Navigating through 200 Years of Historical Newspapers*

Yannick Rochat
Digital Humanities Laboratory
CH-1015 Lausanne
yannick.rochat@epfl.ch

Maud Ehrmann
Digital Humanities Laboratory
CH-1015 Lausanne
maud.ehrmann@epfl.ch

Vincent Buntinx
Digital Humanities Laboratory
CH-1015 Lausanne
vincent.buntinx@epfl.ch

Cyril Bornet
Digital Humanities Laboratory
CH-1015 Lausanne
cyril.bornet@epfl.ch

Frédéric Kaplan
Digital Humanities Laboratory
CH-1015 Lausanne
frederic.kaplan@epfl.ch

ABSTRACT

This paper describes the processes which led to the creation of an innovative interface to access a digital archive composed of two Swiss newspapers, namely *Gazette de Lausanne* (1798–1998) and *Journal de Genève* (1826–1998). Based on several textual processing steps, including lexical indexation, n-grams computation and named entity recognition, a general purpose web-based application was designed and implemented; it allows a large variety of users (e.g. historians, journalists, linguists and the general public) to explore different facets of about 4 million press articles spanning an almost 200 hundred years period.

Keywords

Digital humanities, historical newspapers, innovative interface, language evolution, named entity recognition

1. INTRODUCTION

Newspapers are essential sources for the exploration of the past [4]. From a historical point of view, they document aspects and events of our societies from the perspective of contemporary actors and, from a linguistic point of view, they constitute (once digitized) large corpora that can be used to e.g. investigate the evolution of language(s). Both researchers and the general public benefit from online access to cultural heritages such as newspaper archives [15].

Many newspapers digitisation projects [1] have been realised in the last ten years [16, 22] thanks to the facilitated acquisition of larger storage amenities and higher computing power. Most projects provide access to the scanned documents but do not offer more than basic search through the textual content.

In Switzerland, the Swiss National Library has contributed to the digitisation of more than thirty newspapers. [2] The library centralises some of these projects [3], while others are hosted by public or private partners. [4]

---

*Supported by the Swiss National Library.
 1. http://en.wikipedia.org/wiki/Wikipedia:List_of_online_newspaper_archives Accessed on April 24th, 2016.
 2. http://www.nb.admin.ch/themen/02074/02076/03887/?lang=en Accessed on April 24th, 2016.
 3. http://newspaper.archives.rero.ch/ Accessed on April 24th, 2016.
 4. http://www.nb.admin.ch/public/04506/04514/index.html?lang=en Accessed on April 24th, 2016.

In 2008, all original issues of the three journals composing the archives of *Le Temps* [5]–*Gazette de Lausanne, Journal de Genève* and *Le Nouveau Quotidien* [6] (1991–1998)–were digitised and made available for consultation to the public through a website. [7] Texts have been extracted from scanned pages using optical character recognition (OCR) and layout detection algorithms, allowing visitors to search through a corpus composed of close to 1 million pages and 4 million articles [8], covering 200 years of local, national and global news as seen from the French part of Switzerland. [9]

This article describes a web application offering a new interface to navigate this 200 year corpus. It was developed as part of a collaboration between *Le Temps*, the Swiss National Library and the Digital Humanities Laboratory of the Swiss Federal Institute of Technology of Lausanne (EPFL). Features formerly available like lexical search, editable time intervals or the possibility to look for a given issue based on the date were implemented. An image viewer that situates articles in their original contexts was developed allowing to browse full newspaper issues from the first to the last page without leaving the interface. Each page can be zoomed into up to a level allowing to see small details of graphics or comfortable on-screen reading. In addition, two methods stemming from natural language research to improve the navigation in the corpus, namely n-grams viewing and named entities, were adapted.

The remainder of the paper is organised as follows. In section 2 we describe the corpus composed of the two main newspapers, *Gazette de Lausanne* and *Journal de Genève*. Next, we present the text processes applied on the digital archive with the computation of n-grams (section 3) and the recognition of named entities (section 4). In section 5 we detail theoretical and technical aspects of the public interface and finally we conclude and consider future work in section 6.

2. LE TEMPS CORPORA

In this section, we present a few quantitative descriptors for this corpora (publication frequency, statistics of words and pages), then we display front pages for key moments

---

 5. A Swiss newspaper launched in 1998.
 6. At the time of writing, the inclusion of *Le Nouveau Quotidien* in the new website is ongoing.
 7. It will be removed in the future. At the time of writing, it is accessible at old.letempsarchives.ch
 8. Including images with captions, and advertisements.
 9. These newspapers were written in French.

in the history of these newspapers and sketch their stylistic evolution over time. Eventually, we discuss the encoding of the data.

