2019 IFLA WLIC Athens

Submitted on: 02/09/2019

Historical Newspaper User Interfaces: A Review

Maud Ehrmann
Digital Humanities Laboratory (DHLAB), EPFL, Lausanne, Switzerland.
maud.ehrmann@epfl.ch

Estelle Bunout
Centre for Contemporary and Digital History (C2DH), Luxembourg University, Luxembourg.
estelle.bunout@uni.lu

Marten Düring
Centre for Contemporary and Digital History (C2DH), Luxembourg University, Luxembourg.
marten.duering@uni.lu

Abstract:

After decades of large-scale digitization, many historical newspaper collections are just one click away via online portals developed and supported by various public or private stakeholders. Initially offering access to full text search and facsimiles visualization only, historic newspaper user interfaces are increasingly integrating advanced exploration features based on the application of text mining tools to digitized sources. As gateways to enriched material, such interfaces are however not neutral and play a fundamental role in how users perceive historical sources, understand potential biases of upstream processes and benefit from the opportunities of datafication. What features can be found in current interfaces, and to what degree do interfaces adopt novel technologies? This paper presents a survey of interfaces for digitized historical newspapers with the aim of mapping the current state of the art and identifying recent trends with regard to content presentation, enrichment and user interaction. We devised 6 interface assessment criteria and reviewed twenty-four interfaces based on ca. 140 predefined features.

Keywords: digitized historical newspapers, user interfaces, digital scholarship

1. Introduction

Historical newspapers are mirrors of past societies. They reflect the political, moral, and economic environments in which they were produced and they hold dense, continuous, and multi-level information which can help us understand how contemporaries experienced their present. This makes them indispensable sources for research, for both academic and non-academic users.

Their value is reflected in digitization efforts in recent years: regional and national libraries as well as transnational bodies and commercial operators have made considerable investments in newspaper digitization with the aim of both making them available to larger audiences and ensuring the preservation of sometimes fragile paper originals (Stroeker and Vogels 2012; Center for Research Libraries 2015). This effort for systematic digitization has yielded large-scale collections of digitized newspapers at regional, national and international levels. Remote access to such collections significantly lowers the bar for academic and non-academic users alike to select them as sources for their research, but also opens up new opportunities and transforms research practices (Bingham 2010; Milligan 2013; Putnam 2016).

These digitized sources are subject to extensive processing. There are de facto standard treatments such as OCR, OLR and metadata collection. More recently, however, advanced techniques from the field of natural language processing (NLP) are being deployed with the intention to facilitate interactions that go far beyond basic keyword search, browsing and close reading. These include n-gram frequencies, named entity recognition and disambiguation, techniques for the detection of latent semantic structures such as topic modelling, content recommendations as well as event and text re-use detection. Arguably, these tools enrich the digitized sources and promise to change how users engage with newspapers in particular, and digitized documents in general. Potential enhancements include advanced search and discovery functionalities, recommendation services and topic filters.

As a result, and adopting the view put forward by Pelle Snickars (Jarlbrink and Snickars 2017), we understand digitized newspapers as complex objects determined by multiple layers of processing and datafication. However, users need to be aware of the consequences, resulting biases and opportunities of processing and datafication. User interfaces, whose purpose is to serve as gateways to this enriched content and to relay such information, play a major role in this regard. Not only do they control what users can learn about the digitized content; they also actively shape user workflows by offering different selections of tools and features for searching and exploring that content. These two aspects make user interfaces an exciting field of study, and they deserve to be assessed critically with regard to the opportunities they create and the problems they raise: What features do existing newspaper interfaces offer and what criteria should be used to assess them? What are the key aspects interface design should focus on in order to accommodate text analysis research tools and their usage by humanities scholars? A closer look at the feature sets incorporated by newspapers interfaces reveals the degree to which novel technologies are being adopted by institutions and how they are being envisioned to serve their respective audiences.

Given that only few published user studies exist (Crymble 2016), although a lot of effort is put into collecting information to improve the user's interaction with the collections (Rautiainen 2016), this paper presents a survey of interfaces for digitized historical newspapers with the aim of mapping the current state of the art and identifying recent trends with regard to content presentation, enrichment and user interaction. To this end we have identified six assessment criteria which we believe capture both current and prospective intentions of content providers and user requirements, namely: source criticism, content search, content filtering, generosity, user content management and exploration, and connectivity.

Our own interest in newspaper interfaces stems from an ongoing interdisciplinary research project. In *impresso – Media Monitoring of the Past*[1], computational linguists, designers and historians are working together to enrich a corpus of Swiss and Luxembourgish newspapers with named entities, topic models, image search, text re-use detection and query suggestions based on word embeddings. Access to these enhancements is provided via a newly developed user interface, designed specifically to suit the needs of scholars. This endeavour reflects a wider movement fuelled by several research projects and initiatives based on historical newspaper processing and enrichment, which makes the assessment of user interfaces all the more relevant.[2]

The remainder of this paper is organized as follows: after the discussion of related work (Section 2), we present our assessment approach (Section 3), considering assessment criteria and methodological aspects. We then outline the interface survey (Section 4), with a close examination of various features, and a general evaluation. Finally, we briefly discuss needs and priorities and conclude (Section 5).

2. Related work

In the cultural heritage domain user interfaces are mainly developed in relation with digital libraries, which greatly vary in terms of size, scope and usage. User interfaces are usually examined from two perspectives: study of user behaviour by interface designers and developers on the one hand, and review of interfaces by interface users on the other.

User behaviour studies typically rely on two main approaches: quantitative, with automatically generated data such as query logs and web analytics, and qualitative, with surveys via questionnaires and interviews. Despite the relatively easy access to user-generated data of digitized newspaper interfaces, only few analyses have been conducted – or published – by libraries. In such studies, libraries are interested in getting to know their users, who usually correspond to educated “laymen”, genealogists or academic users (Ayres 2013; Geiger and Zarndt 2013). Besides user profiles, understanding how users search is key to understand how to improve and best parametrize interfaces. Analysis of user behaviours often rely on the detection of usage patterns in query logs collected over a particular period of time, together with information about visited pages, metadata filters, and visit times. In this respect, it has been shown that keywords very often include named entities (Chardonnens et al. 2017; Sumikawa et al. 2019; De Wilde and Hengchen 2016), that faceted search prevails over non-faceted search (Bogaard et al. 2019), and that some time periods and titles are more searched than others (Gooding 2016; Sumikawa et al. 2019). It is also possible to study which interface features are used most (Marschall 2017), and the relationship between metadata, clicks and downloads: if one period is very popular but download counts are low, this may indicate poor quality of the scanned documents (Chardonnens et al. 2017). Although web analytics are made on partial information (for technical and privacy reasons), such studies are helpful to better understand user needs and improve their experience.

Another source of feedback on the use of digitized newspapers interfaces are user interviews and interface reviews by users. User interface reviews are highly informative, indicating which

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

[2] See e.g. the projects *Europeana Newspapers* (Neudecker and Antonacopoulos 2016), *Living with Machines* (www.turing.ac.uk/research/research-projects/living-machines), *Oceanic exchanges* (Cordell and Smith n.d.) and *NewsEye* (https://www.newseye.eu).

features are prominently perceived by users and collecting explicit evaluations about positive aspects and problems encountered during the use (Nicholson 2015; Natale 2019). User interviews are somehow more constrained, but equally useful to understand needs and expectations, such as the online survey conducted by the national libraries of Austria, Finland and France which demonstrates the wish for different tools (e.g. named entity recognition, topic modelling, keyword suggestions) in order to get better or more specific results (Oberbichler et al. 2019). The user testing sessions conducted on the very first version of the Europeana Newspaper portal emphasized similar conclusions: users appreciate fine-grained faceting more than browsing (Atanassova 2014).

