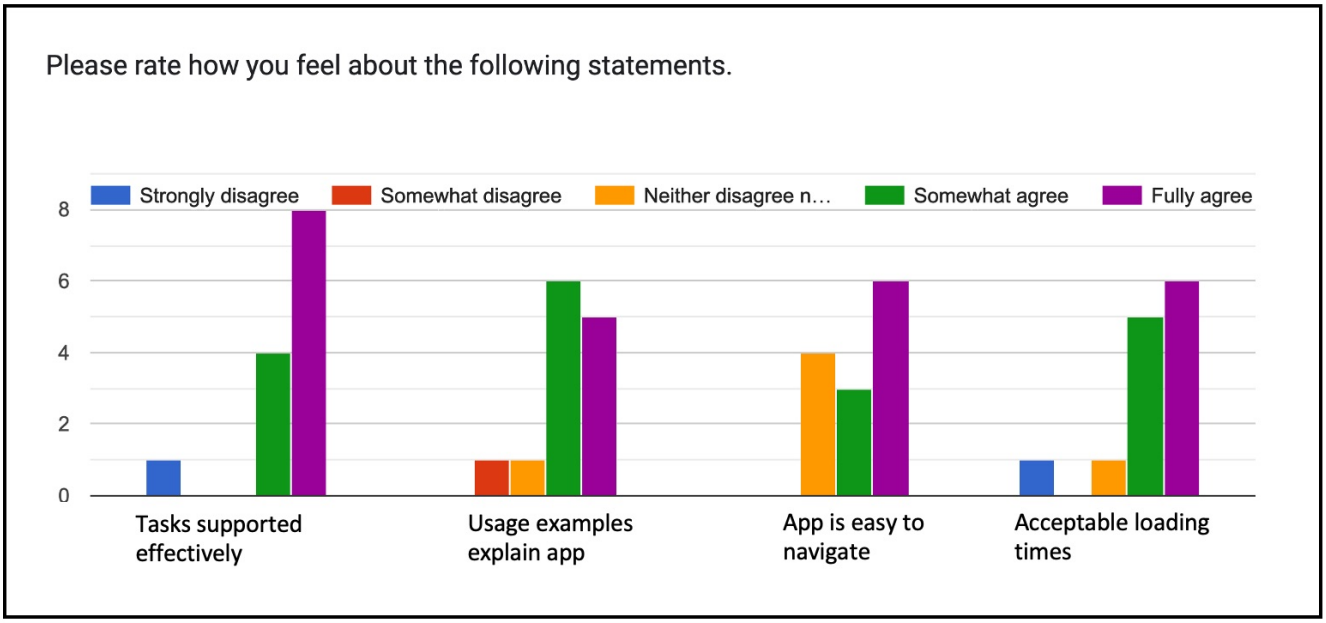

Figure 11. Overall rating of the interface.

development addressed the already foreseen integration of *impresso*'s Inspect & Compare component for side-by-side comparisons of article sets, higher speed for the creation of collections, API access to the data, and new filters based on a yet to be created taxonomy of text reuse types.

6 CONCLUSION AND FUTURE WORK

In this paper we have presented the prototype of the Text Reuse at Scale interface, the to our knowledge first interface which integrates text reuse data with other forms of semantic enrichment to enable a versatile and scalable exploration of intertextual relations in historical newspaper corpora. The interface was developed as part of the *impresso* project and combines powerful search and filter operations with close and distant reading perspectives. We reported on high-level research objectives as well as common user tasks for the analysis of historical text reuse data and presented the prototype interface together with the results of a user evaluation.