2.1 General Statistics

Gazette de Lausanne and *Journal de Genève* reached regular and similar publication frequencies in the 1850s. Before that time, the situation was less harmonious. *Gazette de Lausanne* appeared rather regularly, around 100 times a year from 1804 to 1846 (see figure 1) while the number of issues per year of *Journal de Genève* varied from 52 issues (1828) to 246 issues (1834) between 1826 and 1850 (see figure 2).
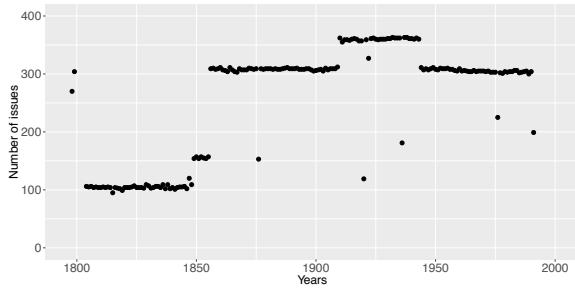
Figure 1: The number of issues per year of *Gazette de Lausanne.*
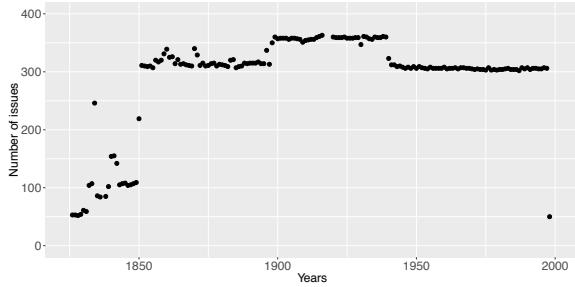
Figure 2: The number of issues per year of *Journal de Genève.*

There are eight outlying years in our dataset for *Gazette de Lausanne* (described in table 1, and no equivalent for *Journal de Genève*. With the exception of years 1798 and 1799, which are composed of issues from *Gazette de Lausanne*'s ancestors, it appears that these outliers are mostly years with missing data inherited from the original data set.[10] The task of retrieving the parts currently lacking is ongoing.

In our corpus, there are in total 441′579 printed pages for *Gazette de Lausanne* from 1798 to 1998[11] and 495′986 for *Journal de Genève* from 1826 to 1998. In addition, figures 3

---

10. For example, all issues are missing : from 1800 to 1803, from July 1876 to December 1876, from May 1920 to December 1920, from January to June 1936.
11. During years 1991 to 1998, the two newspapers were merged into a single one whose name was *Journal de Genève et Gazette de Lausanne* (see figure 16).

Table 1: Outlying years from figure 1.

Year	# of published issues
1798	270
1799	304
1876	153
1920	119
1922	320
1936	181
1976	225
1991	199

and 4 show the average number of pages per issue for these two newspapers. With exception of the very first years and 1830s for *Gazette de Lausanne*, both newspapers were printed on 4 pages (one large sheet of paper, folded) until 1900s for *Journal de Genève* and 1940s for *Gazette de Lausanne*. Then the number climbed with a slowing down in the 1970s for both newspapers.
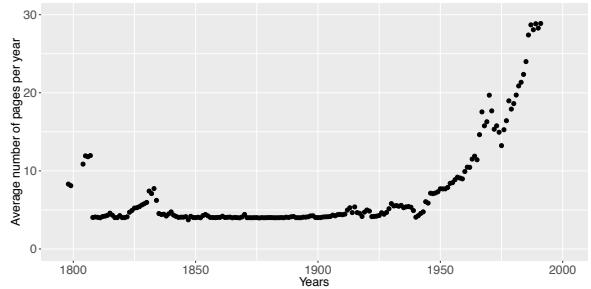
Figure 3: Average number of pages by issue, per year, in *Gazette de Lausanne.*
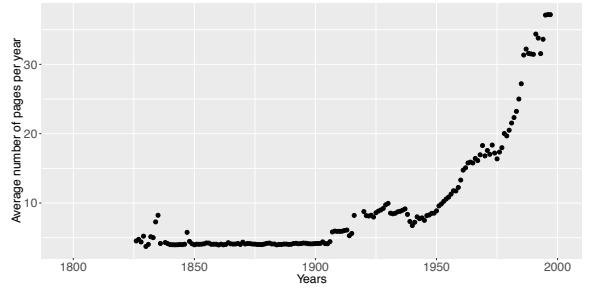
Figure 4: Average number of pages by issue, per year, in *Journal de Genève.*

Selected front pages from *Gazette de Lausanne* and *Journal de Genève* at births (see figures 12, 13 and 14) and deaths of these journals (see figures 15 and 16) are shown in the appendix.

