Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment

Jean-Philippe Moreux

News Media Section

Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment

Facilitating Access for various Profiles of Users

Jean-Philippe Moreux
Preservation dpt, Digitization service, Bibliothèque nationale de France, Paris, France.
jean-philippe.moreux@bnf.fr

Abstract:

In this age of Big Data this paper describes how digital librairies can apply at large scale innovative approaches to better valorize and bring better experiences of old newspapers.

On the first hand, the state-of-the-art OLR (optical layout recognition) technique in one of the largest heritage press digitization projects in Europe (Europeana Newspapers, www.europeana-newspapers.eu, 2012-2015) was used in a data mining experiment. Data analysis was applied to quantitative metadata derived from a 850K pages subset of six XIXth-XXth c. French newspaper titles from the BnF collection. The METS/ALTO XML data was analyzed with data mining and data visualization techniques that show promising ways for the production of knowledge about historical newspapers that are of great interest for library professionals (digitization programs management, curation and mediation of newspaper collections) and for end-users, particularly the digital humanities community.

On the other hand, the Retronews web portal showcases how advanced semantic annotation techniques can improve the retrieval efficiency on a digital newspapers collection; thus the rediscovery and reappropriation of these documents by various types of users: teachers, students, researchers, general public.

Keywords: heritage newspapers; document analysis; OCR/OLR; metadata; data mining; data visualisation; semantic enrichment; named entities recognition; digital mediation; digital humanities.

1 INTRODUCTION

Libraries are full of digital data and everyday they produce new data: bibliographic metadata are created or updated in catalogs describing collections [1],[2]; usage data on libraries and their audience are collected; digital documents are produced by the digitization of content stored in heritage libraries.

But can library data and metadata fit with the concept of big data? Are they legitimate targets for data mining? Their relatively small volume (12 millions of records for BnF catalog) does not encourage some caution? The criterion of the volume is irrelevant, if we believe Viktor Mayer-Schoenberger and Kenneth Cukier: "(…) big data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value (…)" [1]. On a large scale, but set against the activity ("my big data is not your big data" [2]), with methods different from those satisfying the nominal business needs, and with the aim to "create something new": new links (author, place, date, etc.) are built on top of catalogs (OPAC) [3]; libraries management can be backed by the analysis of attendance and reading data [4]; a history of newspapers' front pages can be written on data extracted from digital libraries (DLs) [5],[6].

E.g., does it make sense to data mine quantitative metadata of the daily newspapers digitized and refined during the Europeana Newspapers project [7]? What lessons can be learned from the Retronews portal, which makes extensive use of semantically enriched data? We attempt to answer these questions by first presenting the process of creating new metadata; then some methods of analysis, interpretation and reuse of these metadata; and finally data quality issues.

2 THE EUROPEANA NEWSPAPERS LIVE CASE

2.1 Creating new Metadata

Six national and regional newspapers (1814-1945, 880K pages, 140K issues) of BnF collections are part of the data set OLR'ed (Optical Layout Recognition) by the project Europeana Newspapers. The OLR refinement consists of the description of the structure of each issue and article (spatial extent, title and subtitle, etc., using METS/ALTO formats [8]) and the classification of content types (MODS format).

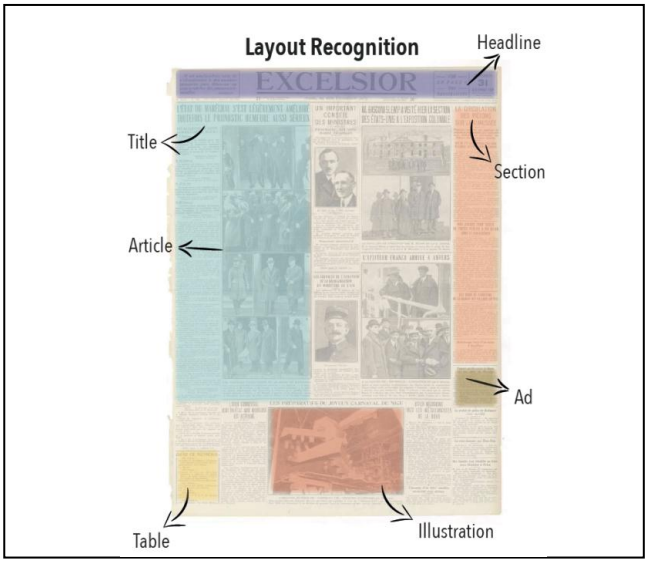


Fig. 1. OLR process

2.1.1 From the Digital Documents to the Derived Data

OCR and OLR files are full of interesting objects marked up into the XML:

- OCR (ALTO) is a source for quantitative metadata: number of words, illustrations & tables; paper format…

- OLR (METS) is a valuable source too for high level informational objects: number of articles, titles…; identification of sections (groups of articles); content types classification (ads, judicial, stock market…)

Based on this finding, from each digital document a set of bibliographical and descriptive metadata related to content and layout is derived, both at issue and page levels (date of publication, number of pages, articles, words, illustrations, etc.). XSLT or Perl scripts (Fig. 2) are used to extract some metadata from METS manifest (e.g. number of articles) or OCR files (e.g. number of words). The complete set of derived data contains about 5.5M atomic metadata values expressed with XML, JSON or CSV formats.

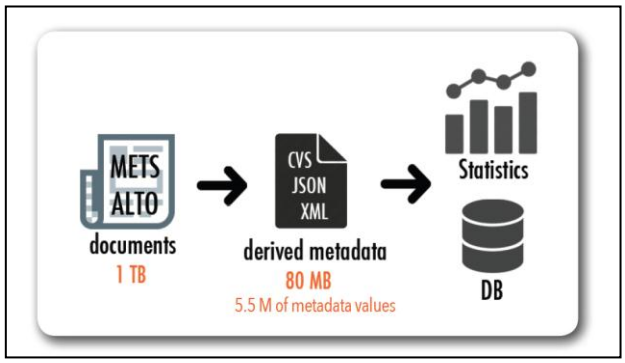


Fig. 2. Derived data production process

This operating principle has many advantages:

- It makes light derived data sets, rather than heavy XML corpora available to end-users.
- It's not rocket science and it's fast (30,000 pages/hour with an optimized NoXML parsing Perl script).

2.2 WHEN THE DATA TALK

2.2.1 Producing Knowledge

Some data describe a reality of which the analyst has prior knowledge or intuition. This is the case of statistical information helping to pilot digitization or curation actions. The data set could then be a representative sample of the collection, because the information sought are mostly statistical measures.

Digitization programs: What is the density in articles of these newspapers (Fig. 3)? What is the potential impact on OLR processing costs?

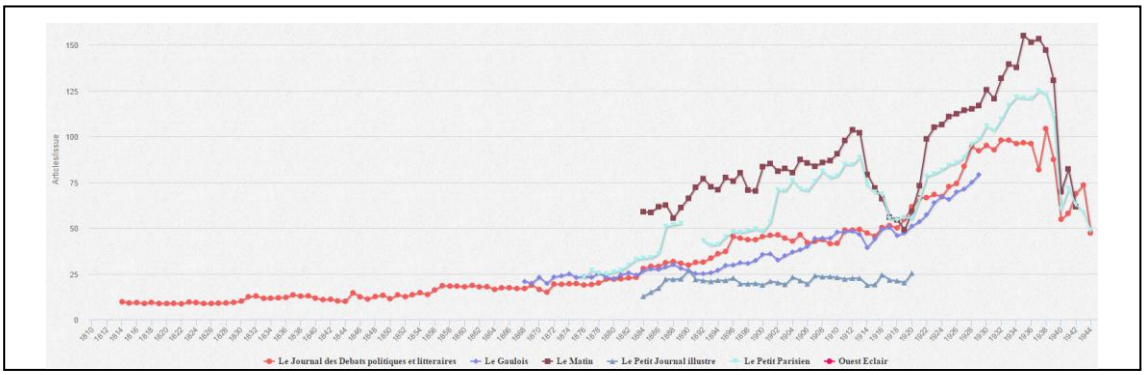


Fig. 3. Average number of articles per issue

Image bank: What titles contain illustrations (Fig. 4)? What is the total number of images one can expect?