We use examples to illustrate how the integration of text reuse data with semantic enrichments (content type, language, topics, named entities) has proven advantageous. First, enrichments serve as a means to effectively filter for relevant sets of text reuse data, second to identify different types of text reuse and third to gain overviews of the content of text reuse data. We have also demonstrated the interface's ability to retrieve text reuse following temporal patterns such as rapidly spreading content of different types as well as content rediscovery after long time periods. Examples include the coverage of the attack on Hiroshima, the event anniversary reprints of the same article, and the reprint of the 1945 article in 1998. Further, the interface reveals systematic co-publication independently of content as in the example of the Journal de Geneve and Gazette de Lausanne. We have also shown its ability to give insight into the content captured by one or more text reuse cluster(s) using topics and to shift between distant and close reading operations. Finally, we have shown its usage for a critical assessment of corpora and variations in the performance of TRD based on the distributions of passages, cluster sizes and lexical overlap over time.

At this stage of its development, the Text Reuse at Scale interface supports many but not all of the previously discussed tasks for the exploration of historical text reuse data (see Table 2). Future development of the prototype will address the integration with the *impresso* Inspect & Compare component to enable side-by-side comparisons of article sets which contain text reuse passages in support of Task 4 - Compare clusters and Task 7 - Connections and better support for temporal dimensions of text reuse data 1. Furthermore we will take into account the need to improve the legibility and documentation of the statistics tab and work

to resolve the observed difficulties in the passages tab together with smaller bugs discovered during the evaluation.

CONFLICT OF INTEREST STATEMENT

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

AUTHOR CONTRIBUTIONS

DG was responsible for UI and UX development and integration into the *impresso* interface. MD, MR, and DG contributed to conception and design of the user workshop. ME and MR revised the *impresso* technical infrastructure to suit the needs of the interface. MD, MR and DG developed the evaluation procedure. MD and MR wrote the first draft of the manuscript. PA, KB and BD wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

The workshop was funded by the Luxembourg Centre for Contemporary and Digital History (C^2DH). This work is building on the research project "*impresso*. Media Monitoring of the Past" funded by the Swiss National Science Foundation (SNSF) under grant ID CR- SII5_173719.

ACKNOWLEDGMENTS

We wish to express our gratitude to fred Pailler, Matteo Romanello and Jana Keck for their contributions to the workshop and to the Luxembourg Centre for Contemporary and Digital History (C^2DH) for agreeing to host it.

SUPPLEMENTAL DATA

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

DATA AVAILABILITY STATEMENT

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

REFERENCES

Büchler, M., Burns, P. R., Müller, M., Franzini, E., and Franzini, G. (2014). Towards a Historical Text Re-use Detection. In *Text Mining*, eds. C. Biemann and A. Mehler (Springer International Publishing), Theory and Applications of Natural Language Processing. 221–238

Cordell, R. (2015). Reprinting, Circulation, and the Network Author in Antebellum Newspapers. *American Literary History* 27, 417–445. doi:10.1093/alh/ajv028

Keck, J., Oiva, M., and Fyfe, P. (2022). Lajos Kossuth and the Transnational News: A Computational and Multilingual Approach to Digitized Newspaper Collections. *Media History* 0, 1–18. doi:10.1080/13688804.2022.2146905

Liebl, B. and Burghardt, M. (2020). "Shakespeare in the Vectorian Age" – An evaluation of different word embeddings and NLP parameters for the detection of Shakespeare quotes. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (Online: International Committee on Computational Linguistics), 58–68

Manjavacas, E., Long, B., and Kestemont, M. (2019). On the Feasibility of Automated Detection of Allusive Text Reuse. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics*

for Cultural Heritage, Social Sciences, Humanities and Literature (Minneapolis, USA: Association for Computational Linguistics), 104–114. doi:10.18653/v1/W19-2514

Moritz, M. and Steding, D. (2018). Lexical and Semantic Features for Cross-lingual Text Reuse Classification: An Experiment in English and Latin Paraphrases. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (Miyazaki, Japan: European Language Resources Association (ELRA)), 1976–1980

Oiva, M., Nivala, A., Salmi, H., Latva, O., Jalava, M., Keck, J., et al. (2020). Spreading News in 1904. *Media History* 26, 391–407. doi:10.1080/13688804.2019.1652090