2.2 Encoding

The whole archive, including text and images, weighs 22 terabytes. The structure of each newspaper issue is encoded in a master XML file which lists articles [12] along with metadata information (e.g. article boxes and ordering, used font, etc.). In turn, the content of each article, that is to say words and their positions, is encoded in a proper XML file. Besides the XML text material, each page image is saved in TIFF format and all pages of an issue are saved in a single PDF containing the OCRed text.

The quality of the OCR has not been evaluated at this stage, but some mistakes are immediately noticeable, mostly due to bad conservation of paper, to the digitisation process (transparency, creases, stains), and to ink drips at the time of printing, all common phenomena in this type of project.

3. N-GRAMS

An *n-gram* is an ordered sequence of *n* consecutive words. For instance, given the phrase *La Gazette de Lausanne*, *La Gazette*, *Gazette de* and *de Lausanne* are 2-grams, whereas each word taken separately is a 1-gram.

N-grams were extracted from the XML text files, by considering alphanumeric tokens only. In order to compute the n-grams relative frequencies, we chose a time granularity of one year and divided the number of occurrences of each n-gram by the total number of n-grams occurrences for the same period. Obviously, as *n* increases, so does the likeliness of any n-gram to be unique, and thus the number of distinct n-grams converges towards the number of words in the corpus. For that reason, storing the n-grams becomes rapidly costly in terms of volume for large values of *n*.

Visualising n-grams frequency distributions on a given corpus allows to test hypotheses about linguistic and sociolinguistic evolutions, as preceding works demonstrated [18, 23]. In order to help users gather knowledge for a given query on the whole corpora, a viewer allowing to display the variations of n-grams relative frequencies over time was created. Examples of this n-gram viewer can be seen in figures 5 and 6. N-grams distributions are influenced by linguistic and socio-cultural factors, but also by constraints related to the journals themselves (e.g. the diversity of covered topics, article sizes, etc.). As an example, the behaviour of the n-gram *1914* is greatly impacted by the First World War, which is not a linguistic factor. On the other hand, the lexical diversity might be influenced by the length of articles as well as linguistic evolution. All these factors contribute, in different proportions, to the n-grams frequencies evolution and have to be considered together.

Uses of n-grams are manifold. From a set of n-grams distributed over time, we can extract linguistic, semantic and sociocultural information. Several researches are currently underway and use the extraction of absolute and relative frequencies of n-grams. For example, a study explored the different typologies of n-grams curves identifying core processes and classifying n-grams in these archetypical categories [6]. This study considered the question of reversing the n-gram viewer paradigm, searching in the space of n-grams frequencies curves instead of searching in the space of n-grams. Another study defined the notion of n-gram cores and resilience, allowing to compare corpora and study linguistic


**Figure 5: Visualisation of 1-grams "russie" (Russia) and "urss" (USSR).**
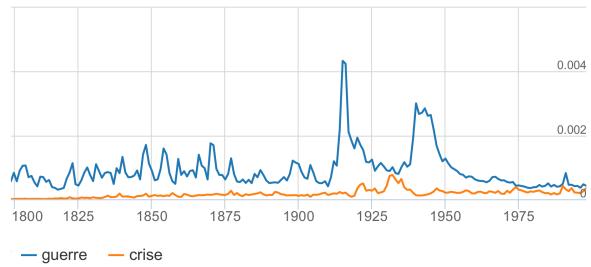
Figure 6: Visualisation of 1-grams "guerre" (war) and "crise" (crisis).

evolution through the concept of words resilience instead of linguistic changes [5].

4. NAMED ENTITIES

Recognition and processing of real-world entities is essential for enabling effective text mining. Indeed, referential units such as names of persons, organisations and locations underlie the semantics of texts and guide their interpretation. Known as named entities (NE), these units are major bearers of information and can help answering the questions of *Who did What to Whom, Where and When*? (known as the *5Ws* in journalism). First introduced during the 6th Message Understanding Conference [12], named entity processing have evolved significantly over the last two decades, from entity recognition and classification to entity disambiguation and linking [13] [10, 21]. More recently, NE processing has been called upon to contribute to the research area of Digital Humanities where algorithms have to deal with OCRed documents [25, 26] and languages and documents of earlier stages [7, 11, 30].

