

Article

Impresso Inspect and Compare. Visual Comparison of Semantically Enriched Historical Newspaper Articles

Marten Düring ^{1,*}, Roman Kalyakin ^{2,*}, Estelle Bunout ^{1,3} and Daniele Guido ^{2,t}

¹ Luxembourg Centre for Contemporary and Digital History (C²DH), Digital History and Historiography, 2, Avenue de l'Université, L-4365 Esch-sur-Alzette, Luxembourg; estelle.bunout@zzf-potsdam.de

² Luxembourg Centre for Contemporary and Digital History (C²DH), Digital Research Infrastructure, 2, Avenue de l'Université, L-4365 Esch-sur-Alzette, Luxembourg; daniele.guido@uni.lu

³ Leibniz-Zentrum für Zeithistorische Forschung Potsdam (ZZF), Am Neuen Markt 1, 14467 Potsdam, Germany

* Correspondence: marten.during@uni.lu (M.D.); roman@kalyakin.com (R.K.)

† These authors contributed equally to this work.

Abstract: The automated enrichment of mass-digitised document collections using techniques such as text mining is becoming increasingly popular. Enriched collections offer new opportunities for interface design to allow data-driven and visualisation-based search, exploration and interpretation. Most such interfaces integrate close and distant reading and represent semantic, spatial, social or temporal relations, but often lack contrastive views. Inspect and Compare (I&C) contributes to the current state of the art in interface design for historical newspapers with highly versatile side-by-side comparisons of query results and curated article sets based on metadata and semantic enrichments. I&C takes search queries and pre-curated article sets as inputs and allows comparisons based on the distributions of newspaper titles, publication dates and automatically generated enrichments, such as language, article types, topics and named entities. Contrastive views of such data reveal patterns, help humanities scholars to improve search strategies and to facilitate a critical assessment of the overall data quality. I&C is part of the *impresso* interface for the exploration of digitised and semantically enriched historical newspapers.

Keywords: historical newspapers; semantic enrichment; user interface; metadata; comparison; small multiples; search; data criticism; diverging bar charts; digital humanities; digital history



Citation: Düring, M.; Kalyakin, R.; Bunout, E.; Guido, D. *Impresso Inspect and Compare. Visual Comparison of Semantically Enriched Historical Newspaper Articles*. *Information* **2021**, *12*, 348. <https://doi.org/10.3390/info12090348>

Academic Editor: Willy Susilo

Received: 15 July 2021

Accepted: 10 August 2021

Published: 27 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Text mining produces an overwhelming amount of data which promises new perspectives on large-scale document collections. Such data help researchers to identify patterns but also call for visual support in the analysis [1]. Digitised historical newspapers, in particular, lend themselves to text mining since they are available in large quantities and offer rich information for diverse research interests [2,3]. For example, recent data-driven analyses traced content flows across titles [4–6] and detected semantic shifts for a data-driven, bottom-up approach to the history of ideas [7,8]. The majority of users in the humanities and in history in particular, however, focus on search and exploratory content retrieval tasks [9–11], where exploration is understood as “comprehending the information space and expressing evolving search intents for iterative exploration and retrieval of information” [12], p. 1. While the idea of comparison—“one of the most basic scholarly operations” [13]—is at least implied in the majority of all exploratory interfaces, dedicated platforms for the side-by-side comparison of text document sets are rare.

In this paper, we introduce Inspect and Compare (I&C), which is part of a novel interface for the exploration of semantically enriched historical newspapers. Its main contribution to the current state of the art in interface design for such document collections is its ability to reveal similarities, differences and overlaps across user-curated sets of historical

newspaper articles based on metadata and semantic enrichments (<https://impresso-project.ch/app/compare> (accessed on 10 August 2021)). The component was developed for historians with very limited expertise in data science/visualisation, based on their needs expressed during user workshops to identify relevant content from the corpus and to gather different perspectives on the materials they were working on. This translated into a versatile, yet simple interface, which offers multiple search and filter functions to allow users to compile and compare article sets for a wide range of research interests and help them generate a multitude of contrasting views on the data they study. In this paper, we exemplify these interactions with three usage scenarios: query improvement, content exploration and data criticism. We prioritise these needs over the integration of more advanced data science techniques. The latter would yield insights into specific research topics, such as rumour spreading or misinformation [14] at the expense of overall versatility and would result in a significantly steeper learning curve for our target users.

I&C is part of the *impresso* web interface for the exploration of digitised and semantically enriched historical newspapers (<https://impresso-project.ch/app/> (accessed on 10 August 2021)). The interface integrates text-mining-based semantic enrichments, design and data visualisations in generic workflows for content discovery and (digital) source criticism [15]. It was developed by the interdisciplinary research project *impresso*. Media Monitoring of the Past (<https://impresso-project.ch> (accessed on 10 August 2021)) and brought together computational linguists, designers, software developers and historians to open up digitised collections of multilingual historical newspapers for data-driven content search, discovery and data criticism. In partnership with the national libraries of Switzerland and Luxembourg, the Swiss Economic Archives, the State Archives of Valais and newspapers *Neue Zürcher Zeitung* and *Le Temps*, the project created a multilingual (fr, de, lu) corpus of Swiss and Luxembourgish newspapers and developed a technical architecture that reflects the idiosyncrasies of digitised historical newspapers to facilitate data storage, enrichment and access, which is described in [16]. Table 1 offers an overview of the corpus and its enrichment; the interface offers more detailed information on its composition (<https://impresso-project.ch/app/newspapers/> (accessed on 10 August 2021)). The main contribution that the *impresso* interface makes to the current status quo in interface design for historical newspapers is an iterative query-building workflow across multiple inter-linked components that is driven by semantic enrichments generated during the project. This workflow encourages users to weave together insights gathered through advanced search, the exploration of ngram frequencies, named entity and topic distributions, text reuse and image similarity detection as well as article recommendations.

Table 1. Overview of the size of the *impresso* corpus and its enrichment. These numbers reflect the status quo at the time of writing and prior to a planned corpus expansion.

| Item | Count |
|-----------------------------------|----------------|
| Newspaper titles | 76 |
| Newspaper issues | 600,919 |
| Pages | 5,429,656 |
| Content (articles, adverts, etc.) | 47,798,468 |
| Words | 12,493,358,703 |
| Linked named entities | 530,086 |
| Topics per language (fr, de, lu) | 100 |

Related Work

Interfaces for historical newspapers that also target non-academic audiences use data visualisations mostly to represent article and keyword distributions over time [17,18]. More analytic data visualisations for expert audiences in media history and libraries focus on the distribution of metadata within corpora (e.g., number of images, text length, missing data) [19], often also in combination with spatial data [20,21].