Apart from user studies, publications analyzing interfaces for digitized newspaper collections tend to focus on the opportunities (and drawbacks) offered by digitization and text mining techniques. One has to turn towards more generic studies to find prospective surveys, such as the one conducted by Gibbs an Owens on the use of various digital tools by historians (Gibbs and Owens 2012). Among other things, they underline the need for a "social contract" between 'tool builders' and users with further dissemination and pedagogy. In this regard, there is currently little indication about interface developers' attempts to support digital literacy.

It should also be noted that in some institutions – especially national libraries – digitized newspaper collections share the same interface with other types of collections. In this context interfaces need to weigh newspaper-specific against general-purpose interface features (Whitelaw 2015). In return, newspaper interfaces can gain from the development of cultural heritage interfaces, especially in terms of visualization and exploration (Glinka, Meier, and Dörk 2015).

Overall, work on historical newspaper interfaces are rather few and focus almost exclusively on user studies conducted in isolation. They reveal a strong appreciation of the availability of newspaper portals as well as the need for advanced search capacities. We could not find an extensive analysis of historical newspaper interfaces and wish to contribute toward this objective with the present survey.

3. Interface assessment approach

Below we present the criteria we propose to use to assess historical newspaper interfaces and outline our evaluation methodology.

3.1. Assessment criteria

As mentioned earlier, newspaper interfaces offer access to objects that are complex, by virtue of their very nature and as a result of the application of multiple pre- and post-processing techniques. Presenting such transformed, layered, enriched sources in ways that best meet user needs and requirements and accommodate the potential of text and image processing is challenging, but essential. In order to assess interface capabilities from this perspective, we propose a set of six criteria which we believe cover present and future challenges for newspaper interfaces:

1. **Source criticism** (or *What am I looking at?)* describes the critical assessment of documents based on provenance and awareness of context. Digitized and datafied sources require additional transparency in the form of digitized provenance information but also actionable

information on the kind of processing and enrichment a source has undergone. We consider information about the compilation of the corpus, metadata about newspaper collections, titles, and individual units (issues, pages, articles), as well as performance scores for OCR or entity linking.[3]

2. **Content search** (or *How do I engage with the material/content?*), to evaluate the extent to which interfaces can help understand the information space and identify relevant content. Plain (OCRed) text is the primary entry point into digitized newspaper material, but semantic enrichment and linking via information extraction and text analysis tools provide valuable and complementary alternatives. We consider the search functionalities offered by interfaces on various types of (enriched) material.

3. **Content filtering** (or *How do I select?*), to determine the extent to which interfaces support the narrowing down of search results from large corpora. Together with search tools that locate items, content filtering is a powerful tool to iteratively hide unwanted items and refine the scope of exploration. We consider result sorting and result filtering functionalities.

4. **Generosity** (or *How do I discover?*), to appraise whether interfaces feature functionalities which, beyond keyword search and content browsing, help users discover relevant content they had not anticipated to find. We consider corpus presentation, result display modes and recommendation techniques. Contrary or complementary to content filtering, these features help expand the list of results.

5. **User content management and exploration** (or *How do I work?*), to evaluate the extent to which interfaces allow scholars to collect, organize, tag and compare their own collections of material so as to be able to work on specific research questions, in isolation or in collaboration. We consider here personal work space functionalities.

6. **Connectivity** (or *How do I go beyond?*), to assess how interfaces interlink their collections so as to allow the study of digitized newspapers and other sources across institutional silos. We consider interlinking at the level of content (e.g. entity linking) and metadata.

3.2. Methodology

The assessment of these six high-level criteria was based on 139 properties that were compiled for 24 interfaces (see Annex A). These 139 properties encompass what is technically feasible and already available for interfaces for collections of digitized newspapers.

Interface selection – Our selection of interfaces was iterative and guided by several (pragmatic) considerations. Since we are involved in a project on digitized newspapers, we were able to make an initial selection of interfaces we already knew. This list was then expanded with suggestions from colleagues and by referring to the Wikipedia list of online newspaper archives.[4]

[3] Transparency naturally also applies to various enrichment processes, but those are still quite rare among existing interfaces.

[4] https://en.wikipedia.org/wiki/Wikipedia:List_of_online_newspaper_archives

Our main criteria for inclusion were: the interface should not be too old nor too ‘basic’, it should be in a language that at least one of the reviewers understands, and the total number of interfaces should not be too high so that they can be reviewed in a reasonable amount of time. The final selection included a majority of public institution-supported interfaces, with nine national libraries, seven regional or city libraries (including US states), and one at European level. Other interfaces included commercial portals (1) and interfaces developed by publishers of still-existing titles (3) and by semi-public consortia (3). With regard to country distribution, the most represented are federal states (the United States and Switzerland) and countries where subnational institutions develop their own interfaces. All the other countries (in Europe, plus Australia) have one portal from their national libraries. The list of reviewed interfaces is presented in Table 1.

Feature selection – Our criteria for choosing the interface properties we wanted to observe were based on a mix of initial knowledge of what should be expected and iterative refinement during the investigation. These features were further organized into 14 ‘families’ of functionalities, which were then mapped to the high-level assessment criteria. The full list of features, organized per family and with their corresponding mapping to high-level criteria, is presented in Table 2 in Annex A.

Survey – Finally, the survey itself was conducted by the authors, who visited each interface and recorded their observations in a spreadsheet. Beyond the clarification of some properties, the survey did not raise any major difficulties.

3.3. Observations

Collecting information about newspaper interfaces is not an error-free process, and our study has some inherent limitations. First, our interface selection is not a statistically representative sample of all currently available digitized newspaper interfaces. One option would have been to spend more time inventorying interfaces, but we did not deem it relevant for our purpose and it would have been almost impossible to know if we had discovered all the interfaces available. Second, our reviews were not verified by a third person to ensure agreement. But we believe that despite the potential for human error, an interface survey does not leave too much room for interpretation and is a rather neutral process. Finally, interfaces are ‘living’ digital objects, and our review may well be out of date within a few months or years.

Despite these imperfections, we believe that our interface sample and review process are sufficiently diverse, balanced and sound, and that they provide a solid basis to estimate the current state of the art with regard to digitized newspaper interfaces.

The survey material (dataset and Jupyter notebook) is available on the software development platform GitHub[5], and also published on the open access repository Zenodo[6].

[5] https://github.com/impresso/impresso-interface-review

[6] https://doi.org/10.5281/zenodo.3369875

Name	Country	Approx. creation date	Access
ANNO	AT	2003	free
Ancestry	US	unk.	paywall
British Newspaper Archives	UK	2008	free
California Digital Newspaper Collection	US	unk.	free
Chronicling America	US	2003	free
Colorado Historical Newspaper Collection	US	unk.	free
Delpher	NL	unk.	free
DigiPress	DE	at least 2016	free
Difmoe	DE	unk.	free
E-luxemburgensia	LU	unk.	free
E-newspaperarchives	CH	2018	free
Europeana newspapers	EU	2019	free
L'Express	CH	2013	free
Gallica (newspapers and journals)	FR	1997	free
Georgia Historic Newspapers	US	2007	free
Libraria (Ukrainian online periodicals archive)	UA	2012	on site
New York Times	US	unk.	paywall
Polona	PL	unk.	free
Retronews	FR	2016	freemium
Scriptorium	CH	2012	free
StaBi	DE	unk.	free
Tessmann	IT	unk.	free
Le Temps archives	CH	2016	free
Trove	AU	2007	free

Table 1: List of reviewed interfaces.