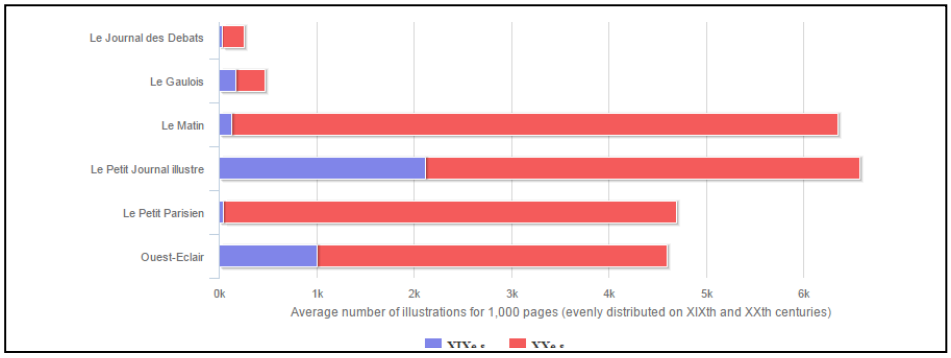


Fig. 4. Average number of illustrations for 1,000 pages

Invited to comment on these results, the collections curators easily establish links with the documentary reality they know:

> "Of course, *Le Matin* is a daily which was published during the golden age of modern newspapers (1890-1914) and emblematic of the age's innovations: it is highly structured and illustrated." (Fig. 3: brown curve; Fig. 4: 6k illustrations for 1k pages)
>
> "The *Journal des Débats politiques et littéraires* (JDLP) founded in 1789 is an heir of the first newspapers (gazettes): it retains throughout its history a rubric based layout, and in which the illustration is rare." (cf. Fig. 3: orange curve, Fig. 4: only 225 illustrations for 1k pages)

The collected statistical measures help to enrich this knowledge with actual data (mean, total, maximum, distribution…) of great value for digitization program managers but librarians as well, in the case where such techniques were applied to the DL collection as a whole.

2.2.2 Discovering knowledge through visualization

Data visualization allows researchers (digital humanities, history of press, information science) to discover meaning and information hidden in large volumes of data. Moreover, OLR content types classification feature helps researchers to spot on specific types of content.

The History of the Press

Front page: The role of the image in the daily press is a classic research subject [5],[10] that data mining analysis and visualization tools can enrich with micro-facts as well as macro-trends. Thus, the singular curve describing a supplement of *Le Petit Journal illustré* (Fig. 5) highlights the appearance of the full front page illustration on Nov. 29, 1890.

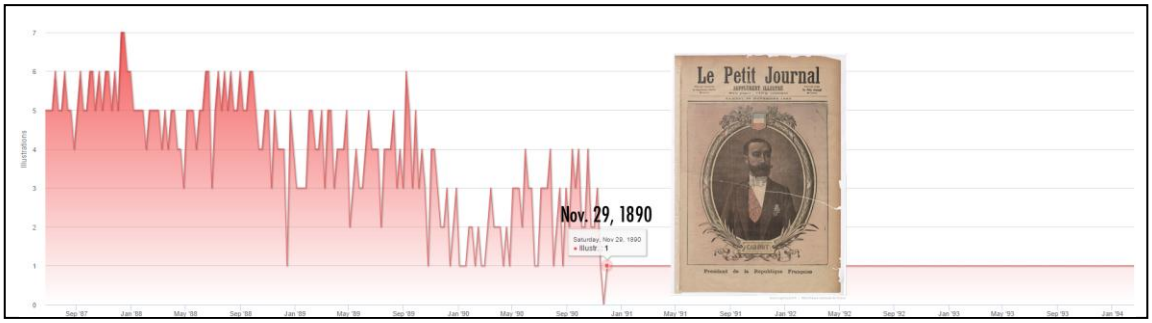


Fig. 5. Average number of illustrations on front page (*Le Petit Journal illustré*)

Figure 6 highlights that the number of illustrations on *Le Petit Parisien* front page (blue curve) exceeds the average by 1902, and then follow an exponential growth: in the 1930s, the front page contains 45% of the illustrations of a 8 to 10 pages issue.