Within the visual analytics research community, there is an overwhelmingly large literature on the usage of interactive visualisations for the exploration of textual data which justifies the creation of a survey of surveys on the topic [1]. For an overview of visualizations of text-based data, see the Text Visualization Browser (<https://textvis.lnu.se/> (accessed on 10 August 2021)). While exploratory comparisons are among the most common tasks for such systems, to our knowledge, I&C is the first component of its kind that allows side-by-side comparisons of semantically enriched sets of text documents. Its development was inspired by product comparison websites, which typically represent different products in columns and group their exact specifications by categories in rows below. For an example see <https://www.pricerunner.com/> (accessed on 10 August 2021)). This principle is, of course, not new and dates back to the early 1990s, when “interactive tables” first allowed rapid overviews of product specifications, making it easier to compare them and to identify outliers [22,23]. Such side-by-side ranked lists are also used to compare, for example, search engine results [24] and, in combination with Venn diagrams, to compare gene lists [25]. Regarding the data-driven comparison of document sets, more distantly related works focus on statistical corpus composition in linguistics [26], the (visual) exploration of probabilistic document classifications, such as topic modelling [27,28] or content overlap, e.g., in text reuse [29].

Our usage scenario on search query improvement shares the same underlying motivation as work on query expansion, which comprises of techniques that “reformulat[e] the user’s original query to enhance the information retrieval effectiveness” [30], p. 3. Earlier work in the field dating back as far as the mid 1990s already used data visualisation to indicate the significance of search terms and experimented with enhanced bar charts to visualise keyword relevance per retrieved document [31–33], included histograms and tag clouds to illustrate keyword relevance [34], made use of circular layouts to allow for interactive query comparison [35] and made use of interactive 2D scatter plots [36].

The focus of I&C instead lies on article sets and less on search terms. We compare the features of the retrieved documents and emphasise overlaps and dissimilarities between different types of article sets which can be either query results or user-curated article sets, called “Collections” within the *impresso* interface. Queries in I&C can also be significantly more complex compared to those analysed above and may include multiple filters and Boolean expressions.

Data quality problems, understood as instances when data do not have sufficient fitness for use [37], can severely affect the outcomes of analyses. Despite efforts to design strategies for improved data quality, to detect flaws and to compensate for them, they continue to be a concern for researchers [38]. While much work in the field focuses on data cleansing [39], others have demonstrated that interactive visualisations can play an important role to flag different types of anomalies [40,41]. I&C offers far less sophisticated analytical tools and was not designed for data cleansing operations but, as we seek to demonstrate below, offers an accessible interface for its target audience which can nevertheless reveal important insights into overall data quality and its impact on content retrieval and exploration.

2. Materials and Methods

2.1. Interface Design

The side-by-side comparisons of article sets enables both a content-based comparison and an awareness of the spatial, temporal, linguistic distributions of the compared queries and collections. I&C is split into two views: “Inspect” (Figure 1) displays the relative distribution of metadata and semantic enrichments using small multiples of bar charts and a three-column layout for the two sets to compare (A) on the left and (B) on the right, as well as the intersection with articles that are part of both sets in the centre. The bar length is proportionate to the highest value per object group and column. By default, (A) displays the last query that users were working on to encourage them to compare it with either a new query or a precompiled collection of articles in (B). Generally, any

combination of queries and collections can be compared against each other. Queries and collections can be further refined using search and filter operations based on topics, named entities, newspaper titles, countries of origin, dates, languages and user-curated collections (Figure 2).

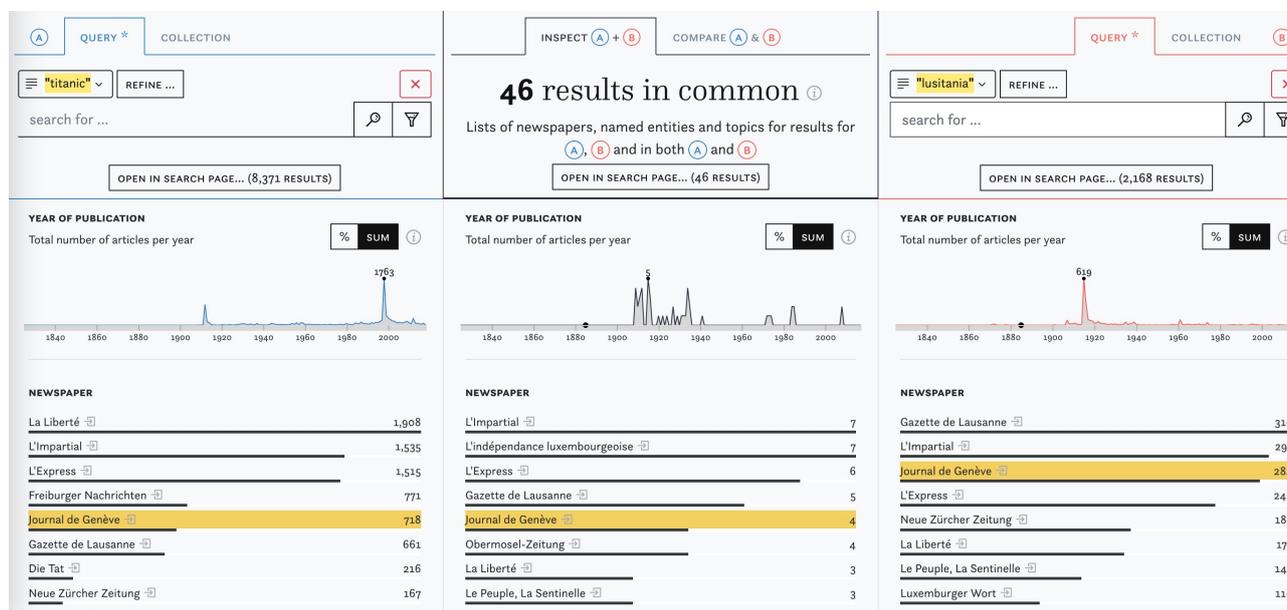


Figure 1. Overview of the I&C interface with an example query for “titanic” and “lusitania”.

