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 aims to describe and explain the processes behind the creation of a digital library composed of two Swiss newspapers, namely *Gazette de Lausanne* (1798–1998) and *Journal de Genève* (1826–1998), covering an almost two-century period. We developed a general purpose application giving access to this cultural heritage asset ; a large variety of users (e.g. historians, journalists, linguists and the general public) can search through the content of around 4 million articles via an innovative interface. Moreover, users are offered different strategies to navigate through the collection : lexical and temporal lookup, n-gram viewer and named entities.

CCS Concepts

•Information systems → Digital libraries and archives; •Applied computing → Arts and humanities;

Keywords

Digital humanities, historical newspapers, innovative interfaces, language evolution, named entities recognition

1. INTRODUCTION

Newspapers are essential sources in the exploration of the past [4]. On the historical side, they document aspects and events of our societies from the point of view of contemporary actors, while on the linguistic side they provide, once digitised, large corpora to researchers. 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].

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]. All words in the texts were recognised using optical character recognition (OCR) and layout detection algorithms, allowing visitors to search through a corpus comprised 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 in 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 the full newspaper issue 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 . These two dimensions will be the core to research based on this corpus in the near future.

*Supported by the Swiss National Library.

1. http://en.wikipedia.org/wiki/Wikipedia:List_of_online_newspaper_archives Accessed on April 24th, 2016.

??????

© 2016 ACM. ISBN ???...$???

DOI: ???

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.

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 following sections, firstly we describe the corpus composed by the two main newspapers, *Gazette de Lausanne* and *Journal de Genève*, with some context and statistics about the numbers of words and pages and the frequencies at which the newspapers were published. Then, we discuss the elements of text processing that we included into this digital library : the n-gram viewer and the named entities search engine. The final section presents theoretical and technical aspects of the public interface and selected software components.

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 in the history of these newspapers and sketching their stylistic evolution with 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*.

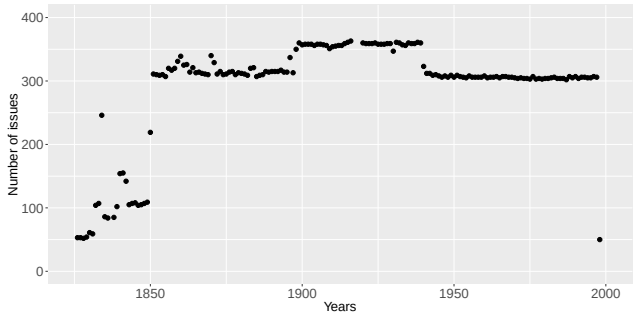


Figure 2: The number of issues per year of *Journal de Genève*.

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.

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

Table 1: Outlying years from figure 1.

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 and 4 show the average number of pages per issue for these two newspapers. With exception of the very first years and 1830s of *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.

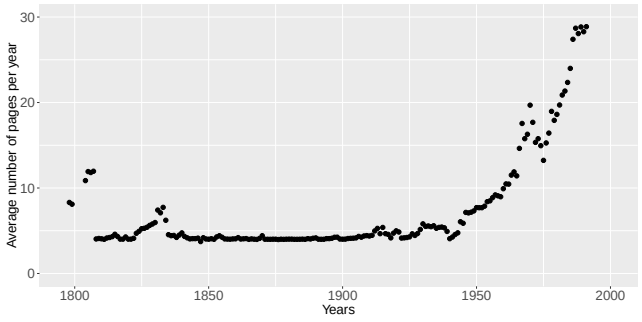


Figure 3: Average number of pages by issue, per year, in *Gazette de Lausanne*.

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

We migrated data from a previous web application. The whole archive, including text, images and all issues, weighs 22 TB. Each page of a newspaper issue is structured by an

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).