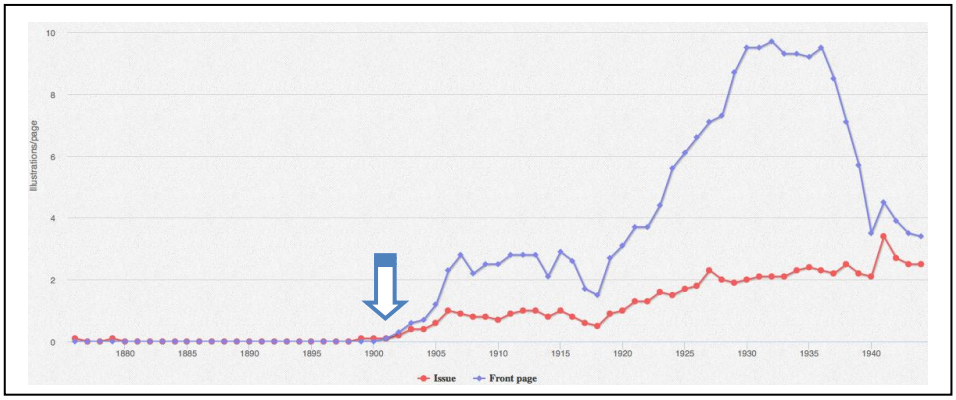


Fig. 6. Average number of illustrations per page (*Le Petit Parisien*)

Activity: The content classification performed during OLR refinement allows an analysis in terms of types of content (text, table, ad…). Figure 7 shows the impact of the Great War on the activity and assesses the period of return to pre-war level activity (roughly 10 years).

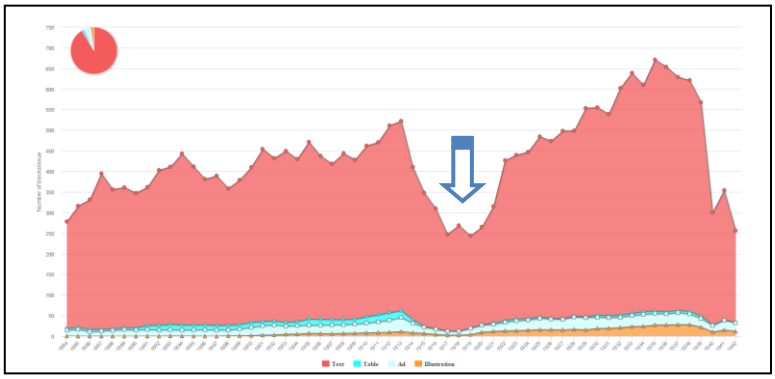


Fig. 7. Types of content per issue (*Le Matin*)

Layout*:* Form factors and layout of Dailies have varied considerably over time. Fig. 3 allows us to locate a major transition in the 1880s, with two families of newspapers, the "old", poorly structured into articles (*Le Gaulois, Journal des Débats politiques et littéraires*) and the "modern" (*Le Matin, Le Petit Parisien, Ouest-Éclair, Le Petit Journal illustré*) borned with a structured layout. Combining in a bubble chart (Fig. 8) the three form factors of "modernity" which are the average number of articles per page (*x*), illustrations per page (y) and illustrations on front page (z) illustrate this typology.

Stock market section in Daily [11]: The content classification performed during OLR can help researchers focusing on specific newspapers content, e.g. Stock Market section (quotes and analysis). The quantitative metadata are of a great help because "tables" in newspapers are predominantly used in such quotes (Fig. 9).