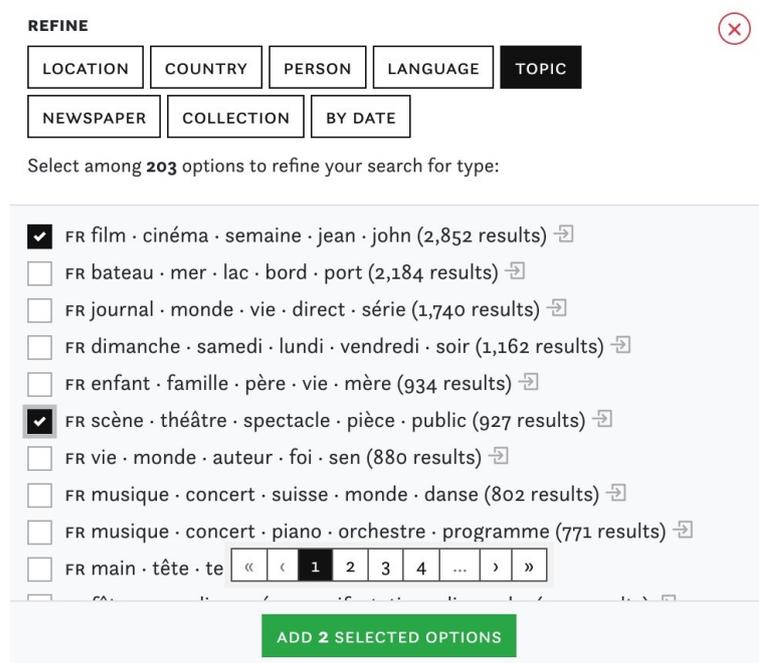


Figure 2. “Refine” view to add additional filters to queries and collections.

The results in all three columns are ranked by frequency and yellow highlights reveal the position of corresponding hits across all columns. Like the rest of the *impresso* interface, I&C allows users to shift between distant and close reading views. Article sets in each column can be inspected in detail and additional information about each result item—e.g., biographical details about a person entity and its distribution across the corpus—can be obtained with a single click (Figure 3). Where the “Inspect” function reveals similarities, “Compare” (Figure 4) uses diverging bar charts to highlight the discrepancies between two

sets. To increase legibility for highly divergent values, users can choose square root scale over the default linear scale. Bars can be sorted by largest overlap in percent or by absolute number of articles. The two views complement each other: “Inspect” offers overviews of the distribution of enrichments and metadata in (A), (B) and their intersection, while “Compare” emphasises the relative proportions of value distributions.

Usage scenarios best illustrate I&C’s abilities; in the next section, we present three distinct yet related usages of the component: (1) improving search query composition, (2) content exploration, i.e., to observe (dis)similarities and overlaps in the distribution of metadata and semantic enrichments, and (3) data criticism, i.e., to assess the composition of the corpus and the quality of the semantic enrichments.

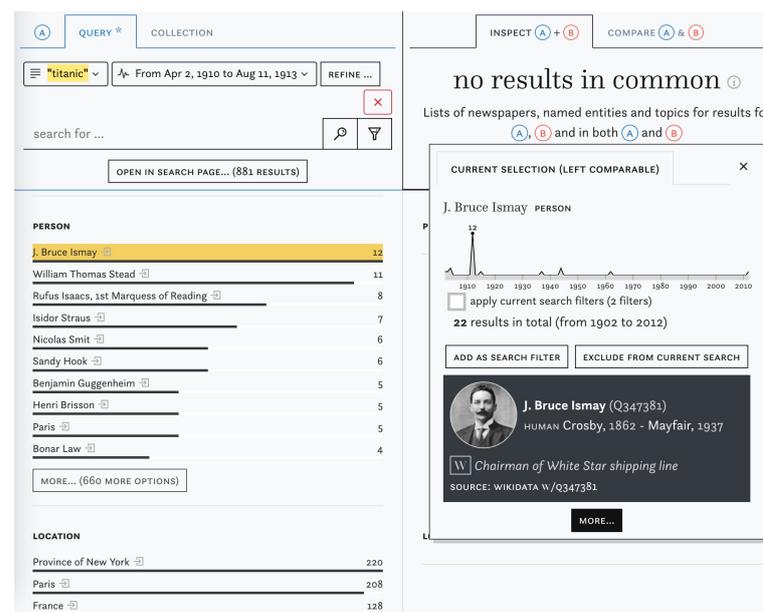


Figure 3. Contextual information about a linked named entity.

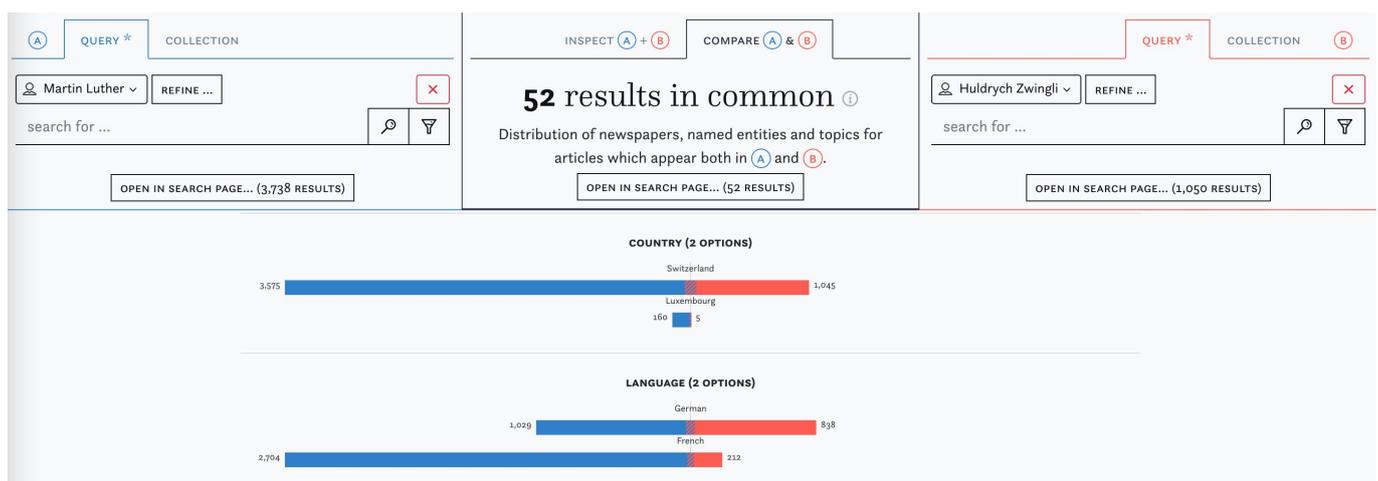


Figure 4. Comparison of the linked entities “Martin Luther” and “Huldrych Zwingli”.