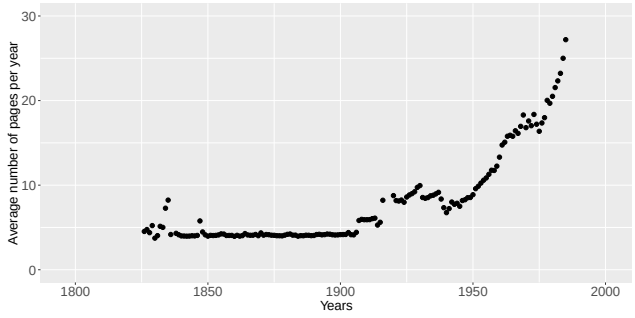


Figure 4: Average number of pages by issue, per year, in *Journal de Genève*.

XML file providing the positions of articles [12]. Each article is encoded by a proper XML file. All the words and special characters were detected and their positions saved.

The quality of the OCR has not been evaluated at this stage, but some mistakes are immediately visible, 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. Each page is saved in TIFF format. Each issue, including all its pages, is saved in a PDF file containing the OCRed text.

3. N-GRAMS

3.1 Initial preprocessing

Due to the error rate resulting from unavoidable OCR errors, preprocessing needs usually to be applied to the raw corpus before indexation in order to improve the overall quality of the results. Indeed, OCR data may include special characters irrelevant for linguistic purposes, like bullet points, the character "l" being recognised as "|" (a pipe), and graphics or paper stains wrongly interpreted as textual data.

As a result, the raw text outputted by the OCR includes a large number of special characters that need to be ruled out. For all text processing tasks, we thus applied a preprocessing filter that recognises only alphanumeric characters.

3.2 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. 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 both 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 corpora 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.

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

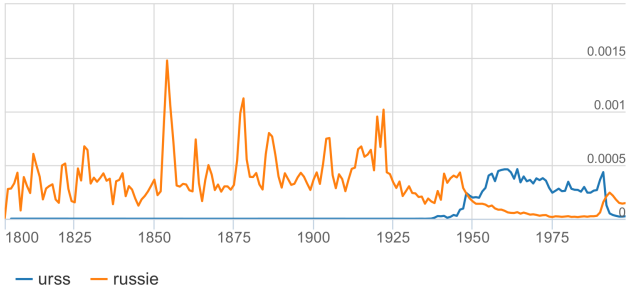


Figure 5: Visualisation of 1-grams "russie" (Russia) and "urss" (USSR).

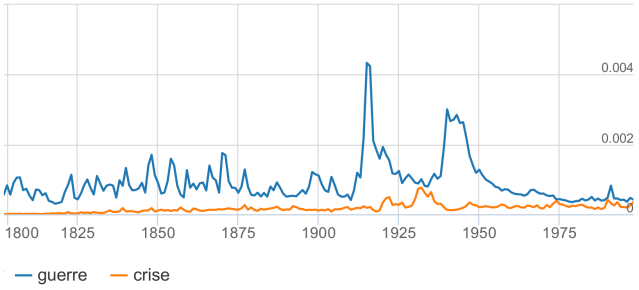


Figure 6: Visualisation of 1-grams "guerre" (war) and "crise" (crisis).

In order to compute the n-grams relative frequencies, we chose a time granularity of one year and divided the number of occurrences of every n-gram by the total number of n-grams counted on 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.

3.3 Possible uses and interpretations

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 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* ? First introduced during the 6^{th} 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 [10, 21]. More recently, NE processing is being 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 by private companies (e.g. Open Calais [13], Zemanta [14], Alchemy [15]) are most of the time 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 (named entities 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 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 developed/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 pre-processing (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 (*Simonetta*) and a last (*Sommaruga*) 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*), expression 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 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 to 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 [16] 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, NIF, OLiA and LexInfo 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

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

We thus needed to define a set of basic requirements that

13. www.opencalais.com
14. www.zemanta.com
15. www.alchemyapi.com

16. JavaScript Object Notation.

would follow the most generic possible use case, yet providing modern and powerful features to journalists, historians and information scientists. The core features that were outlined by preceding studies on similar archives were :

— 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 hierarchical organisations 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 be easily derived from the way contents are organised, as there are no links between articles that may allow to guess their relative importance, nor clear way to predict which kind of content might be interesting to the considered user. To answer this question, we thus introduced the n-gram visualisation as a very part of the search results. This way, each search query is presented starting with its frequency through time on both journals, allowing any user 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.