In the context of designing and developing a new interface to enable users to search through two of the newspapers composing *Le Temps* archive, implementing a named entity recognition system appeared as an obvious desideratum. Although many NE processing tools are now available almost "off-the-shelf", they can hardly be applied on *Le Temps* documents for various reasons. Tools developed

---

12. The word "article" represents articles, images and advertisements.

---

13. Entity linking corresponds to the task of linking an entity metnion to a unique identifier in a knowledge base, e.g. DBpedia.

by private companies (e.g. Open Calais [14], Zemanta [15], Alchemy [16]) are most of the time intended for English language and, when available for French, are only accessible through limited web services–a framework unsuitable when analysing millions of documents. Moreover, APIs and tag sets (i.e. named entity categories) of those tools are regularly updated, which results in undesirable maintenance problems. On the academic side, various entity linking tools are being developed by the Natural Language Processing and Semantic Web communities. DBpedia Spotlight [8, 17], AIDA [32] and BabelFy [20] are dedicated to the spotting of entity mentions in texts and their linking to entities stored in knowledge bases (KBs). If they are able to assign referents to entities in text (i.e. entity disambiguation), these tools do not however perform real named entity recognition in the sense that they can only spot names of entities which are present in the given KB. Besides, background KBs are for the most part derived from Wikipedia and thereby contain primarily VIPs, which is unsuitable for recognising the *John Doe(s)* of past and present days from *Le Temps* collection. Finally, those tools are well developed and maintained for English language; it is possible to deploy them on new languages but it requires a huge effort for a result which might not meet all needs.

Without discarding the option of using one of these tools at a later stage, as of now we sought a solution able to (1) parse French language, (2) recognise all entity mentions, and (3) be executed offline. To this end, we used a rule-based system using the ExPRESS formalism [24] such as deployed by the *Europe Media Monitor* (EMM) [29] for multilingual NER [28]. ExPRESS is an extraction pattern engine based on finite state automata. It allows to define rules or patterns which, coupled with appropriate lexical resources and preprocessing (tokenization and sentence splitting), can detect and type specific phrases in texts. Named entity recognition is implemented via a cascade of grammar files where units are detected and processed in increasing order of complexity. In concrete terms, NE rules focus on typical patterns of person, location and organisation names, e.g. an adjective (*former*) followed by a function name (*President of the Confederation*), a first (*Ruth*) and a last (*Dreifuss*) name. Units such as *former* and *President* are called trigger words; besides modifiers and function names they cover professions (*guitarist, football player*), demonyms and markers of religion or ethnical groups (*Italian, Genevan, Bambara, Muslim*), expressions indicating age (*42 years-old*), and more. It is worth noticing that this system performs named entity recognition and classification but not disambiguation.

We applied our named entity grammars [17] on articles of *Le Temps* archive for the recognition of Person and Location names (we reserve the Organisation type for future work). In order to speed up the process and ease the debugging, we executed our process in parallel on a very powerful computing node (48-core, 256GB of RAM). Parsing of all files took a couple of hours. In order to allow maximum flexibility with the usage of data, processing results are first stored in JSON [18] format. They are afterwards converted in the Resource Description Framework (RDF) so as to allow final

data publication as Linked Data [2, 13]. The ontology used to represent extracted entities revolves around two core elements, *Article* and *EntityMention*, each one being further qualified with specific properties. We made use of classes and properties defined by the Dublin Core terms [19], NIF [20], OLiA [21] and LexInfo [22] vocabularies. The RDF graph is loaded on a triple store (Virtuoso open source) whose SPARQL endpoint is available from the interface, as we shall see in the next section. Users can access about 30 million entity mentions of type Location and 20 million of type Person. Thanks to the extraction of detailed information along with person names and to their RDF representation, it is possible to explore various dimensions of person entities. Examples of queries against the data set include :

— all person mentions having a specific function (e.g. *German Chancellor*) in articles issued between date x and date y ;
— all functions of a specific person mention ordered chronologically, with the possibility to get the source articles ;
— all articles mentioning conjointly 2 or more specific person mentions ;
— all person mentions which occur with a specific title or function ;
— etc.

Future developments regarding this text processing module involve NER evaluation, processing of Organisation entities and entity disambiguation.

5. WEB APPLICATION

5.1 Interface Principles

The interface design we addressed typically falls under a lack of known or typical use cases. With any website, we expect the base of users and their expectations to be very wide and diverse. Regarding the old website, the only statistic we could use would have been the user search history. However, this tells little about their intents and we ignore if the information they found was relevant to them.