4. Digitized newspaper interfaces

Based on our feature inventory, this section surveys digitized newspaper interfaces. We first give a general overview of the main characteristics we observed. Next, we examine each feature family more closely, organizing the interfaces into 'generations'. Finally, in light of this examination, we review the current landscape of digitized newspaper interfaces according to our assessment criteria.

4.1. General comments

Leaving aside information about the interface and about the newspaper collection (group 'a' and 'b' in Table 2), we are left with 125 features belonging to 12 families of functionalities, namely: *newspaper metadata* (12 features), *browsing* (5), *search options* (19), *result display* (8), *result sorting* (10), *result filtering* (22), *viewer* (8), *documentation on digitization* (10), *personal account and user interaction* (14), *connectivity* (3), *content enrichment* (10) and *Code and APIs* (4).

A first observation is that, in general, feature coverage is rather low. Looking at the extremes, we can observe that 78% of the features are covered by less than half of the interfaces, with 8% of all surveyed features not covered by any interface. These low-coverage features are distributed among the 14 families, and those covered by no interfaces belong to *result sorting*, *user interaction*, *documentation on digitization*, *content enrichment* and *APIs*. On the other side, 22% of the features are covered by at least half of the interfaces, and 11% by at least three quarters of them. Only three features figured on all the interfaces: 'keyword search' (*search options* family), 'facsimile display' and 'option to continue to next page' (*viewer*).

Next, we observe disparities among feature families: those best represented are *newspaper metadata*, *search*, *result filtering* and *viewer*, closely followed by *result sorting*, *result display* and *user interaction*. Not surprisingly, *enrichment*, *connectivity*, *information on digitization* and *APIs* are rather weak. Figure 1[7] shows how the 12 feature families 'scored'[8] for each interface.

In the same vein, and as shown on *Figure 2*, the most developed features relate to access and full text search, from *newspaper metadata* to *user interaction*, while more advanced features are still underdeveloped. The objective here is not to rank interfaces but rather to identify trends. It should be noted, though, that the reviewed interfaces were mainly developed by public institutions whose primary objectives at the time of launch were access and preservation rather than advanced functionalities for (scholar) users.

Finally, considering features from group 'a' and 'b' not shown in the Figures, it must be emphasized that almost half of the interfaces (11 out of 24) offer access to multilingual collections, although this contrasts with the absence of cross-lingual text processing. With regard to the access model, most of the portals we reviewed are free (sometimes with

[7] All figures are also available here: https://github.com/impresso/impresso-interface-review/tree/master/charts

[8] A 'score' corresponds to how many features of a specific feature family an interface covers, relatively to the total number of features for that family.

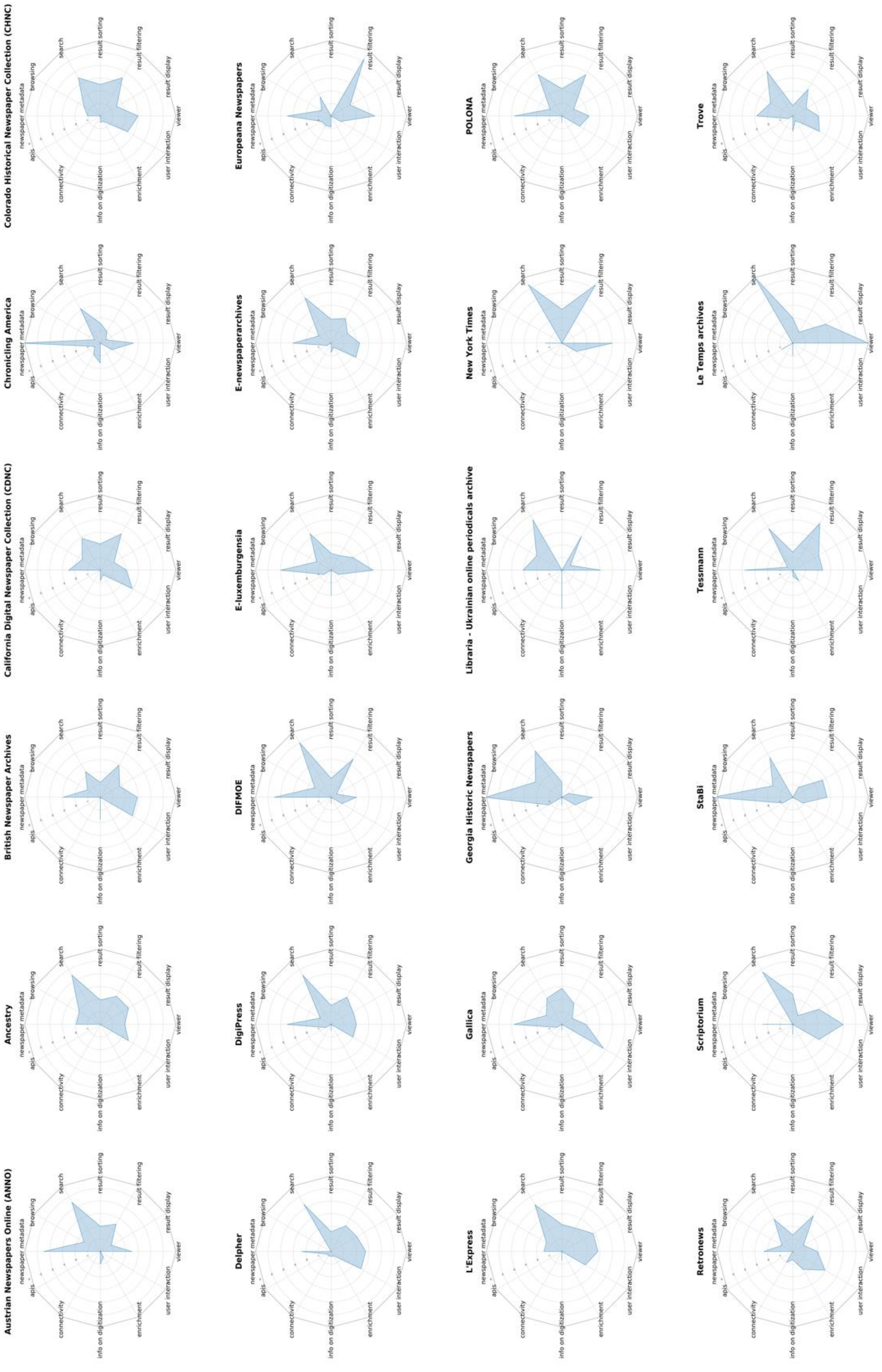

Figure 1: Distribution of the ‘scores’ for the 12 feature families across all interfaces (as percentages).

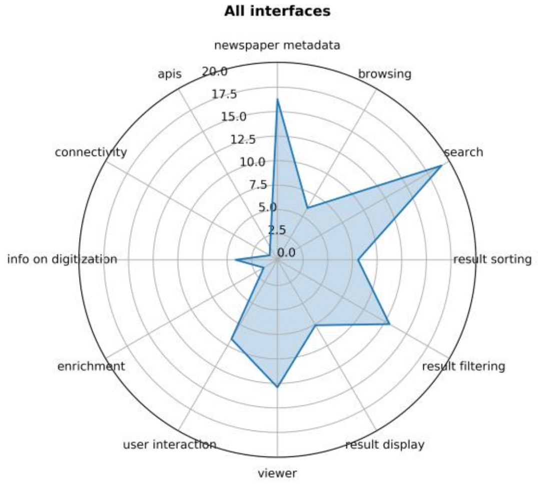

Figure 2: Aggregated count per feature family across all interfaces (in percent).

limitations such as embargoed images), 3 are behind a paywall, 1 offers a freemium plan and 1 is only available for consultation on site at the library.