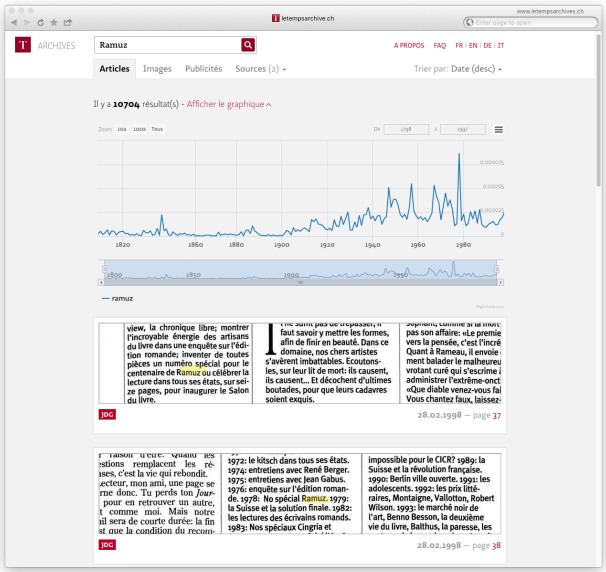


Figure 7: Search results. At the top, period selection using the n-gram viewer. At the bottom, previews from the found articles.

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 [17], 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 send 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 (in order to read from raw files suited for archiving and preservation and deliver web optimised ones instead) and tiling. We selected an image server responding to the *IIIF* norm, for its outstanding interoperability and academic approach [27].

Figures 8 and 9 illustrate the use of the viewing interface.

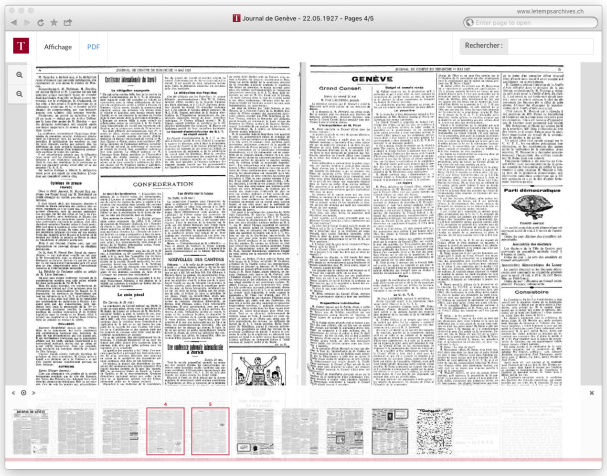


Figure 8: Viewer interface (featuring one full double-page). At the bottom, previews of all pages from the same issue.

The named entities 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 just by clicking on them.

5.2 Application Stack Design

The software setup we decided on was as follows :

— Raw text indexation and search : Apache Solr [18].
— Image Server : Loris IIIF [19].
— Web development frameworks : Laravel [20].
— Internal database engine : PostgreSQL [21].
— Triplestore : Virtuoso Open Source [22].

Figure 11 shows how the different parts interact and are organised. 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

17. A double-page typically weighs about 10 MB.
18. http://lucene.apache.org/solr
19. http://github.com/loris-imageserver/loris
20. http://laravel.com
21. http://www.postgresql.org
22. http://github.com/openlink/virtuoso-opensource

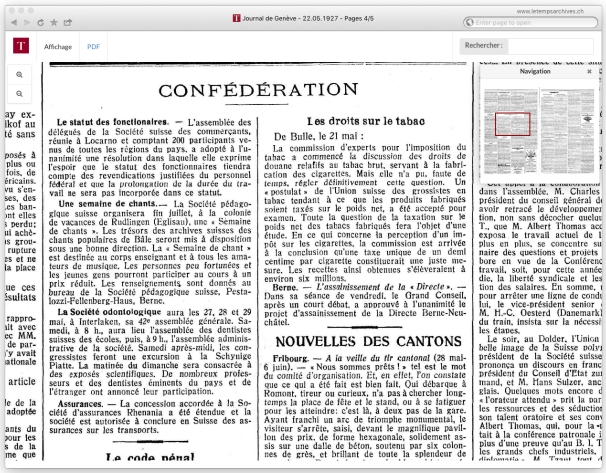


Figure 9: Viewer interface (zooming on one article). In the top right corner, the location of the article in the double page is highlighted.