2.2. Computational Complexity

I&C retrieves its data from the *impresso* system, which facilitates data storage, enrichment and indexing. Software that offers data query interfaces usually keeps its data at different levels of normalisation, which often depends on the size of the data and the type of queries performed. Small data sets can be kept highly normalised in relational SQL databases that offer great flexibility in querying data. As the data size grows,

like in *impresso*, indexes need to be added to the database. If queries are known in advance, they can be built almost entirely around indices and normalisation of data becomes largely redundant. This increases memory requirements but decreases CPU use. Therefore, for *impresso*, we opted for denormalised data structure stored in a Solr cluster (<https://solr.apache.org/features.html> (accessed on 10 August 2021)). All article metadata that can be relevant for researchers is stored alongside the articles and is indexed, allowing us to use the faceting capability of Solr for executing predefined queries. This method allows us to avoid heavy use of CPU and retrieve facets from indices which most of the time are loaded into RAM or can be retrieved from the cache on disk. Table 2 gives an overview of execution times for the queries we discuss throughout this paper.

Table 2. Execution times in seconds for the queries displayed in Figures 1 and 5–8 averaged over three runs.

| Query | Duration | Std. Dev. |
|----------|----------|-----------|
| Figure 1 | 0.92 s | 0.20 s |
| Figure 5 | 1.23 s | 0.19 s |
| Figure 6 | 0.72 s | 0.13 s |
| Figure 4 | 0.81 s | 0.37 s |
| Figure 7 | 0.75 s | 0.18 s |
| Figure 8 | 0.69 s | 0.01 s |

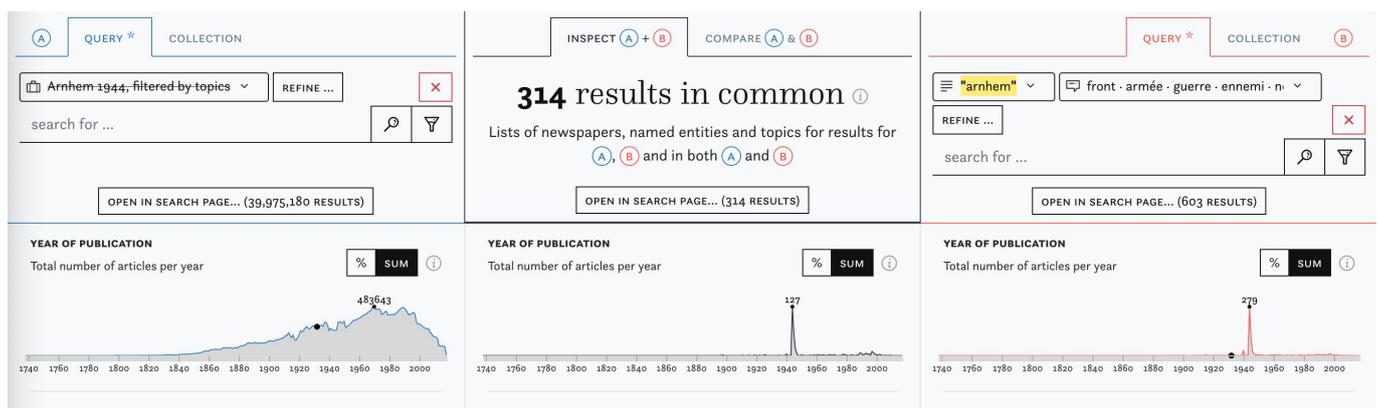


Figure 5. Identifying previously unknown articles by subtracting one set of articles from another.

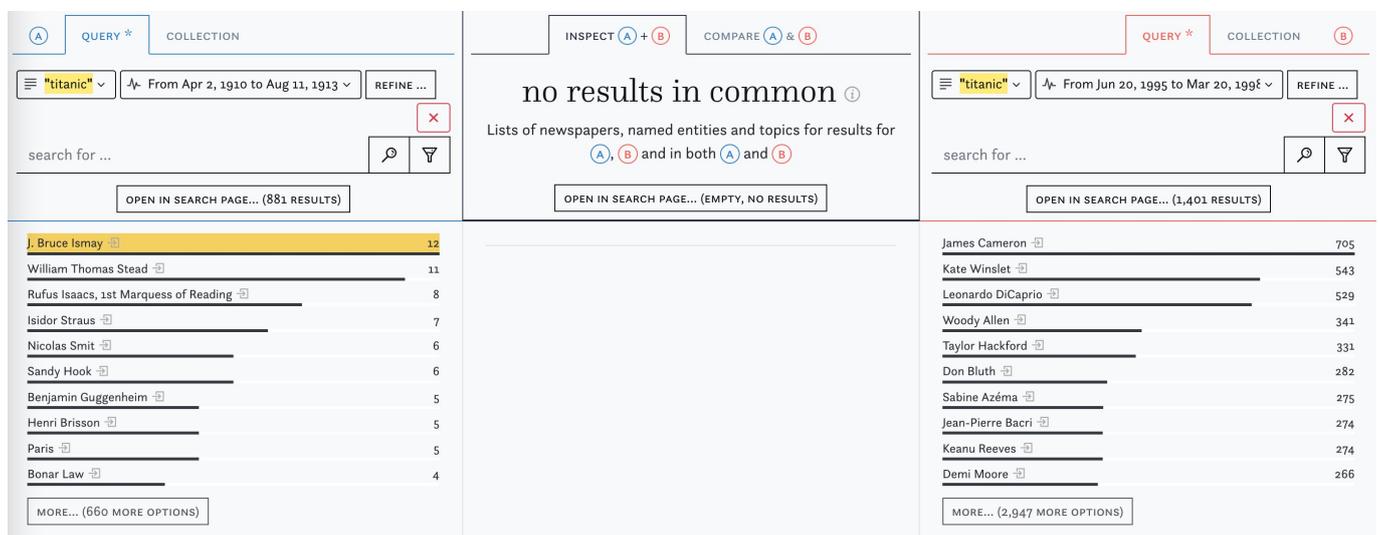


Figure 6. Comparing time periods—“titanic” around 1912 and around 1998.