4.2. Families of features and interface generations

As a means of examining the interfaces more closely, we decided to analyze the survey dataset by identifying families of features and interface 'generations'. Our use of the notion of 'generations' relates to the level of refinement and the comprehensive nature of functionalities rather than to any temporal dimension.[9]

First generation – The first stage in the development of digitized newspaper interfaces focused on **access** to digital sources via a combination of **full text search** and metadata facets, accompanied by the first generation of content viewers. This corresponds to the feature families of *newspaper metadata*, *browsing*, *search options*, *result display*, *result sorting*, *result filtering*, and *viewer*, which we examine hereafter.

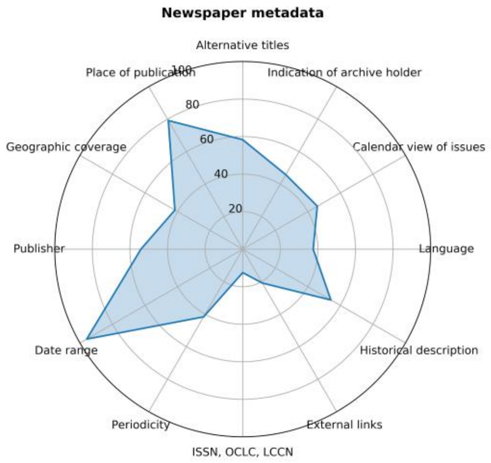

Figure 3: Distribution of features related to newspaper metadata.

[9] The two may coincide, but not necessarily.

As shown previously, *newspaper metadata* is a rather well-implemented feature family among our reviewed interfaces. Looking closer, however, one can observe that individual features are far from being covered by all interfaces (Figure 1). Out of 12 individual features, only 5 are supported by at least half of the interface: newspaper date ranges (95%), newspaper place of publication (79%), alternative, succeeding and related titles (58%), historical description (54%) and newspaper publisher (54%). These features are usually part of library 'traditional' metadata information and it is somehow expected to find them in digital portals. As for the others, if some are more difficult to collect and make available as filter (e.g. periodicity, geographic coverage, political orientation, external links) or not relevant in all cases (e.g. language), others could however be covered rather easily (calendar view of issues and indication of archive holder).

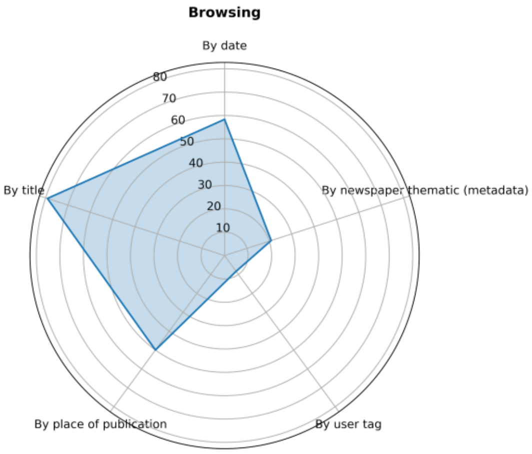

Figure 4: Distribution of browsing features.

Figure 4 consists of 5 single features only, with browsing by title, date and place of publication as the most prominent. Some interfaces have started to include more semantic browsing capacities, drawing on either newspaper themes (based on metadata) or user tags, but this is still not widespread.

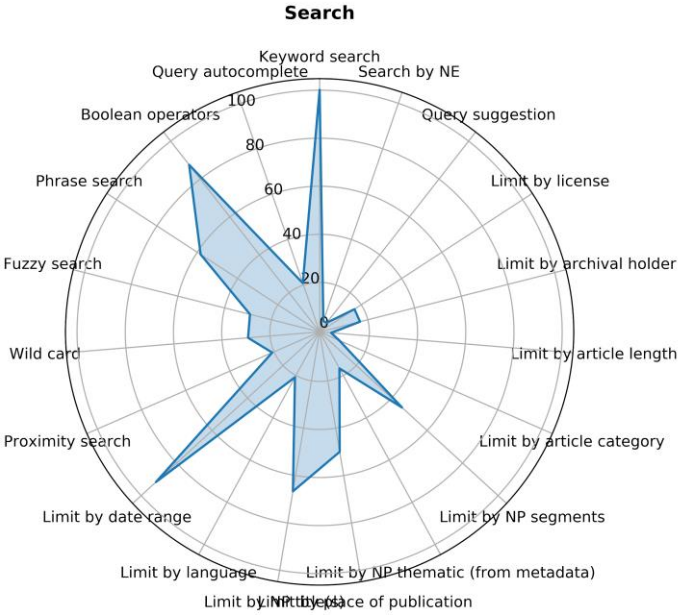

Figure 5: Distribution of features related to search.

The *search* family (Figure 5) is one of the largest, with 19 individual features which encompass basic and also more advanced characteristics (therefore also determining other interface generations). The most common feature is naturally keyword search, supported by 100% of the reviewed interfaces. The leading group also includes the option to limit by date range (91%), to use Boolean operators (87.5%), to limit by newspaper title(s) (66%), to search for multi-word expressions (58%), and to limit by place of publication (50%). All these features can be implemented using OCRed texts and metadata as they are, i.e. without further processing. It should also be noted that fuzzy search, proximity search, wildcards and query auto-completion are surprisingly rare and often hidden in interfaces. This is surprising considering that they are well integrated in out-of-the-box search engine software.

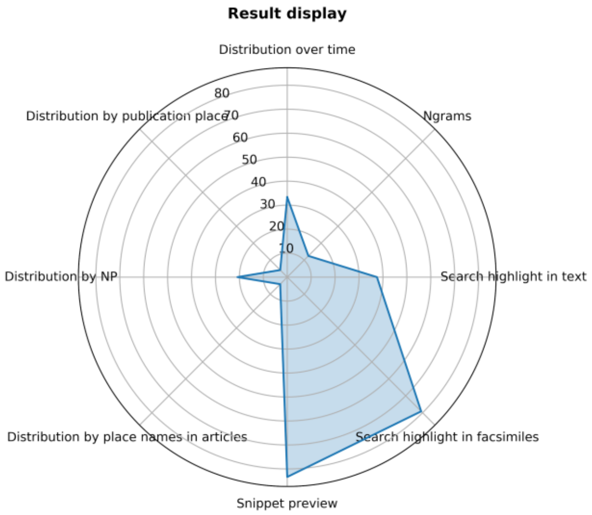

Figure 6: Distribution of features related to the display of search results.

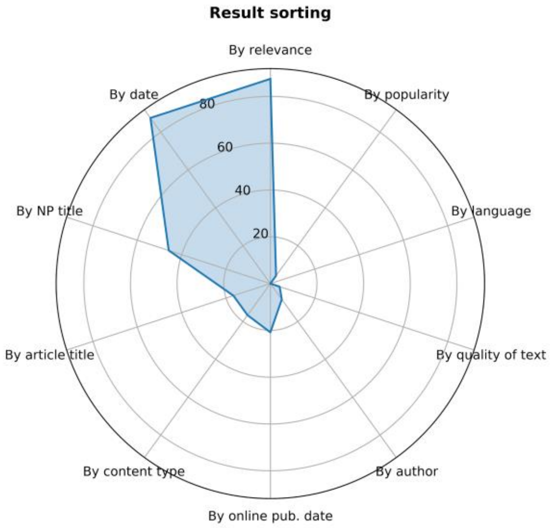

Figure 7: Distribution of features related to the sorting of search results.