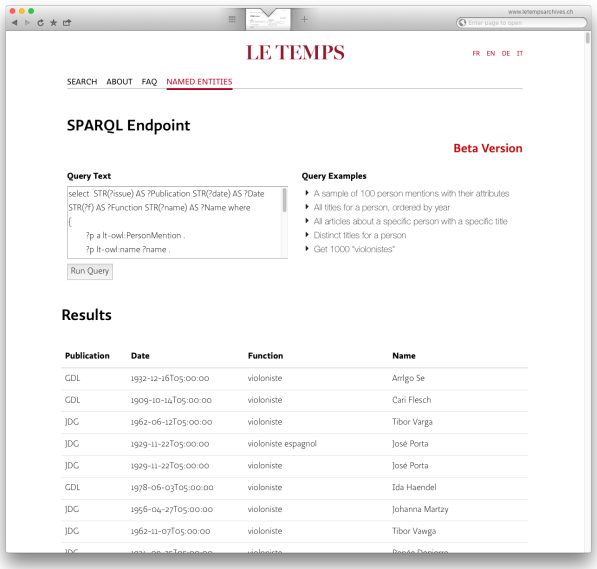


Figure 10: SPARQL endpoint presenting the results of a sample query

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.

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

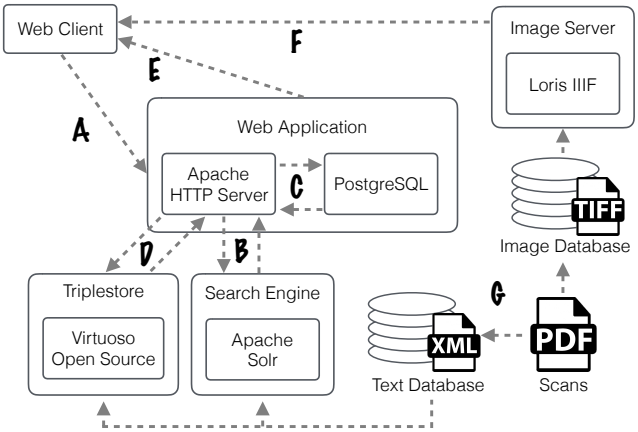


Figure 11: Information workflow and technical components.

using *Cloudflare* services [23].

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 [24], 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.

The access statistics demonstrated great enthusiasm from the general public. On the day of the public launch, *Le Temps* newspaper announced it by a dedicated article [25] and by including a four-pages insert composed mainly of archival articles. Rebounding on the launch, third parties also opened a *Facebook* page [26] to discuss noteworthy findings in the archives, such as century old discussions relevant in regard 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 or to more relevant data, depending on the users' queries. 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].

23. http://www.cloudflare.com
24. http://analytics.google.com
25. http://www.letemps.ch/suisse/2016/03/18/epfl-temps-lancent-un-site-pointe-technologie-faciliter-acces-200-ans-archives Accessed on April 24th, 2016.
26. http://www.facebook.com/groups/Etonnantdansletemps/ Accessed on April 24th, 2016.

— Named entities disambiguation. In use, this allows the user to filter queries that relate to different people or places having the same names.
— Completion of the corpus with the missing journal issues, wherever possible.
— Find new partners and add new collections.

7. ACKNOWLEDGEMENTS

The authors would like to thank the Swiss National Library for their funding of this research, *Le Temps* for sharing their archives, as well as the Scientific Committee composed of Alain Clavien, Marie-Christine Doffey, Gaël Hurlimann, Joëlle Kuntz, Enrico Natale and François Vallotton, and the Geneva Library for their availability. All the actors cited provided precious and constructive feedbacks to our work.