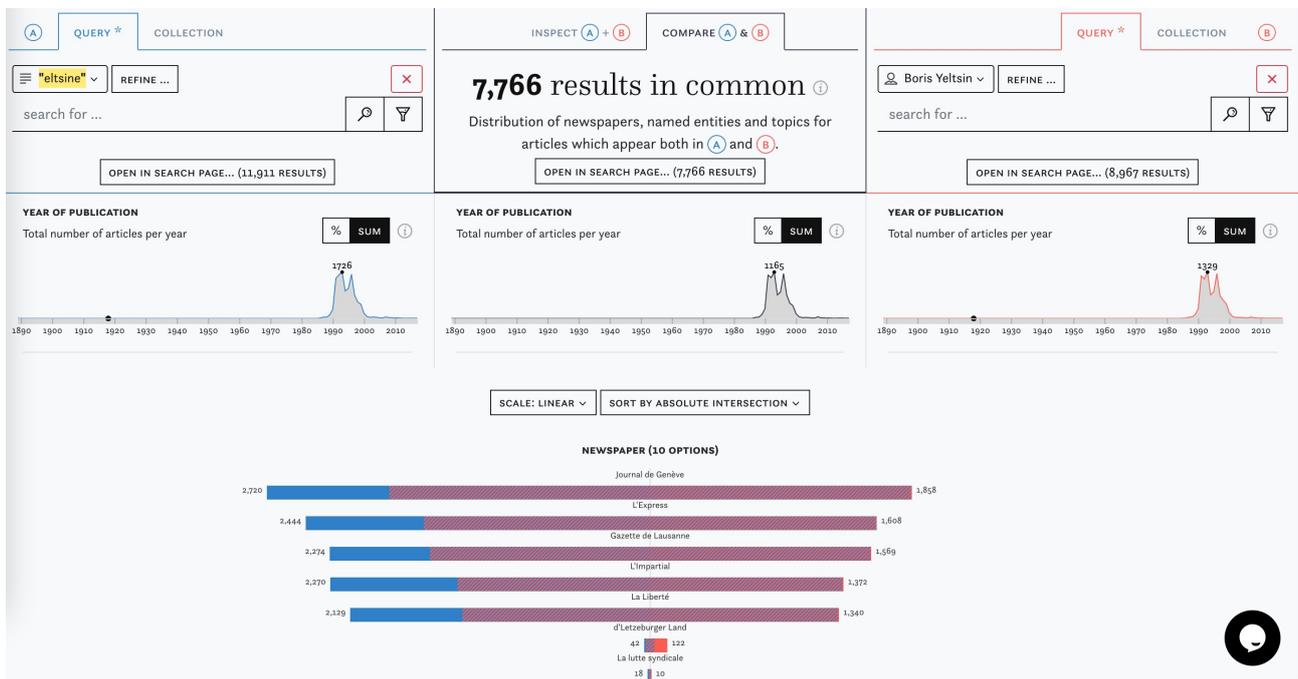


Figure 7. Screenshot from “Compare” which contrasts the string “eltsine” and the linked entity “Boris Yeltsin”.

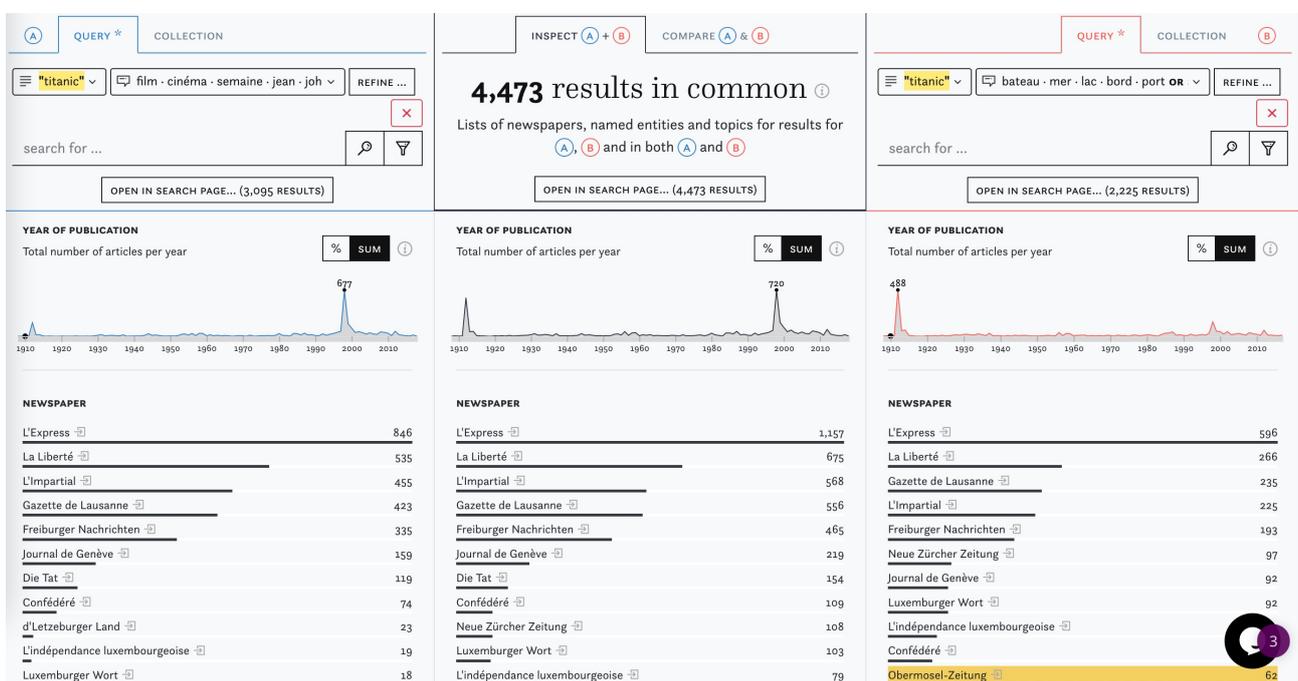


Figure 8. Screenshot from “Inspect” displaying keyword searches for “titanic” filtered by film-related topics (left) and by seafaring-related topics (right).

The back end part of our software is written using Node.js (<https://nodejs.org/> (accessed on 10 August 2021)). It adds little overhead to the Solr engine execution times and, in some cases, reduces them by executing queries in parallel. We use D3.js (<https://d3js.org/> (accessed on 10 August 2021)) for data visualisation on the front end. Data retrieved from the back end are already in the suitable format and do not need to be transformed again, which also reduces the overhead.

3. Results

3.1. Improving Search Queries

Advanced search for content retrieval is considered to be the preferred way for historians to engage with digitised newspapers [9–11] and was one of the focus areas for the design of the *impresso* app. Its development was accompanied by a series of co-design workshops, first to help identify user needs and later to validate previous development (<https://impresso-project.ch/activities/timeline/> (accessed on 10 August 2021)). In one of these workshops, historians described their experiences with related interfaces and stated that small tweaks in the formulation of search queries often yielded significantly different results. Continued experimentation with such tweaks helped them maximise the number of articles they retrieved. However, the process was described as cumbersome and lacking in transparency since it was often unclear exactly where the differences and gains between two queries were.