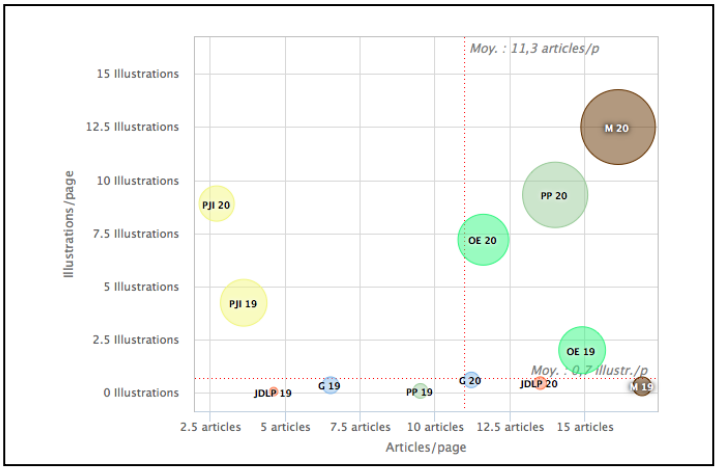


Fig. 8. Newspaper modernity classification

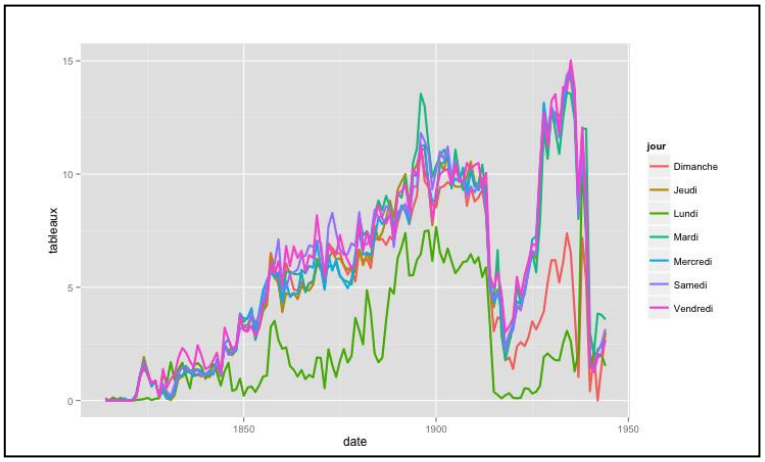


Fig. 9. Tables per week day, 1838-1870 (© P-C Langlais, 2016)

The History of Newspaper's titles

Data visualization on a complete data set (one data per issue) makes possible to focus on a specific daily title.

Digital archeology of papermaking and printing: Page format information can be retrieved from the digitized images. It provides researchers the complete printing history of a title.

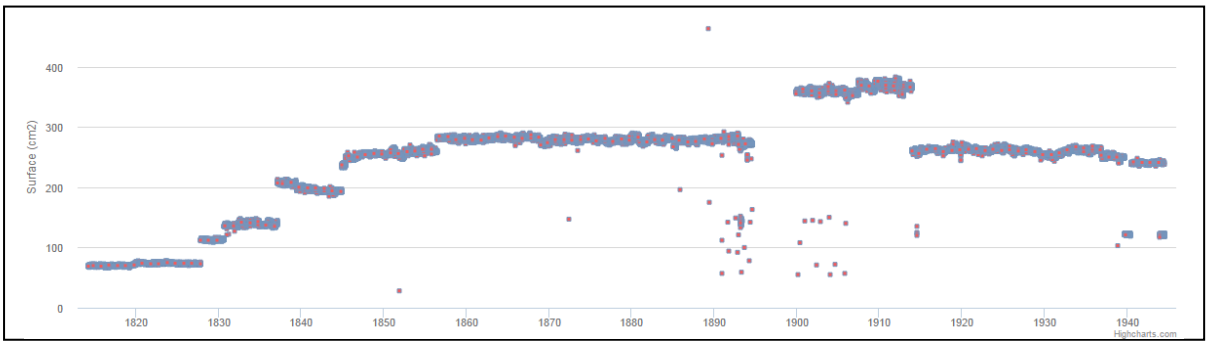


Fig. 10. Page format (*JDPL*, complete data set)

Illustrations: Data visualization of illustration density can reveal outstanding values like these highly illustrated issues of the *Journal des Débats politiques et littéraires* (Fig. 11), which prove to be illustrated supplements (March 27, 1899, 201 illustrations). It also reveals micro-facts such as the first published illustration in this title (within an ad, May 11, 1828).