We thus needed to define a set of basic requirements that would follow the most generic possible use cases, yet providing modern and powerful features to journalists, historians and information scientists. The core features that were outlined by preceding studies on similar archives are the following :

— A global, full-text, high performance search engine is generally the preferred way to access information, both to novice and expert users [9]. The added value of finding aids such as advanced search options or a hierarchical organisation is however subject to debate [31].
— Articles should always be read in their full publication context [4].
— Each page needs to be easily referenced by an unique URL, so it can be quickly stored for later access in a situation of information gathering [1].

In addition to the search engine, we needed to come up with an appealing way to browse search results. Unfortunately, no relevance score can easily be derived from the way

contents are organised, as there are no links between articles that may allow to guess their relative importance, nor a clear way to predict which kind of content might be interesting. To answer this question, we thus introduced the n-gram visualisation as a very part of the search results. In this manner, results of search queries consist of, first, the n-gram viewer featuring the evolution of query term ies over time on both journals and, second, snippet previews of retrieved articles. The n-gram viewer allows users to get a quick hint at periods of interest and to select more precise time frames to dig for interesting results. Figure 7 shows how typical search results are presented.
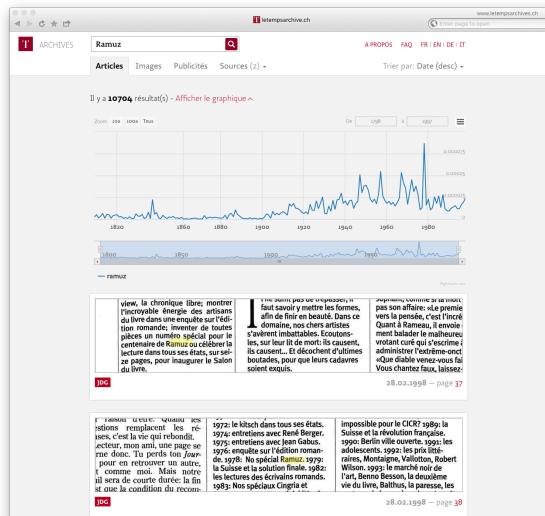
Figure 7: Search results. At the top, period selection using the n-gram viewer. At the bottom, previews from the found articles.

The necessity of viewing full pages with an adequate resolution called for a tailored solution. The technical requirements are as follows :

— The search engine results need to access previews that can be anywhere in the pages, typically showing the found word(s) in a context of a couple sentences.
— The high quality scans of the full pages are too big to be loaded as they are [23], yet we need to be able to present them as a whole to the user and to enlarge the relevant parts, possibly up to the highest definition available.
— All the images sent out to the client must be optimised to keep low loading times and acceptable server loads.

Those can be addressed in a nice way using a web image server supporting multiple image formats (raw files for archiving and preservation and optimised ones for web delivery) and tiling. We selected an image server responding to the International Image Interoperability Framework (*IIIF*) norm, for its outstanding interoperability and academic approach [27]. Figures 8 and 9 illustrate the use of the viewing interface.

---

23. A double-page typically weighs about 10 MB.
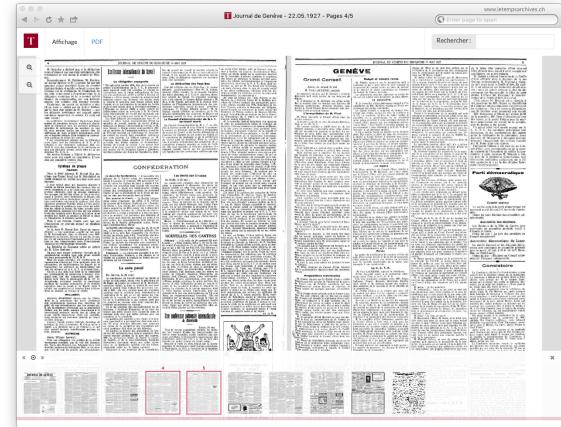
Figure 8: Viewer interface (featuring one full double-page). At the bottom, previews of all pages from the same issue.
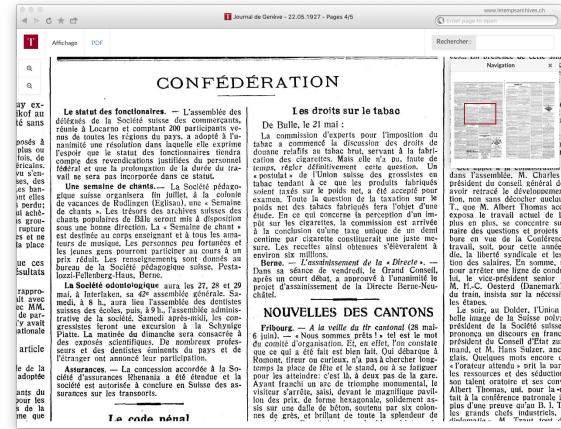
Figure 9: Viewer interface (zooming on one article). In the top right corner, the location of the article in the double page is highlighted.