I&C allows a more systematic approach to the optimisation and comparison of search queries. See Figure 5 for an example: in (A), we exclude a previously created user collection with articles on the Battle of Arnhem (1944). This means that (A) now contains a pool of all 40 million articles, except for those we already know to be relevant. In (B), we run an experimental new query which uses the keyword “arnhem” in combination with a war-related topic, which yields 603 articles. In this case, the intersection column shows that 314 of these articles were also found in (A) and, consequently, are not yet part of the collection. As close reading confirms, many of them are indeed relevant hits and should be added to the collection. With this setup, users can try out many different queries in rapid succession, observe differences between them, thereby iteratively improving queries or building collections of articles.

3.2. Content Exploration

Comparisons can offer valuable contextual information which help with the interpretation of patterns observed in the data. In Figure 6, we illustrate how I&C reveals change over time by applying a date filter on top of the query for “titanic” around the years 1912 (A) and 1998 (B). In the latter column, mentions of the Titanic movie cast and production team and other players in the film industry completely overshadow those of the people involved in the historical event (note that we did not filter out any coverage of the historical event here).

For the next example, we compare the coverage of German reformer Martin Luther and his Swiss counterpart Huldrych Zwingli in our Swiss–Luxembourgish corpus. Figure 4 shows a comparison based on automatically detected linked person entities and reveals that Luther was detected significantly more often, but that Zwingli was identified disproportionately less often in the Luxembourgish press and in French-language newspapers in general. Also of note is the unexpectedly small overlap of 52 articles in which both entities were detected.

3.3. Data Criticism

Historical source criticism emphasises the importance of the circumstances under which historical artefacts of any kind were created and preserved. Knowledge of these circumstances is key to their interpretation. The same principles apply for the analysis of digitised sources and the use of tools and interfaces, forcing historians to adjust their methodological toolkits and to reflect on the consequences of the digital transformation of their sources [15,42] and how they interact with them [43,44]. The *impresso* interface sought to reflect these debates and to offer transparency, especially regarding the composition and enrichment of corpora and the techniques for exploring them (<https://impresso-project.ch/theapp/usage/> (accessed on 10 August 2021)). As with any similar system, output quality inevitably varies depending on the quality of the input: newspapers were digitised at different points in time and especially the quality of OCR and layout recognition can vary significantly, also impacting the quality of any subsequent semantic enrichment [45].

Within I&C and the *impresso* interface in general, data quality problems are therefore considered a given and inevitable. The data quality assessment workflow we exemplify here is a crucial step in obtaining an understanding of the opportunities that such flawed data offer and to identify their limits. In other words, interfaces such as *impresso* offer answers but, much like Greek oracles, it is not always clear to which questions, meaning that they may lead astray those who do not exercise cautious interpretation and self-reflexivity.

To exemplify how I&C can help assess the varying quality of enrichments, we have chosen the examples of named entity linking and topic modelling, beginning with linked named entities and the example of former Russian President Boris Yeltsin. In Figure 7, we compare a string search for “eltsine” (French spelling) with a search for the linked entity “Boris Yeltsin”. Blue segments on the bars in (A) indicate the proportion of articles for different newspaper titles which would be missed by a search for the linked entity alone. While combinations of keywords, such as “eltsine” + “jelzin” (German spelling), could increase recall, they are more likely to reduce precision and retrieve false positives (in this case, these could for example include references to a vodka brand). On the other hand, entity linking will also capture synonyms associated with the linked entity, but at the expense of a significant bias toward social elites, which is caused by repositories used to coreference entities. There is no obvious choice to make in this instance; the task at hand will determine which search strategy should be chosen or whether a combined approach will yield better results.

A variation of the previous example shows how I&C can help shed light on the impact of topics as filters. In Figure 8, (A) shows a keyword search on “titanic” filtered by film-related topics. In (B) we also run a query for “titanic” but filter for seafaring-related topics. The frequencies over time reveal the consequences: (A) lists, for the most part, articles around the time that the 1998 film came out, whereas (B) shows a clear spike around the time the ship sank in 1912. A closer inspection of the noticeable spike in (A) around the year 1912 reveals that film-related topics were assigned to 199 articles, even though most of them are not directly related to films. Topics are generally a salient means to identify semantic themes in a corpus but they remain fuzzy because of their probabilistic nature. This experiment gives users an idea of how this fuzziness manifests itself.

4. Discussion

In spring 2020, we asked 10 domain experts (8 historians and 2 librarians) to review the *impresso* interface to collect feedback that could guide the development of the final version. All reviewers had previously worked with different interfaces for digitised historical newspapers, and two reviewers were themselves involved in projects similar to *impresso* and had a more specific interest in its development. To ensure sufficient time for familiarisation and realistic conditions for testing, all reviewers were asked to test the app in their own time, using a topic they were familiar with, and to document their experiences while doing so. They received a form which listed the main interface components and offered brief explanations of their functionality together with illustrations of their intended usage, links to example queries and links to all available educational materials. For each component, the reviewers were encouraged to answer questions on perceived relevance, validity of results obtained and possible improvements, and to mention the research topic they chose to test the interface on.

Positive feedback on I&C pointed to the opportunity to obtain a high-level yet precise overview of search results, overlaps and contrastive views, which served to provide meaningful insights and guide further exploration of the corpus, not least because of the close link between distant and close reading views. One reviewer considered I&C an enrichment of their historical research workflow, while others named it their favourite component in the interface. Negative feedback included criticism of the complexity of the I&C interface and the high density of information, which prompted us to offer additional documentation and learning resources (<https://impresso-project.ch/theapp/usage/> (accessed on 10 August 2021)). Suggestions included a more granular approach to date filtering (which

was since implemented), the optional removal of the centre column, and the possibility of comparing more than two queries/collections, which we did not implement because it was likely to lead to a significant increase in complexity.

5. Conclusions

In this paper, we presented Inspect and Compare, a component which facilitates the comparison of newspaper article sets within the *impresso* interface for semantically enriched historical newspapers. I&C targets humanities scholars, especially historians, and was designed to facilitate a seamless interaction between close and distant reading whilst also helping researchers to identify overlaps and (dis)similarities across queries and collections. We demonstrated the capabilities of the simple yet versatile interface with three scenarios which covered search query improvement, content exploration and data quality assessment. The feedback we obtained during a preliminary review and during more informal one-to-one demo sessions suggests that I&C serves researchers by helping them to identify overlaps and (dis)similarities, as expected. The opportunities related to query and search strategy improvement as well as data criticism are less obvious to users and need to be actively advertised. They yield important insights that are part of an iterative learning process during which users familiarise themselves with the opportunities and limitations of the semantic enrichments and the capabilities of the interface.