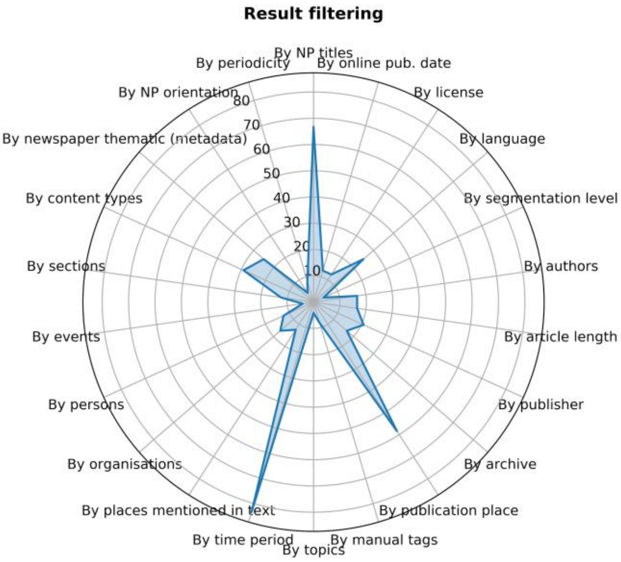

Figure 8: Distribution of features related to the filtering of search results.

The feature families *result display*, *result sorting* and *result filtering* (Figure 6, Figure 7 and Figure 8) comprise 8, 10 and 22 features respectively. Display options include snippet previews and highlighting in facsimiles in most interfaces (83% and 79%). However, rendering of result distribution (over time with n-grams, or by newspaper title(s) or place of publication) and highlights in text are not yet widespread. With regard to result filtering, this large family contains 3 top-ranked functionalities, namely filtering by time period, by title and by place of publication (83%, 66% and 58% respectively), all of which are metadata facets. The remaining features score below 30% and characterize other interface generations.

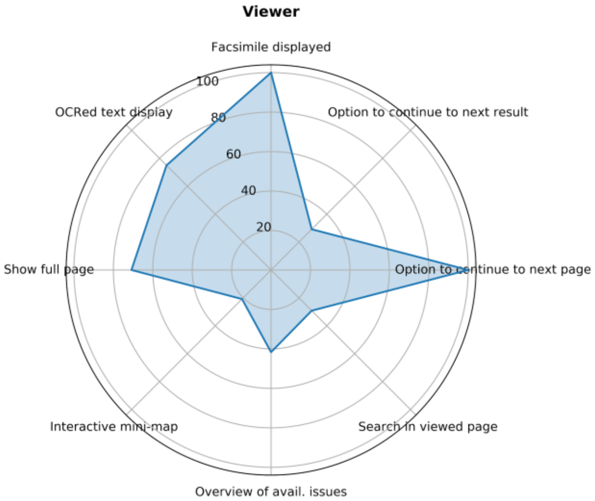

Figure 9: Distribution of features related to content viewers.

Finally, the *viewer* family (Figure 9) is rather well covered, with 4 features out of 8 scoring above 50% (facsimile display, OCR display, option to continue to next page and full-page view).

To conclude this first examination, it clearly appears that the most common features characterizing what we refer to as first-generation interfaces are based on raw textual material and metadata information already available in library bibliographical notices. Although these functionalities represent a significant step forward in terms of preservation and access, further advanced components are both needed and, considering the technical state of the art, to be expected.

Second generation – The second generation is characterized by **user interaction** functionalities. Figure 10 shows the selected features of this family, with the option of obtaining a permalink (50%), saving articles to favorites (45%), organizing articles into collections (41%) and screenshotting images (37%) among the highest-scoring features. A quarter of the interfaces also allow users to correct OCR and to save queries to favorites.

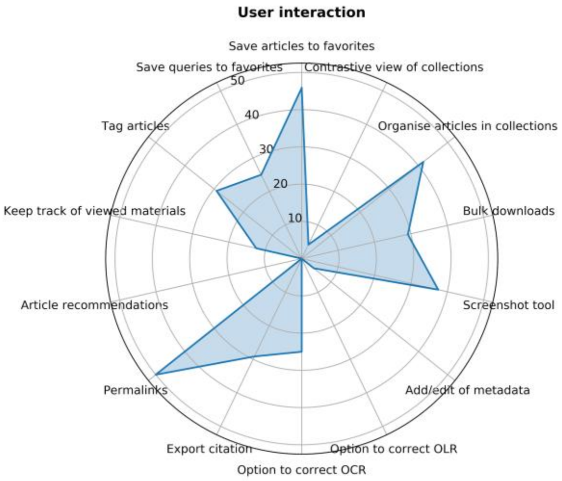

Figure 10: Distribution of features related to user interaction.

Third generation – The third generation refers to a more advanced set of functionalities, still in its infancy, which enable users and machines to explore newspaper content based on **semantic enrichment** acquired automatically via natural language processing. Here we naturally touch upon the feature families of *enrichment*, *connectivity* and *APIs*, which also impact *search* and *filtering* functionalities.

The feature set for *enrichment* is composed of text improvement capabilities (automatic and crowd-sourced post-OCR correction) and semantic enrichment capabilities based on information extraction (named entity recognition and classification, entity linking, event recognition, sentiment analysis) and document collection processing (topic modelling, text re-use), and recommendations derived from this information. As shown in Figure 11, although 16% of interfaces support crowd-sourced OCR post-correction, other features all scored below 8%, with a couple of recent interfaces offering named entity processing, topic modeling (not documented) and automatic post-OCR correction. Text re-use and sentiment analysis are not supported.

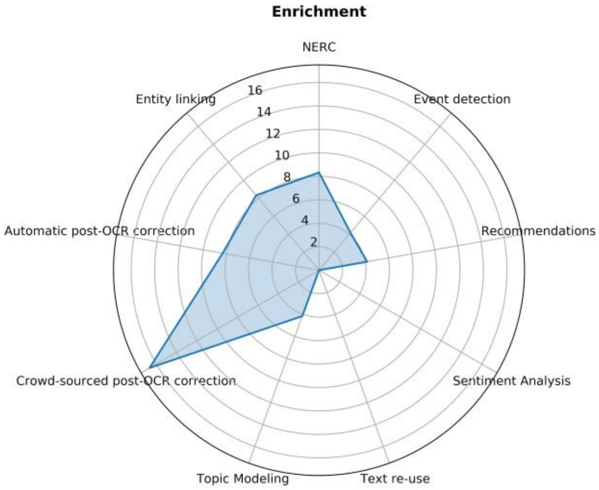

Figure 11: Distribution of features related to enrichment.

For connectivity criteria, the situation is not better: one interface provides third-party identifiers and 6 offer links to other repositories (2 of which are based on semantic web technologies). With respect to APIs, 5 offer an endpoint to programmatically query their collections, 7 implement the IIIF Image API, and none the Presentation API (at the time of inventory).[10]

As mentioned for the first generation, search options and filtering capacities are also impaired by the lack of content enrichment, as it could be used to directly search or facet upon semantic annotations.