The named entity query engine features a simple interface, in the form of a SPARQL endpoint, showed in figure 10. In order to make it more accessible to non-technical users, we included 5 sample queries that can be tried out with a simple click.

5.2 Application Stack Design

The software setup we decided on is as follows :

— Raw text indexation and search : Apache Solr. [24]
— Image Server : Loris IIIF. [25]
— Web development frameworks : Laravel. [26]

---

24. http://lucene.apache.org/solr
25. http://github.com/loris-imageserver/loris
26. http://laravel.com

Figure 10: SPARQL endpoint presenting the results of a sample query.

— Internal database engine : PostgreSQL. [27]
— Triplestore : Virtuoso Open Source. [28]

Figure 11 shows the organisation of the different components. The typical web client issues a search request (A) that the web application forwards to the search engine (B) to find out the relevant pages, and to the internal database to load the necessary metadatas (C). Alternatively (in response to a SPARQL query), it will load data from the triplestore database (D). It then returns an HTML page (E) including URLs to the images that will be provided by the image server (F). Finally, new journal issues may be added to the archive using a publication workflow (G) that extracts image and textual representations from the scans.
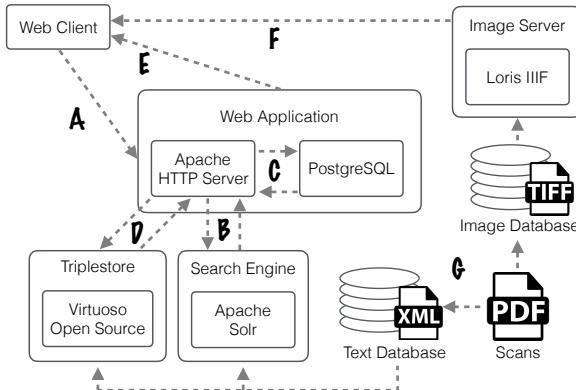
Figure 11: Information workflow and technical components.

5.3 Public Release

The final application has been released to the public on a dedicated web server running the full software stack described earlier. For improved performance, the website is cached using *Cloudflare* services. [29]

During the first month (March 18th to April 20th, 2016), about 35'000 search queries were made (hence more than 1'000 a day). Out of those, 2200 were direct accesses to specific dates, and the rest represented 21'350 unique words. According to Google Analytics [30], the new site was seen by 18'300 people, out of which more than 90% accessed it at least twice.

6. CONCLUSIONS

The new website and tools immediately received significant interest from researchers of several Swiss universities and state libraries. We received many constructive feedback, and answered questions from users having long-time use cases they needed to reproduce with the new web application.

Consultation statistics demonstrated great enthusiasm from the general public. On the day of the public launch, *Le Temps* published a dedicated article [31] and included a four-pages insert mainly composed of archival articles. Building on the launch, third parties also opened a *Facebook* page [32] to discuss noteworthy findings in the archives such as century old discussions relevant to current events or advertisements seen as comical from today's perspective.

Future works will focus on updating the contents and refining our tools to provide access to a wider range and even more relevant data depending on queries from users. Several improvement techniques have already been considered and are on their way :

— Improvement of raw data with a set of tools aiming to correct the OCR results, especially for the earlier years. Multiple approaches are possible including the use of language models [3], semi-automated statistical correction and crowdsourcing [14].
— Named entity disambiguation. In use, this allows the user to filter results that relate to different entities sharing the same names.
— Completion of the corpus with the missing journal issues, wherever possible.
— Find new partners and add new collections.

7. ACKNOWLEDGEMENTS

8. REFERENCES

[1] S. Attfield and J. Dowell. Information seeking and use by newspaper journalists. *Journal of Documentation*, 59(2) :187–204, Apr. 2003.