There are multiple possible avenues to expand a system like I&C. One of them is more interactions between the system and the user, e.g., to guide user attention towards potential data problems or statistically significant patterns as exemplified by [40]. Another direction are opportunities for more advanced data analytics. While semantically enriched historical newspapers do not offer data as equally rich and diverse as contemporary social media, skilled researchers can nevertheless take inspiration from analyses such as those discussed in [14], which may also inform future efforts in the enrichment of historical document collections.

Author Contributions: Conceptualisation, E.B., M.D., D.G., R.K.; data curation, D.G.; funding acquisition, M.D.; investigation, E.B., M.D.; methodology, E.B., M.D., D.G. and R.K.; project administration, M.D.; software, D.G., R.K.; supervision, M.D.; validation, E.B., M.D.; visualisation, D.G. and R.K.; writing—original draft, E.B., M.D., D.G., R.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Swiss National Science Foundation under grant CRSII5_173719.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Research data generated by the *impresso* project is available on Zenodo: <https://zenodo.org/communities/impresso/> (accessed on 5 August 2021).

Acknowledgments: We are grateful for constructive feedback on an earlier version of this document by two anonymous reviewers. We would also like to thank Stefan Jänicke for input in the preparation of the paper. Finally we would like to acknowledge the contributions of the rest of the *impresso* team (in particular Thijs van Beek, Simon Clematide, Maud Ehrmann, Matteo Romanello, Paul Schroeder, Philip Ströbel as well as supervisors Andreas Fickers, Frédéric Kaplan, Martin Volk and Lars Wieneke and all associated partners for their contributions to the design and creation of the *impresso* app.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alharbi, M.; Laramée, R.S. SoS TextVis: An Extended Survey of Surveys on Text Visualization. *Computers* **2019**, *8*, 17. [CrossRef]
2. Windhager, F.; Federico, P.; Schreder, G.; Glinka, K.; Dörk, M.; Miksch, S.; Mayr, E. Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges. *IEEE Trans. Visual. Comput. Graph.* **2019**, *25*, 2311–2330. [CrossRef] [PubMed]

3. Glinka, K.; Meier, S.; Dörk, M. Visualising the 'Un-seen': Towards Critical Approaches and Strategies of Inclusion in Digital Cultural Heritage Interfaces. In *Kultur und Informatik: Cross Media*, 1st ed.; Verlag Werner Hülsbusch, Glückstadt, 2015; pp. 105–117. Available online: https://uclab.fh-potsdam.de/wp/wp-content/uploads/Visualising_the_Unseen_Kul15.pdf (accessed on 15 June 2021).
4. Smith, D.A.; Cordell, R.; Mullen, A. Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers. *Am. Lit. Hist.* **2015**, *27*, E1–E15. [[CrossRef](#)]
5. Pinson, G. La Réimpression dans la Presse Francophone du 19e Siècle—Numapresse. Available online: <http://www.numapresse.org/2017/10/13/la-reimpression-dans-la-presse-francophone-du-19e-siecle-g-pinson-j-schuh-avec-p-c-langlais/> (accessed on 15 June 2021).
6. Oiva, M.; Nivala, A.; Salmi, H.; Latva, O.; Jalava, M.; Keck, J.; Domínguez, L.M.; Parker, J. Spreading News in 1904. *Media Hist.* **2020**, *26*, 391–407. [[CrossRef](#)]
7. Marjanen, J.; Zosa, E.; Hengchen, S.; Pivovarov, L.; Tolonen, M. Topic Modelling Discourse Dynamics in Historical Newspapers. In Proceedings of the 5th Conference Digital Humanities in the Nordic Countries (DHN 2020), Riga, Latvia, 21–23 October 2020; Schloss Dagstuhl Leibniz Center for Informatics: Riga, Latvia, 2020; pp. 63–77.
8. Martínez-Ortiz, C.; Kenter, T.; Wevers, M.; Huijnen, P.; van Eijnatten, J. Design and implementation of ShiCo: Visualising shifting concepts over time. *HistoInformatics* **2016**, *16*, 9.
9. Huistra, H.; Mellink, B. Phrasing history: Selecting sources in digital repositories. *Hist. Methods A J. Quant. Interdiscip. Hist.* **2016**, *49*, 220–229. [[CrossRef](#)]
10. Willems, M.; Atanassova, R. Europeana Newspapers: Searching Digitized Historical Newspapers from 23 European Countries. *Insights* **2015**, *28*, 51–56. [[CrossRef](#)]
11. Allen, R.B.; Sieczkiewicz, R. How Historians use Historical Newspapers. *Proc. Am. Soc. Inf. Sci. Technol.* **2010**, *47*. [[CrossRef](#)]
12. Liu, S.; Wang, X.; Collins, C.; Dou, W.; Ouyang, F.; El-Assady, M.; Jiang, L.; Keim, D.A. Bridging Text Visualization and Mining: A Task-Driven Survey. *IEEE Trans. Vis. Comput. Graph.* **2019**, *25*, 2482–2504. [[CrossRef](#)]
13. Unsworth, J. Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This? Available online: <https://johnunsworth.name/Kings.5-00/primitives.html> (accessed on 15 June 2021).
14. Thai, M.T.; Wu, W.; Xiong, H. (Eds.) *Big Data in Complex and Social Networks*; Chapman and Hall/CRC: New York, NY, USA, 2016. [[CrossRef](#)]
15. Fickers, A. Towards A New Digital Historicism? Doing History in the Age of Abundance. *VIEW J. Eur. Telev. Hist. Cult.* **2012**, *1*, 19–26. [[CrossRef](#)]
16. Matteo, R.; Ehrmann, M.; Clematide, S.; Guido, D. The Impreso System Architecture in a Nutshell. Technical Report, EuropeanaTech Insights. 2020. Available online: <https://infoscience.epfl.ch/record/283595> (accessed on 15 June 2021).
17. Ehrmann, M.; Bunout, E.; Düring, M. Historical Newspaper User Interfaces: A Review. IFLA WLIC 2019. 2017. Available online: <http://library.ifla.org/2578/> (accessed on 15 June 2021).
18. Hechl, S.; Langlais, P.C.; Marjanen, J.; Oberbichler, S.; Pfanzelter, E. Digital Interfaces of Historical Newspapers: Opportunities, Restrictions and Recommendations. *HistoInformatics* **2021**. [[CrossRef](#)]
19. Moreux, J.P. Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment. IFLA News Media Section. 2016. Available online: <https://hal-bnf.archives-ouvertes.fr/hal-01389455> (accessed on 15 June 2021).
20. Viola, L.; Verheul, J. The GeoNewsMiner: An Interactive Spatial Humanities Tool to Visualize Geographical References in Historical Newspapers. *Dig. Human.* **2020**. [[CrossRef](#)]
21. Franke, M.; John, M.; Knabben, M.; Keck, J.; Blascheck, T.; Koch, S. LilyPads: Exploring the Spatiotemporal Dissemination of Historical Newspaper Articles. In Proceedings of the 11th International Conference on Information Visualization Theory and Applications, Valletta, Malta, 27–29 February 2021; pp. 17–28. [[CrossRef](#)]
22. Spenke, M.; Beilken, C.; Berlage, T. FOCUS: The Interactive Table for Product Comparison and Selection. In Proceedings of the 9th Annual ACM Symposium on User Interface Software and Technology—UIST'96, Seattle, DC, USA, 6–8 November 1996; ACM Press: Seattle, DC, USA, 1996; pp. 41–50. [[CrossRef](#)]
23. Ahlberg, C.; Williamson, C.; Shneiderman, B. Dynamic Queries for Information Exploration: An Implementation and Evaluation. *CHI* **1992**. [[CrossRef](#)]
24. Ochigame, R.; Ye, K. Search Atlas: Visualizing Divergent Search Results Across Geopolitical Borders. In *Designing Interactive Systems Conference 2021*; ACM: New York, NY, USA, 2021; pp. 1970–1983. [[CrossRef](#)]
25. Sun, L.; Dong, S.; Ge, Y.; Fonseca, J.P.; Robinson, Z.T.; Mysore, K.S.; Mehta, P. DiVenn: An Interactive and Integrated Web-Based Visualization Tool for Comparing Gene Lists. *Front. Genet.* **2019**. [[CrossRef](#)] [[PubMed](#)]
26. Ren, X.; Lv, Y.; Wang, K.; Han, J. Comparative Document Analysis for Large Text Corpora. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, Cambridge, UK, 6–10 February 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 325–334. [[CrossRef](#)]
27. Jähnichen, P.; Oesterling, P.; Heyer, G.; Liebmann, T.; Scheuermann, G.; Kuras, C. Exploratory Search Through Visual Analysis of Topic Models. *Dig. Human. Quart.* **2017**, *11*. Available online: <http://www.digitalhumanities.org/dhq/vol/11/2/000296/000296.html> (accessed on 15 June 2021).