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* 9^{th} *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* 10^{th} *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* 7^{th} *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* 6^{th} *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* 8^{th} *International Conference Recent Advances in Natural Language Processing (RANLP'2011)*, Hissar, Bulgaria, September 2011.

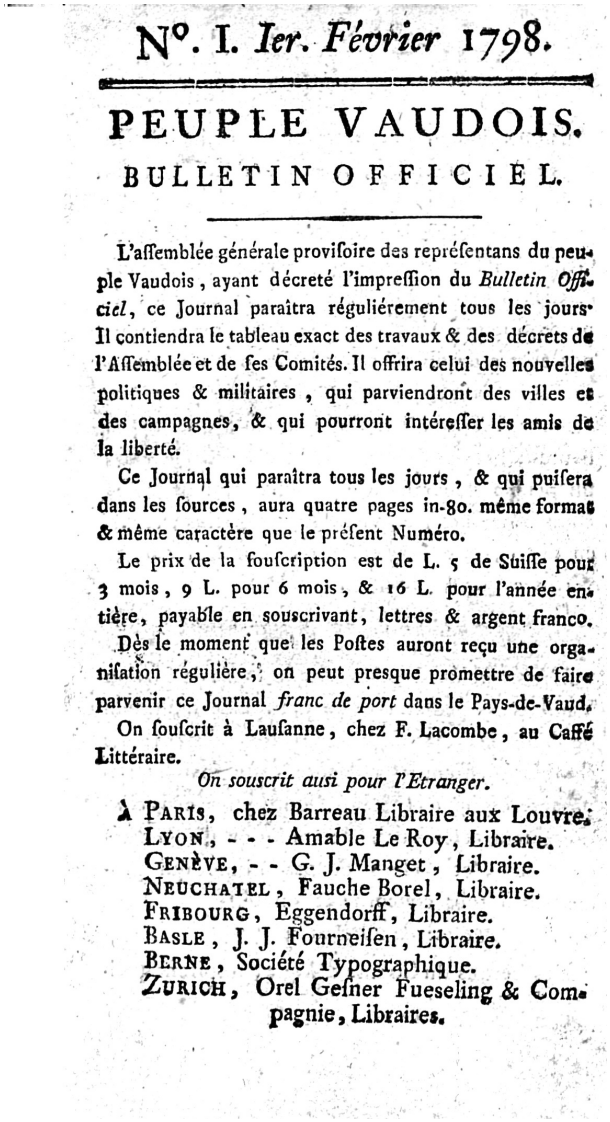
[29] R. Steinberger, B. Pouliquen, and E. van der Goot. An introduction to the europe media monitor family of applications. In . J. K. F. Gey, N. Kando, editor, *Information access in a multilingual world — Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR)*, Boston, USA, July 2009.

[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* 37^{th} *International Conference on Very Large Databases* 9^{th}, page 1450–1453, Seattle, USA, 2011.

APPENDIX



N°. I. *Ier. Février* 1798.

PEUPLE VAUDOIS.

BULLETIN OFFICIEL.

L'aſſemblée générale proviſoire des repréſentans du peuple Vaudois, ayant décreté l'impreſſion du *Bulletin Officiel*, ce Journal paraîtra réguliérement tous les jours. Il contiendra le tableau exact des travaux & des décrets de l'Aſſemblée et de ſes Comités. Il offrira celui des nouvelles politiques & militaires, qui parviendront des villes et des campagnes, & qui pourront intéreſſer les amis de la liberté.

Ce Journal qui paraîtra tous les jours, & qui puiſera dans les ſources, aura quatre pages in-8o. même format & même caractère que le préſent Numéro.

Le prix de la ſouſcription est de L. 5 de Suiſſe pour 3 mois, 9 L. pour 6 mois, & 16 L. pour l'année entière, payable en ſouscrivant, lettres & argent franco.