Fourth generation – Finally, we place in the fourth-generation interfaces which do not yet exist but are under development in research projects, offering more transparency regarding corpus compilation, visual exploration, and personalization in the form of recommendations and query suggestions. With regard to transparency, this includes some newspaper metadata features which could be improved, as well as documentation of the digitization process and of automatic semantic enrichment. Figure 12 shows that apart from information about newspaper copyright and whether there is article-level segmentation, not a lot is shared.

Overall, the examination of these 12 feature families with 125 individual features provided us with an accurate picture of what have become 'standard' features (keyword search and viewer), what exists but could still be improved (newspaper metadata, result display, sorting and filtering and user interaction), and what is yet to be developed (enrichment and transparency). Interface generations follow a gradual development, determined by what is already available (first generation), aspects that merely regard interface development (second generation), features that require advanced processing of digital material, often in collaboration with specialists (third generation), and options that need both additional processing and communications (fourth generation).

[10] International Image Interoperability Framework: https://iiif.io/

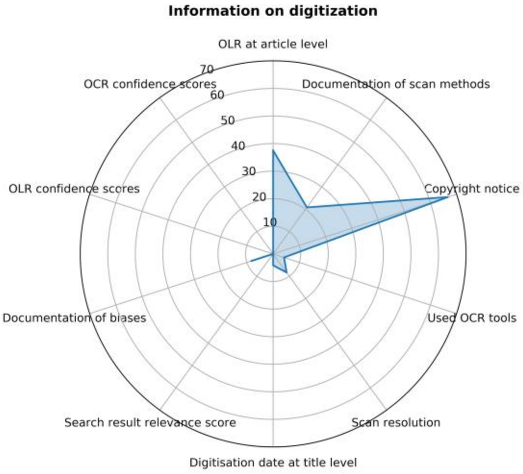

Figure 12: Distribution of features related to Information on digitisation.

4.3. Assessment

After our close examination of the survey dataset, this section zooms out from the many features and returns to the six high-level assessment criteria defined in Section 3.1. To do so, in Figure 13 we mapped each individual feature to high-level criteria, as presented in Table 1. Since some (extra) features could not be mapped or were not relevant, the mapping occurred only for 133 features (37 for *source criticism*, 39 for *content search*, 32 for *content filtering*, 10 for *generosity* and 4 for *connectivity*).

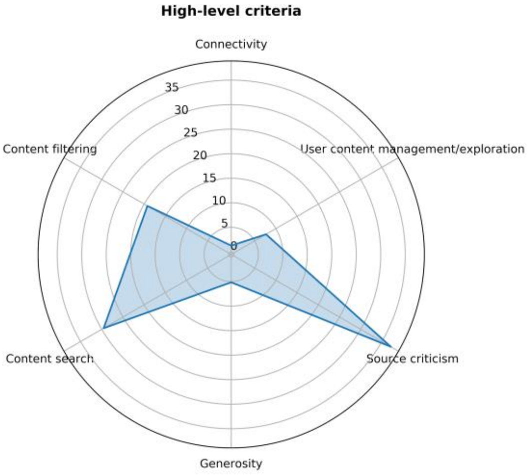

Figure 13: Mapping of relevant features to the high-level assessment criteria.

This view confirms the observations made in the previous sections. The latest state-of-the-art interfaces emphasize search and filtering functionalities based on provenance-related metadata which is typically collected by libraries. Features linked to connectivity, user content management/exploration and generosity (the three areas which we have identified to be representative for third and fourth generation interfaces) are comparably rare. They are disproportionately common in the most recent interfaces, however, which is indicative of a trend to develop such features further in future.

5. Discussion and outlook

In this paper we presented results from a survey of 24 user interfaces for digitized historical newspapers. Based on the analysis of ca. 140 features grouped into 12 families, we identified four generations of interfaces: the first focuses primarily on making content available online, the second on advanced user interaction with the content, the third on automated enrichment and the fourth on personalization and increased transparency. Interfaces were further assessed with regard to six high-level criteria we introduced.

Main findings can be summarized as follows. First, the set of surveyed interfaces sparsely covers the considered features, with many of them (ca. three quarters) being present in less than half of the interfaces. This confirms the gap between growing user expectations, encouraged by text mining progresses, and current interface capacities. Next, and it does not come as a surprise, the top-covered feature families include *newspaper metadata*, *search*, *result filtering* and *viewer* while the least covered one are *enrichment*, *connectivity*, *information on digitization* and *APIs*. More surprisingly, it appears that there is still quite some room for improvement amongst the top-covered families.

This survey, which outlines current strengths, weaknesses and trends among historical newspaper interfaces, also brings up a number of open questions and challenges. There is first the problem of finding the best trade-off between different aspects. With respect to audiences, how can we reconcile interfaces made for scholars vs. the general public? Should there be dedicated interfaces for each groups? With respect to the complexity of an interface, should all features and enrichments be visible and accessible? If not, which are the most valuable ones? To which extent should they integrate tools for the analysis of image and text mining outputs, at the risk of complexifying the uses, vs. an externalization via download functionalities for further analysis outside the interface environment? There are no clear answers to these questions but current initiatives around digitized newspapers – which build on numerous previous digitization efforts[11] – will certainly contribute clarifying these points.

Next, search and browsing would certainly benefit from additional metadata fields, for example on the size of print runs, areas of distribution, or a newspaper's readership. Librarians are in our experience generally open to such suggestions but rightfully point to a long-term process of standardization and consolidation with other libraries in contrast to the often more short-term perspectives of researchers. We find that there is a need for a continued discussion between the fields on new metadata standards but also concerning the process of data collection: Which contributions can be made by libraries, scholars, crowds and automated processing? How can such a collaboration be organized?

Stepping back from interfaces and tools, the question of the role of stakeholders also arises. How libraries, NLP researchers, designers and humanities scholars can best partner is not always evident. There seems to be a consensus on tasks situated at the extremes of the processing spectrum, such as digitization (libraries) and text mining (NLP researchers), but things are less clear for e.g. services around data (management, pre-processing, formats and standards, entry point maintenance), design and digital literacy training. Frontiers are moving and stakeholders are revisiting their roles (Moiraghi 2018; Claeyssens et al. 2019).

[11] Cf. the IMPACT (Balk and Ploeger 2009; Balk and Conteh 2011) and Europeana (Neudecker and Antonacopoulos 2016) projects.

Every step in the digitization process, and especially advanced computational processing such as topic modelling or named entity linking requires far-reaching decisions which shape the enriched content. In topic modelling for example, the algorithms, the parameters chosen (e.g. number of topics) as well as the composition of the underlying corpus can lead to very different results. Likewise, the disambiguation of named entities with the help of an e.g. Wikipedia-based system will inherently reproduce its temporal and socio-cultural biases. Users of enriched newspaper repositories should be able to take into account such underlying decisions and biases but also inherent imperfections. Interface design can convey this information and educate users to a limited extent. Classical documentation in combination with other means to train users are currently being explored.[12]

Finally, digitization brings with it the opportunity to break nation-based institutional silos and to take a global perspective on an infrastructure for newspaper collections. The challenges here are not technical but political (Zaagsma 2019): How can such operations be funded and questions of copyright be resolved when moving beyond the missions of individual institutions? What do merged collections mean for the operations of contributing institutions?

We need to stress again that any survey of this kind and perhaps also the questions we raise above are doomed to be outdated soon after their publication. We can only offer a snapshot of the state of the art in June 2019. We nevertheless believe that the interest in access to digitized newspapers and the trend towards more enrichment, interactivity, connectivity and personalization will persist in the foreseeable future.