28. Sievert, C.; Shirley, K. LDavis: A Method for Visualizing and Interpreting Topics. In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, MD, USA, 27 June 2014; Association for Computational Linguistics: Baltimore, MD, USA, 2014; pp. 63–70. [[CrossRef](#)]
29. Jänicke, S.; Geßner, A.; Büchler, M.; Scheuermann, G. Visualizations for Text Re-use. In Proceedings of the 2014 International Conference on Information Visualization Theory and Applications (IVAPP), Lisbon, Portugal, 5–8 January 2014; pp. 59–70. [[CrossRef](#)]
30. Azad, H.K.; Deepak, A. Query expansion techniques for information retrieval: A survey. *Inf. Process. Manag.* **2019**, *56*, 1698–1735. [[CrossRef](#)]
31. Veerasamy, A.; Belkin, N.J. Evaluation of a Tool for Visualization of Information Retrieval Results. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, 18–22 August 1996; Association for Computing Machinery: New York, NY, USA, 1996; pp. 85–92. [[CrossRef](#)]
32. Hearst, M.A. TileBars: Visualization of Term Distribution Information in Full Text Information Access. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; ACM Press/Addison-Wesley Publishing Co.: New York, NY, USA, 1995; pp. 59–66. [[CrossRef](#)]
33. Hoeber, O.; Yang, X.D. Evaluating WordBars in Exploratory Web Search Scenarios. *Inf. Process. Manag.* **2008**, *44*, 485–510. [[CrossRef](#)]
34. Hoeber, O.; Liu, H. Comparing Tag Clouds, Term Histograms, and Term Lists for Enhancing Personalized Web Search. In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Los Alamitos, CA, USA, 31 August–3 September 2010. [[CrossRef](#)]
35. Havre, S.; Hetzler, E.; Perrine, K.; Jurrus, E.; Miller, N. Interactive Visualization of Multiple Query Results. In Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01), Sacramento, CA, USA, 28 October–1 November 2007; IEEE Computer Society: Washington, DC, USA, 2001; p. 105.
36. Klouche, K.; Ruotsalo, T.; Micallef, L.; Andolina, S.; Jacucci, G. Visual Re-Ranking for Multi-Aspect Information Retrieval. In Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, Oslo, Norway, 7–11 March 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 57–66. [[CrossRef](#)]
37. Wang, R.Y.; Strong, D.M. Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manag. Inf. Syst.* **1996**, *12*, 5–33. [[CrossRef](#)]
38. Taleb, I.; Serhani, M.A.; Dssouli, R. Big Data Quality: A Survey. In Proceedings of the 2018 IEEE International Congress on Big Data (BigData Congress), San Francisco, CA, USA, 2–7 July 2018; pp. 166–173. [[CrossRef](#)]
39. Liu, S.; Andrienko, G.; Wu, Y.; Cao, N.; Jiang, L.; Shi, C.; Wang, Y.S.; Hong, S. Steering Data Quality with Visual Analytics: The Complexity Challenge. *Vis. Inf.* **2018**, *2*, 191–197. [[CrossRef](#)]
40. Kandel, S.; Parikh, R.; Paepcke, A.; Hellerstein, J.M.; Heer, J. Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment. In Proceedings of the International Working Conference on Advanced Visual Interfaces, Capri Island, Italy, 21–25 May 2012; Association for Computing Machinery: New York, NY, USA, 2012; pp. 547–554. [[CrossRef](#)]
41. Bors, C.; Gschwandtner, T.; Miksch, S. Visually Exploring Data Provenance and Quality of Open Data. *Posters* **2018**, *3*. [[CrossRef](#)]
42. Hitchcock, T. Confronting the Digital: Or How Academic History Writing Lost the Plot. *Cult. Soc. Hist.* **2013**, *10*, 9–23. [[CrossRef](#)]
43. Hoekstra, R.; Koolen, M. Data Scopes for Digital History Research. *Hist. Methods A J. Quant. Interdiscip. Hist.* **2019**, *52*, 79–94. [[CrossRef](#)]
44. Koolen, M.; van Gorp, J.; van Ossenbruggen, J. Toward a Model for Digital Tool Criticism: Reflection as Integrative Practice. *Digit. Scholarsh. Humanit.* **2019**, *34*, 368–385. [[CrossRef](#)]
45. Bunout, E. Collections of Digitised Newspapers as Historical Sources—Parthenos Training. 2019. Available online: <https://training.parthenos-project.eu/sample-page/digital-humanities-research-questions-and-methods/collections-of-digital-newspapers-as-historical-sources/> (accessed on 15 July 2021).