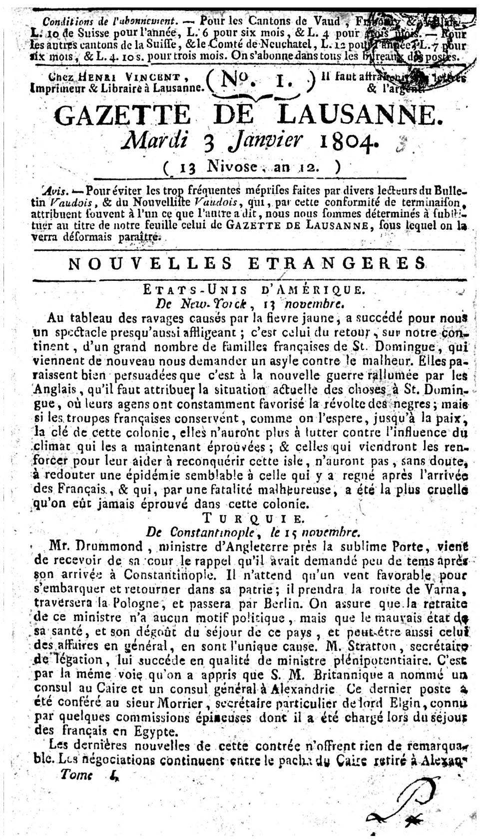
Dès le moment que les Poſtes auront reçu une organiſation régulière, on peut presque promettre de faire parvenir ce Journal *franc de port* dans le Pays-de-Vaud.

On ſouſcrit à Lauſanne, chez F. Lacombe, au Caffé Littéraire.

On souscrit aussi pour l'Etranger.

À PARIS, chez Barreau Libraire aux Louvre.
LYON, - - - Amable Le Roy, Libraire.
GENÈVE, - - G. J. Manget, Libraire.
NEUCHATEL, Fauche Borel, Libraire.
FRIBOURG, Eggendorff, Libraire.
BASLE, J. J. Fourneiſen, Libraire.
BERNE, Société Typographique.
ZURICH, Orel Geſner Fueseling & Compagnie, Libraires.

Figure 12: On February 1st, 1798, the front page of the first issue of what would later become *Gazette de Lausanne* after bearing nine other names [19]. "Bulletin officiel" approximately means "Official news report". This page shows letters s printed as f's.



Conditions de l'abonnement. — Pour les Cantons de Vaud, Fribourg & ... L. 10 de Suisse pour l'année, L. 6 pour six mois, & L. 4 pour trois mois. — Pour les autres cantons de la Suisse, & le Comté de Neuchatel, L. 12 pour l'année, L. 7 pour six mois, & L. 4. 10 s. pour trois mois. On s'abonne dans tous les bureaux des postes.

Chez HENRI VINCENT, Imprimeur & Libraire à Lausanne. (N°. I.) Il faut affranchir les lettres & l'argent.

GAZETTE DE LAUSANNE.

Mardi 3 Janvier 1804.

(13 Nivose, an 12.)

Avis. — Pour éviter les trop fréquentes méprises faites par divers lecteurs du Bulletin *Vaudois*, & du Nouvelliste *Vaudois*, qui, par cette conformité de terminaison, attribuent souvent à l'un ce que l'autre a dit, nous nous sommes déterminés à substituer au titre de notre feuille celui de GAZETTE DE LAUSANNE, sous lequel on la verra désormais paraître.

NOUVELLES ETRANGERES.

ETATS-UNIS D'AMÉRIQUE.

De New-Yorck, 13 novembre.