[2] S. Auer, J. Lehmann, A.-C. Ngonga Ngomo, and A. Zaveri. Introduction to Linked Data and Its Lifecycle on the Web. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, S. Rudolph, G. Gottlob, I. Horrocks, and F. van Harmelen, editors, *Reasoning Web. Semantic Technologies for Intelligent Data Access*, pages 1–90. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[3] A. Bhardwaj, F. Farooq, H. Cao, and V. Govindaraju. Topic based language models for OCR correction. In *Proceedings of the SIGIR 2008 Workshop on Analytics for Noisy Unstructured Text Data*, pages 107–112. ACM Press, 2008.

[4] A. Bingham. 'The Digitization of Newspaper Archives : Opportunities and Challenges for Historians'. *Twentieth Century British History*, 21(2) :225–231, June 2010.

[5] V. Buntinx, C. Bornet, and F. Kaplan. Studying linguistic changes on 200 years of newspapers. In *DH2016 - Annual Conference of the Alliance of Digital Humanities Organizations*, 2016.

[6] V. Buntinx and F. Kaplan. Inversed N-gram viewer : Searching the space of word temporal profiles. In *DH2015 - Annual Conference of the Alliance of Digital Humanities Organizations*, 2015.

[7] K. Byrne. Nested named entity recognition in historical archive text. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pages 589–596. IEEE, 2007.

[8] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 121–124. ACM, 2013.

[9] M. Daniels and E. Yakel. Seek and You May Find : Successful Search in Online Finding Aid Systems. *The American Archivist*, 73(2) :535–568, Sept. 2010.

[10] M. Ehrmann, D. Nouvel, and S. Rosset. Named Entities Resources - Overview and Outlook. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'10)*, Portorož, Slovenia, 2016 (to appear).

[11] F. Frontini, C. Brando, and J.-G. Ganascia. Semantic Web based Named Entity Linking for digital humanities and heritage texts. pages 1–12, 2015.

[12] R. Grishman and B. Sundheim. Design of the MUC-6 evaluation. In *Sixth Message Understanding Conference (MUC-6) : Proceedings of a Conference Held in Columbia, Maryland*, 1995.

[13] T. Heath and C. Bizer. Linked data : Evolving the web into a global data space. *Synthesis lectures on the semantic web : theory and technology*, 1(1) :1–136, 2011.

[14] R. Holley. Crowdsourcing : how and why should libraries do it ? *D-Lib Magazine*, 16(3) :4, 2010.

[15] L. James-Gilboe. The challenge of digitization : Libraries are finding that newspaper projects are not for the faint of heart. *The Serials Librarian*, 49(1-2) :155–163, 2005.

[16] E. Klijn. The current state-of-art in newspaper digitization : A market perspective. *D-Lib Magazine*, 14(1) :5, 2008.

[17] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia spotlight : shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM, 2011.

[18] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, , J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 2010.

[19] L. Monnet. La gazette de lausanne à l'origine. *Le Conteur vaudois*, 39(32) :1, 1901.

[20] A. Moro, A. Raganato, and R. Navigli. Entity Linking meets Word Sense Disambiguation : a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2 :231–244, 2014.

[21] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1) :3–26, 2007.

[22] C. Neudecker and A. Antonacopoulos. Making Europe's historical newspapers searchable. In *Proceedings of the 12th IAPR International Workshop on Document Analysis Systems (DAS2016)*, 2016.

[23] M. Perc. Evolution of the most common English words and phrases over the centuries. *Journal of The Royal Society Interface*, 9(77) :3323–3328, Dec. 2012.

[24] J. Piskorski. ExPRESS – Extraction Pattern Recognition Engine and Specification Suite. In *Proceedings of the International Workshop Finite-State Methods and Natural Language Processing 2007 (FSMNLP 2007)*, Potsdam, Germany, September 2007.

[25] K. Rodriquez, M. Bryant, T. Blanke, and M. Luszczynska. Comparison of named entity recognition tools for raw OCR text. In *KONVENS*, pages 410–414, 2012.

[26] S. Rosset, C. Grouin, K. Fort, O. Galibert, J. Kahn, and P. Zweigenbaum. Structured named entities in two distinct press corpora : Contemporary broadcast news and old newspapers. In *Proceedings. of the 6th Linguistic Annotation Workshop*, pages 40–48. Association for Computational Linguistics, 2012.

[27] R. S. Snydman, Stuart and T. Cramer. The international image interoperability framework (iiif) : A community & technology approach for web-based images. *Stanford University Libraries staff publications and research*, 2015.

[28] R. Steinberger, B. Pouliquen, M. Kabadjov, and E. van der Goot. JRC-Names : A Freely Available, Highly Multilingual Named Entity Resource. In *Proc. of the 8th International Conference Recent Advances in Natural Language Processing (RANLP'2011)*, Hissar, Bulgaria, September 2011.