Acknowledgments

We warmly thank Clemens Neudecker (Berlin State Library), Ralph Marschall (National Library of Luxembourg), Tuula Pääkkönen (National Library of Finland) and Liliane Regamey (Swiss National Library) for having shared with us various sources and pointers, as well as for their valuable feedback based on their experience as librarians involved in the development of historical newspaper interfaces. The authors also gratefully acknowledge the financial support of the Swiss National Science Foundation (SNSF) for the project *impresso – Media Monitoring of the Past* under grant number CR-SII5_173719.

References

Atanassova, Rossitza. 2014. "Improving the Discovery of European Historic Newspapers." In *IFLA WLIC 2014 - Lyon - Libraries, Citizens, Societies: Confluence for Knowledge*. Lyon, France. http://library.ifla.org/1038/.

Ayres, Marie-Louise. 2013. "Singing for Their Supper': Trove, Australian Newspapers, and the Crowd." In *IFLA WLIC 2013 - Singapore - Future Libraries: Infinite Possibilities*. http://library.ifla.org/245/.

Balk, Hildelies, and Aly Conteh. 2011. "IMPACT: Centre of Competence in Text Digitisation." In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, 155–160. HIP '11. New York, NY, USA: ACM. https://doi.org/10.1145/2037342.2037369.

[12] These can take for example the form of digital teaching materials like those provided by the Ranke2 project (https://ranke2.uni.lu/) and the PARTHENOS project (http://www.parthenos-project.eu/#hub). Both projects have forthcoming training materials for digitized newspapers created by the *impresso* project.

Balk, Hildelies, and Lieke Ploeger. 2009. “IMPACT: Working Together to Address the Challenges Involving Mass Digitization of Historical Printed Text.” *OCLC Systems & Services: International Digital Library Perspectives*, October. https://doi.org/10.1108/10650750911001824.

Bingham, A. 2010. “‘The Digitization of Newspaper Archives: Opportunities and Challenges for Historians.’” *Twentieth Century British History* 21 (2): 225–31. https://doi.org/10.1093/tcbh/hwq007.

Bogaard, Tessel, Laura Hollink, Jan Wielemaker, Jacco van Ossenbruggen, and Lynda Hardman. 2019. “Metadata Categorization for Identifying Search Patterns in a Digital Library.” *Journal of Documentation*. https://doi.org/10.1108/jd-06-2018-0087.

Center for Research Libraries. 2015. “The State of the Art. A Comparative Analysis of Newspaper Digitization to Date.” http://www.crl.edu/sites/default/files/d6/attachments/events/ICON_Report-State_of_Digitization_final.pdf.

Chardonnens, Anne, Ettore Rizza, Mathias Coeckelbergs, and Seth van Hooland. 2017. “Mining User Queries with Information Extraction Methods and Linked Data.” *Journal of Documentation*. https://doi.org/10.1108/JD-09-2017-0133.

Claeyssens, Steven, Andreas Degkwitz, Isabel Galina Russel, Silvia Gutiérrez, Hege Hosoien, Marian Lefferts, Sarah Potvin, and Lotte Wilms. 2019. “Libraries As Research Partner in Digital Humanities (DH2019 Workshop).” 2019. https://zenodo.org/communities/libraries-as-research-partner-2019.

Cordell, Ryan, and David Smith. n.d. “Oceanic Exchanges.” *Tracing Global Information Networks In Historical Newspaper Repositories, 1840-1914* (blog). Accessed September 14, 2017. http://oceanicexchanges.github.io/.

Crymble, Adam. 2016. “Digital Library Search Preferences amongst Historians and Genealogists: British History Online User Survey” 10 (4). http://www.digitalhumanities.org/dhq/vol/10/4/000270/000270.html.

De Wilde, Max, and Simon Hengchen. 2016. “Semantic Enrichment of a Multilingual Archive with Linked Open Data,” January.

Geiger, Brian, and Frederick Zarndt. 2013. “What Motivates Library Crowdsourcing Volunteers?” Education presented at the American Library Association Conference. https://www.slideshare.net/cowboyMontana/what-motivates-library-crowdsourcing-volunteers-20130630-ala-lita.

Gibbs, Fred, and Trevor Owens. 2012. “Building Better Digital Humanities Tools: Toward Broader Audiences and User-Centered Designs.” *Digital Humanities Quarterly* 006 (2).

Glinka, Katrin, Sebastian Meier, and Marian Dörk. 2015. “Visualising the ‘Un-Seen’: Towards Critical Approaches and Strategies of Inclusion in Digital Cultural Heritage Interfaces.” *Kultur Und Informatik. Vwh*, 105–118. https://pdfs.semanticscholar.org/28d6/5220f641f8fc4c7d79191599fc086012d4f0.pdf.

https://pdfs.semanticscholar.org/28d6/5220f641f8fc4c7d79191599fc086012d4f0.pdf.

Gooding, Paul. 2016. “Exploring the Information Behaviour of Users of Welsh Newspapers Online through Web Log Analysis.” *Journal of Documentation*, March. https://doi.org/10.1108/JD-10-2014-0149.

Jarlbrink, Johan, and Pelle Snickars. 2017. “Cultural Heritage as Digital Noise: Nineteenth Century Newspapers in the Digital Archive.” *Journal of Documentation* 73 (6): 1228–43. https://doi.org/10.1108/JD-09-2016-0106.

Marschall, Ralph. 2017. “Improving the User Experience of a Digital Content Viewer through Advanced Analytics.” In *2017 IFLA International News Media Conference*. Reykjavík, Iceland. https://ifla2017.landsbokasafn.is/#welcome.

Milligan, Ian. 2013. “Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010.” *The Canadian Historical Review* 94 (4): 540–69. https://muse.jhu.edu/article/527016.

Moiraghi, Eleonora. 2018. “Le projet Corpus et ses publics potentiels.” Report. https://hal-bnf.archives-ouvertes.fr/hal-01739730.

Natale, Enrico. 2019. “Compte Rendu: E-Newspaperarchives.Ch.” Infoclio.Ch. 2019. https://infoclio.ch/fr/compte-rendu-e-newspaperarchivesch.

Neudecker, Clemens, and Apostolos Antonacopoulos. 2016. “Making Europe’s Historical Newspapers Searchable.” In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 405–10. Santorini, Greece: IEEE. https://doi.org/10.1109/DAS.2016.83.

Nicholson, Bob. 2015. “Review of ‘Europeana Newspapers.’” *Reviews in History* review no. 1894. https://doi.org/10.14296/RiH/2014/1894.

Oberbichler, Sarah, Stefan Hechl, Barbara Klaus, Minna Kaukonen, Tuula Pääkkönen, and Marion Ansel. 2019. “Online Research of Digital Newspapers of Three National Libraries: A Survey.” 2019. https://www.newseye.eu/blog/news/online-research-of-digital-newspapers-of-three-national-libraries-a-survey-by-sarah-oberbichler-stef/.

Putnam, Lara. 2016. “The Transnational and the Text-Searchable: Digitized Sources and the Shadows They CastThe Transnational and the Text-Searchable.” *The American Historical Review* 121 (2): 377–402. https://doi.org/10.1093/ahr/121.2.377.

Rautiainen, Juha. 2016. “Getting to Know Users of Digital Newspaper and Journal Library – What Can Statistics of Use Tell Us.” 2016. http://blogs.sub.uni-hamburg.de/ifla-newsmedia/wp-content/uploads/2016/04/Rautiainen-Getting-to-Know-Users-of-Digital-Newspaper-and-Journal-Library.pdf.