Au tableau des ravages causés par la fievre jaune, a succédé pour nous un spectacle presqu'aussi affligeant ; c'est celui du retour, sur notre continent, d'un grand nombre de familles françaises de St. Domingue, qui viennent de nouveau nous demander un asyle contre le malheur. Elles paraissent bien persuadées que c'est à la nouvelle guerre rallumée par les Anglais, qu'il faut attribuer la situation actuelle des choses à St. Domingue, où leurs agens ont constamment favorisé la révolte des negres ; mais si les troupes françaises conservent, comme on l'espere, jusqu'à la paix, la clé de cette colonie, elles n'auront plus à lutter contre l'influence du climat qui les a maintenant éprouvées ; & celles qui viendront les renforcer pour leur aider à reconquérir cette isle, n'auront pas, sans doute, à redouter une épidémie semblable à celle qui y a regné après l'arrivée des Français, & qui, par une fatalité malheureuse, a été la plus cruelle qu'on eût jamais éprouvé dans cette colonie.

TURQUIE.

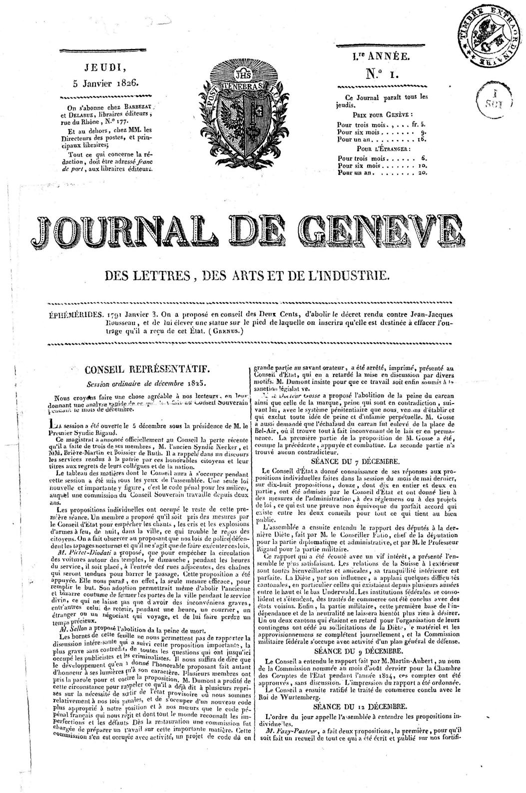
De Constantinople, le 15 novembre.

Mr. Drummond, ministre d'Angleterre près la sublime Porte, vient de recevoir de sa cour le rappel qu'il avait demandé peu de tems après son arrivée à Constantinople. Il n'attend qu'un vent favorable pour s'embarquer et retourner dans sa patrie ; il prendra la route de Varna, traversera la Pologne, et passera par Berlin. On assure que la retraite de ce ministre n'a aucun motif politique, mais que le mauvais état de sa santé, et son dégoût du séjour de ce pays, et peut-être aussi celui des affaires en général, en sont l'unique cause. M. Stratton, secrétaire de légation, lui succède en qualité de ministre plénipotentiaire. C'est par la même voie qu'on a appris que S. M. Britannique a nommé un consul au Caire et un consul général à Alexandrie. Ce dernier poste a été conféré au sieur Morrier, secrétaire particulier de lord Elgin, connu par quelques commissions épineuses dont il a été chargé lors du séjour des français en Egypte.

Les dernières nouvelles de cette contrée n'offrent rien de remarquable. Les négociations continuent entre le pacha du Caire retiré à Alexan-

Tome I.

Figure 13: On January 3rd, 1804, few years after its creation, *Gazette de Lausanne* receives a name it kept for close to two centuries.



JEUDI, 5 Janvier 1826.

1.re ANNÉE. N.° 1.

JOURNAL DE GENEVE

DES LETTRES, DES ARTS ET DE L'INDUSTRIE.

CONSEIL REPRÉSENTATIF.

Session ordinaire de décembre 1825.

SÉANCE DU 7 DÉCEMBRE.

SÉANCE DU 9 DÉCEMBRE.

SÉANCE DU 12 DÉCEMBRE.

Figure 14: *Journal de Genève* was launched on January 5th, 1826.

A lundi!

Lundi 2 septembre 1991 est le jour «J» pour votre quotidien: le «Journal de Genève et Gazette de Lausanne» sort dans sa nouvelle formule, en deux cahiers et avec un graphisme modernisé. Nous vous souhaitons d'ores et déjà bonne lecture et attendons avec intérêt vos réactions. A lundi!