[29] R. Steinberger, B. Pouliquen, and E. van der Goot. An introduction to the europe media monitor family of applications. In . J. K. F. Gey, N. Kando, editor,

*Information access in a multilingual world —
Proceedings of the SIGIR 2009 Workshop
(SIGIR-CLIR)*, Boston, USA, July 2009.

[30] S. van Hooland, M. De Wilde, R. Verborgh,
T. Steiner, and R. Van de Walle. Exploring entity
recognition and disambiguation for cultural heritage
collections. *Digital Scholarship in the Humanities*,
30(2) :262–279, 2015.

[31] E. Yakel. Encoded archival description : Are finding
aids boundary spanners or barriers for users ? *Journal
of Archival Organization*, 2(1-2) :63–77, June 2004.

[32] M. Yosef, J. Hoffart, I. Bordino, M. Spaniol, and
G. Weikum. Aida : An online tool for accurate
disambiguation of named entities in text and tables.
In *Proceedings of the 37th International Conference on
Very Large Databases 9th*, page 1450–1453, Seattle,
USA, 2011.

APPENDIX

**Figure 12: On February 1st, 1798, the front page
of the first issue of what would later become *Ga-
zette de Lausanne* after bearing nine other names
[19]. "Bulletin officiel" approximately means "Offi-
cial news report". This page shows letters s printed
as f's.**
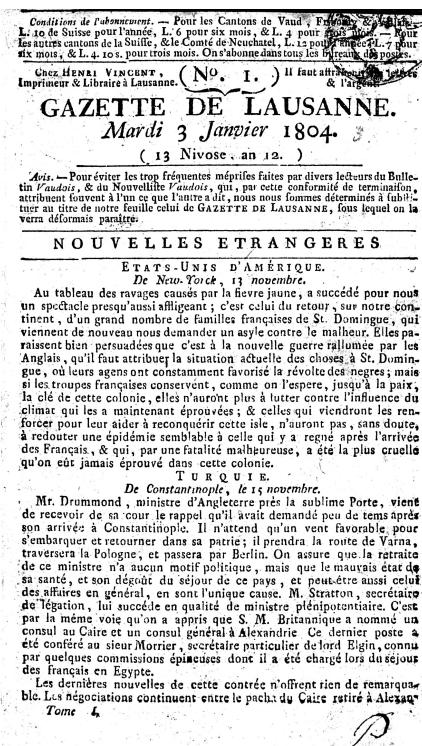
Figure 13: On January 3rd, 1804, few years after its creation, *Gazette de Lausanne* receives a name it kept for close to two centuries.
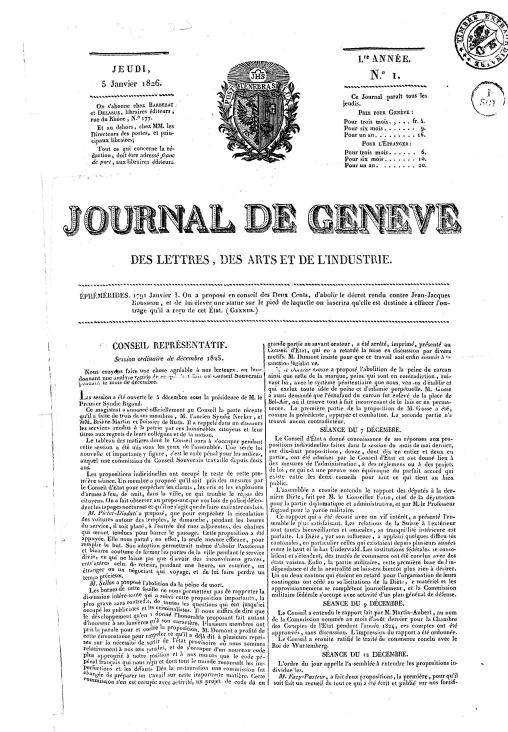
Figure 14: *Journal de Genève* was launched on January 5th, 1826.

Figure 15: (Top.) On August 31st, 1991, a discreet insert at the bottom right corner of the front page announces that the two newspapers are merged into a single one. (Bottom.) On September 2nd, 1991, the result of the merged newspapers is published under the name *Journal de Genève et Gazette de Lausanne*.

Figure 16: February 28th, 1998. The final issue of *Journal de Genève et Gazette de Lausanne*. It would then be merged with *Le Nouveau Quotidien* in order to form *Le Temps*, which was first issued on March 18th, 1998.