Stroeker, Natasha, and René Vogels. 2012. “Survey Report on Digitisation in European Cultural Heritage Institutions 2012.” *Brussels: ENUMERATE Thematic Network, Available Online at Http://Www. Enumerate. Eu/Fileadmin/ENUMERATE/Documents/ENUMERATE-Digitisation-Survey-2012.pdf [Accessed April 20th 2013]*.

Sumikawa, Yasunobu, Adam Jatowt, Antoine Doucet, and Jean-Phillippe Moreux. 2019. “Large Scale Analysis of Semantic and Temporal Aspects in Cultural Heritage Collection’s Search.” In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 77–86. Urbana-Champaign, Illinois: IEEE. https://doi.org/10.5281/zenodo.3243337.

Whitelaw, Mitchell. 2015. “Generous Interfaces for Digital Cultural Collections” 9 (1). http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html.

Zaagsma, Gerben. 2019. “Digital History and the Politics of Digitisation (Presentation DH2019). https://dev.clariah.nl/files/dh2019/boa/0758.html.

Annex A – List of reviewed interface features

Property		High-level criteria*						Property		High-level criteria*					
(a) About the interface		1	2	3	4	5	6	***Newspaper metadata (cont'd)***		1	2	3	4	5	6
1	URL							19	Publisher	✓					
2	Target area							20	Date range	✓					
3	Creator							21	Frequency (i.e. periodicity)	✓					
4	Purpose and scope							22	ISSN, OCLC, LCCN	✓					
5	Approximate date of creation							23	External links						✓
6	Interface is multilingual							24	Description of newspaper (historical)	✓					
7	Access model							25	Language	✓					
8	Interface provider							26	Calendar view of issues	✓					
(b) Information about the newspaper collection								27	Indication of archive holder	✓					
9	Number of newspaper titles	✓						***(d) Browsing***							
10	Number of issues	✓						28	By date			✓			
11	Number of pages	✓						29	By title			✓			
12	Number of articles	✓						30	By place of publication			✓			
13	Indication of the original digitized issue	✓						31	By user tag			✓			
14	New titles continuously added							32	By newspaper theme (metadata)			✓			
15	Languages of the collections							***(e) Search options***							
(c) Newspaper metadata								33	Basic keyword search		✓				
16	Alternative titles, succeeding titles, related titles	✓						34	Query autocomplete				✓		
17	Place of publication	✓						35	Boolean operators (AND, OR, NOT)		✓				
18	Geographic coverage	✓						36	Phrase search		✓				

*High-level criteria: (1) source criticism, (2) content search, (3) content filtering, (4) generosity, (5) user content management and exploration, (6) connectivity.

Property	**High-level criteria**						**Property**	**High-level criteria***					
Search options (cont'd)	1	2	3	4	5	6	***Search result display (cont'd)***	1	2	3	4	5	6
37 Fuzzy search		✓					57 Keyword highlight in facsimiles		✓				
38 Wild card		✓					58 Keyword highlight in OCRed text		✓				
39 Proximity search (near operator)		✓					59 N-grams				✓		
40 Limit the date range		✓					***(g) Search result sorting***						
41 Limit by language		✓					60 By relevance			✓			
42 Limit by newspaper title(s)		✓					61 By date			✓			
43 Limit by place of publication		✓					62 By newspaper title			✓			
44 Limit by newspaper theme (from metadata)		✓					63 By article title			✓			
45 Limit by newspaper segments / zones		✓					64 By content type (ad, article, illustration)			✓			
46 Limit by article category		✓					65 By online publication date			✓			
47 Limit by article length		✓					66 By author			✓			
48 Limit by archive holder / library		✓					67 By quality of text			✓			
49 Limit by license / accessibility		✓					68 By language			✓			
50 Query suggestion				✓			69 By popularity (number of views)			✓			
51 Search by named entities		✓					***(h) Search result filters***						
(f) Search result display							70 By newspaper title			✓			
52 Distribution over time				✓			71 By publishing frequency			✓			
53 Distribution by place of publication				✓			72 By political/religious/etc. orientation of np			✓			
54 Distribution by newspaper coverage				✓			73 By newspaper theme (metadata)			✓			
55 Distribution by place names in articles				✓			74 By content type			✓			
56 Snippet preview (OCR and/or image)	✓						75 By section ("rubrique")			✓			

*High-level criteria: (1) source criticism, (2) content search, (3) content filtering, (4) generosity, (5) user content management and exploration, (6) connectivity.

Property	High-level criteria*						Property	High-level criteria*					
Search result filters (cont'd)	1	2	3	4	5	6	***Viewer (cont'd)***	1	2	3	4	5	6
76 By event			✓				98 Option to continue to next result		✓				
77 By person			✓				***(j) Personal account and user interactions***						
78 By organization			✓				99 Save articles to favorites					✓	
79 By place mentioned in the text			✓				100 Save queries to favorites					✓	
80 By time period			✓				101 Tag articles					✓	
81 By topic			✓				102 Keep track of viewed materials					✓	
82 By manual tag			✓				103 Article recommendations					✓	
83 By place of publication			✓				104 Permalinks					✓	
84 By publisher			✓				105 Export citation					✓	
85 By article length			✓				106 Option to correct OCR					✓	
86 By authors			✓				107 Option to correct OLR					✓	
87 By segmentation level			✓				108 Users can add/edit metadata					✓	
88 By language			✓				109 Screenshot tool					✓	
89 By license			✓				110 Download options (file formats)					✓	
90 By online publication date			✓				111 Bulk downloads					✓	
(i) Viewer							112 Organize articles in collections					✓	
91 Facsimile displayed	✓						113 Contrastive view of personal collections					✓	
92 OCRed text display	✓						***(k) Connectivity***						
93 Show full width/height (=full page)	✓						114 Third party identifiers (e.g. VIAF)						✓
94 Interactive mini-map of page	✓						115 Links to other repositories						✓
95 Overview of available issues	✓						116 Semantic web technologies						✓
96 Search in viewed page		✓											
97 Option to continue to next page				✓									

*High-level criteria: (1) source criticism, (2) content search, (3) content filtering, (4) generosity, (5) user content management and exploration, (6) connectivity.

Property		High-level criteria*						Property		High-level criteria*					
(l) Documentation on digitization		1	2	3	4	5	6	***(n) Content processing***		1	2	3	4	5	6
117	Document layout analysis at article level	✓						131	NERC		✓				
118	Document layout analysis confidence scores	✓						132	Entity linking		✓				
119	Document layout analysis confidence scores	✓						133	Post-OCR correction		✓				
120	Documentation of biases and shortcomings	✓						134	Topic modeling		✓				
121	Search result relevance score	✓						135	Text re-use		✓				
122	Digitization date at title level	✓						136	Sentiment analysis		✓				
123	Scan resolution (in dpi)	✓						137	Query		✓				
124	Information on OCR tools used	✓						138	Recommendations				✓		
125	Copyright notice	✓						139	Event detection		✓				
126	Documentation of scan methods	✓													
(m) APIs and code															
127	Link to source code of the interface	✓													
128	API	✓													
129	IIIF Image API	✓													
130	IIIF Presentation API	✓													

*High-level criteria: (1) source criticism, (2) content search, (3) content filtering, (4) generosity, (5) user content management and exploration, (6) connectivity.

Table 2: Overview of features, feature families and their mapping to assessment criteria (numbers 1-6).