EDITORIAL

Pari tenu: nous voici!

Par Jasmine Audemars

Pari tenu: le «Journal de Genève et Gazette de Lausanne» se présente à ses lecteurs aujourd'hui, lundi 2 septembre 1991. Ce mariage de deux «conjoints» qui totalisent plus de 360 ans d'histoire paraît la chose la plus naturelle du monde. C'est dire l'évolution des esprits dans ce pays qui a le talent de changer tout en cultivant ses racines. Certes – des amis vaudois nous l'ont amicalement reproché – cette union ne respecte pas, dans notre nouveau titre, le principe de l'égalité. Simplement, nous avons dû trouver un compromis entre les exigences d'un graphisme moderne et celles, impitoyables, d'une concurrence croissante. Finalement, à terme, ce qui importe, c'est le fond. Et nos lecteurs vaudois savent déjà qu'ils conserveront leurs pages et que la rédaction vaudoise, à Lausanne, sous la conduite de Daniel S. Miéville, continuera à illustrer la vie du canton. Parallèlement, notre quotidien aura aussi les yeux ouverts sur une réalité régionale qui est en train de bouleverser les cloisonnements traditionnels. Avec ce premier numéro de notre nouvelle formule, nos lecteurs abonnés ou acheteurs au numéro vont découvrir aussi une présentation graphique modernisée, plus aérée et ordonnée.

Elle est destinée à mieux répondre aux vœux d'un public exigeant: un quotidien de qualité se doit d'avoir un habillage de qualité. De même, la répartition des différentes rubriques en deux cahiers séparés répond à des souhaits souvent exprimés. Surtout, les nombreux amateurs du «Samedi littéraire» retrouveront leur supplément culturel en tête de la deuxième section, chaque week-end.

Sur le plan du contenu, la première page sera désormais plus proche de l'actualité, l'éditorial conservant bien sûr sa place privilégiée. Notre quotidien affirmera ainsi toujours haut et fort sa triple vocation de journal d'information, d'analyse et d'opinion. Un sommaire développé, quant à lui, permettra au lecteur de choisir son itinéraire à l'intérieur des pages. La dernière page du «Journal de Genève et Gazette de Lausanne» quant à elle, offrira, nouveauté, un menu varié destiné au lecteur pressé, à la recherche d'une synthèse des principales informations du jour, ou au lecteur désireux de s'accorder quelques minutes «autrement». A l'intérieur du journal, la rubrique culturelle quotidienne et les pages services – cinémas, carnet – prendront un tour résolument inter-cantonal. Chaque vendredi, un agenda culturel romand permettra au lecteur de choisir la meilleure manière possible de passer son temps libre. Et ce n'est là qu'une première étape. La nouvelle formule que nos lecteurs découvrent aujourd'hui va s'affiner et s'enrichir progressivement au fil des semaines. Nous les en informerons au fur et à mesure, car nous préférons tenir, plutôt que promettre. D'emblée, toutefois, nous affirmons notre volonté de développer notre position de quotidien romand de qualité. Nous bénéficions d'acquis précieux: des lecteurs et des annonceurs fidèles, une audience reconnue et une équipe de collaborateurs hors pair dans les domaines rédactionnel, technique et commercial. Avec un tel capital et un dynamisme nouveau, nous voulons et nous allons faire encore beaucoup mieux. ■

Notre nouvelle formule va s'enrichir encore au fil des semaines

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*.

Samedi 28 février, dimanche 1er mars 1998

JOURNAL DE GENÈVE
et Gazette de Lausanne

Fondé en 1826 LE QUOTIDIEN SUISSE D'AUDIENCE INTERNATIONALE

LE DERNIER NUMÉRO

EDITORIAL

Au revoir

PAR IGNACE EBANGERAT

Une journée pour en finir

Vendredi 27 février 1998: le jour où le *Journal de Genève et Gazette de Lausanne* a disparu comme il a vécu.

Fin de partie

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.