Paasikivi, S., Salmi, H., Vesanto, A., and Ginter, F. (2022). Infectious media: Cholera and the circulation of texts in the finnish press, 1860–1920. *Media History* 0, 1–22. doi:10.1080/13688804.2022.2054408. Publisher: Routledge _eprint: https://doi.org/10.1080/13688804.2022.2054408

Paju, P., Rantala, H., and Salmi, H. (2023). Towards an ontology and epistemology of text reuse. In *Reflections on tools, methods and epistemology*, eds. E. Bunout, M. Ehrmann, and F. Clavert (De Gruyter Oldenbourg). 253–274. doi:doi:10.1515/9783110729214-012

Paju, P., Salmi, H., Rantala, H., Lundell, P., Marjanen, and Vesanto, A. (2022). Textual migration across the baltic sea : Creating a database of text reuse between finland and sweden. In *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)*, eds. K. Berglund, M. La Mela, and I. Zwart (The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)), CEUR Workshop Proceedings. 361–369. Publisher: CEUR-WS.org

Romanello, M., Berra, A., and Trachsel, A. (2014). Rethinking Text Reuse as Digital Classicists. In *9th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2014, Lausanne, Switzerland, 8-12 July 2014, Conference Abstracts* (Alliance of Digital Humanities Organizations (ADHO))

Romanello, M. and Hengchen, S. (2020). Detecting Text Reuse with Passim. *The Programming Historian* , /doi:10.46430/phen0092

[Dataset] Romanello, M. and Snyder, R. (2017). Cited Loci of the Aeneid : Searching through JSTOR's content the classicists' way. (Blog post)

[Dataset] Rosson, D., Mäkelä, E., Vaara, V., Mahadevan, A., Ryan, Y., and Tolonen, M. (2023). Reception Reader: Exploring Text Reuse in Early Modern British Publications. (Pre-print)

Salmi, H., Paju, P., Rantala, H., Nivala, A., Vesanto, A., and Ginter, F. (2020). The reuse of texts in Finnish newspapers and journals, 1771–1920: A digital humanities perspective. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 0, 1–15. doi:10.1080/01615440.2020.1803166

Salmi, H., Rantala, H., Vesanto, A., and Ginter, F. (2019). The Long-Term Reuse of Text in the Finnish Press, 1771-1920. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, eds. C. Navarretta, M. Agirrezabal, and B. Maegaard (Copenhagen, Denmark: CEUR Workshop Proceedings), vol. 2364, 253–273

Scheirer, W., Forstall, C., and Coffee, N. (2016). The sense of a connection: Automatic tracing of intertextuality by meaning. *Digital Scholarship in the Humanities* 31, 204–217. doi:10.1093/llc/fqu058

Smith, D. A., Cordell, R., and Dillon, E. M. (2013). Infectious texts: Modeling text reuse in nineteenth-century newspapers. In *2013 IEEE International Conference on Big Data*. 86–94. doi:10.1109/BigData.2013.6691675

Smith, D. A., Cordell, R., and Mullen, A. (2015). Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers. *American Literary History* 27, E1–E15. doi:10.1093/alh/ajv029

Thérenty, M.-E. and Venayre, S. (2021). *Le monde à la une. Une histoire de la presse par ses rubriques* (Anamosa), illustrated édition edn.

Verheul, J., Salmi, H., Riedl, M., Nivala, A., Viola, L., Keck, J., et al. (2022). Using word vector models to trace conceptual change over time and space in historical newspapers, 1840–1914. *Digital Humanities Quarterly* 016

Vesanto, A., Nivala, A., Rantala, H., Salakoski, T., Salmi, H., and Ginter, F. (2017). Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771–1910. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*. 54–58

Walma, L. W. B. (2015). Filtering the "news" : Uncovering morphine's multiple meanings on delpher's dutch newspapers and the need to distinguish more article types. *Tijdschrift voor Tijdschriftstudies*

Yousef, T. and Janicke, S. (2021). A Survey of Text Alignment Visualization. *IEEE Transactions on Visualization and Computer Graphics* 27, 1149–1159. doi:10.1109/TVCG.2020.3028975