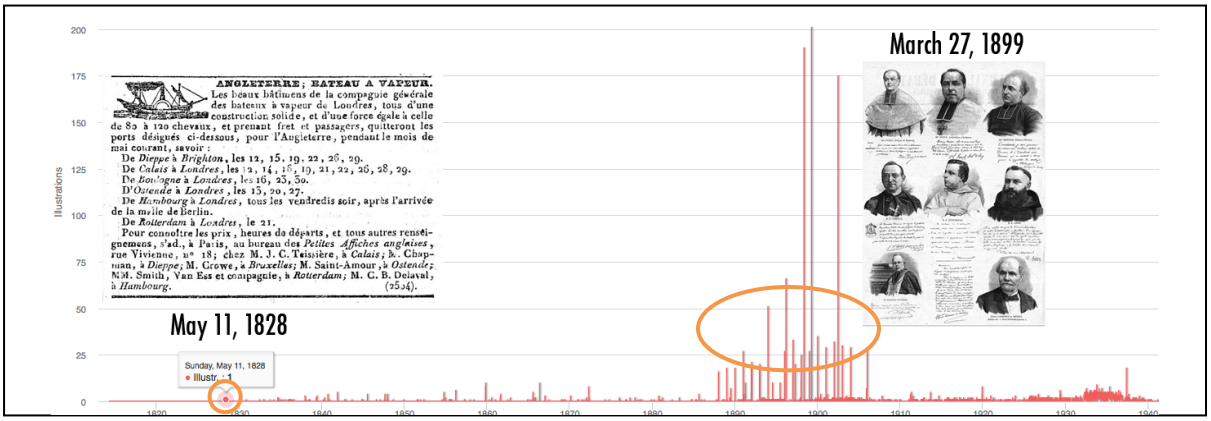


Fig. 11. Number of illustrations per issue (*JDPL*, complete data set)

2.2.3 Engaging new Audience with Datavizualisation

Data visualization facilitates rediscovery and reappropriation of the digital documents described by these data. Such methods and tools can help DLs to improve the access to their collection, in addition to the classic keyword spotting and page flip mode.

Fig. 12 shows an interactive web chart of the word density per page over the complete data set of the *Journal des débats politiques et littéraires* (1824-1944, 45K issues). Data singularities demonstrated by the chart can prompt users to discover and browse the collection differently:

- The significant breaks in the scatter plot chart are linked to the successive changes in layout and/or format (as studied by historians of the press [12]), motivated by technical innovations on papermaking and printing (e.g.: Dec. 1, 1827: 3 columns, 330×450mm; Oct. 1, 1830: 4 col.; March 1, 1837: 400×560mm) or historical events (Aug. 4, 1914: move to 2 p. and 3 col. then back to 6 col. on Aug. 8).
- Outliers can also reveal treasures, likes this 24 words/p. issue (May 2, 1889, Paris Universal Exposition's map) or examples of censorship during the WW1 (22 May, 1915).

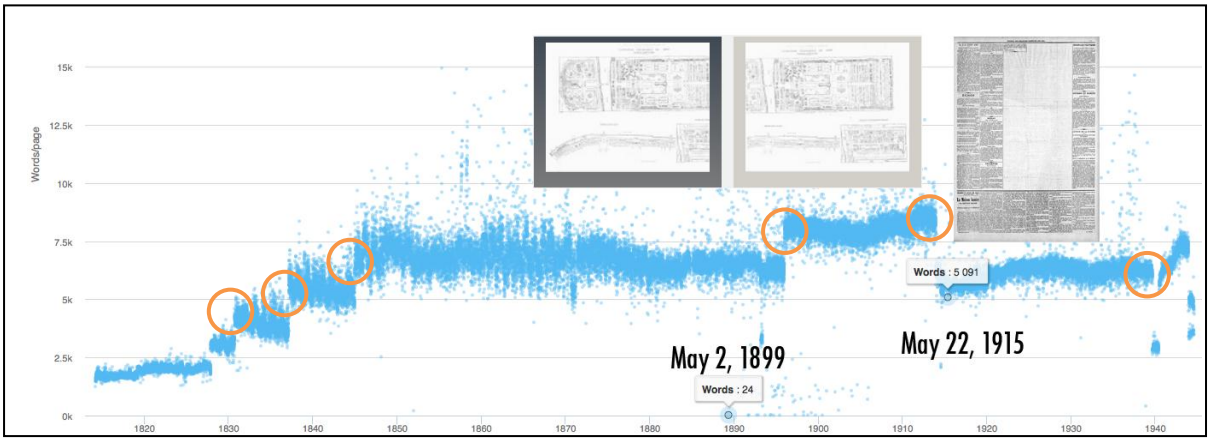


Fig. 12. Average number of words per page (*JDPL*, complete data set)

2.2.4 Querying the Metadata

Improving the effectiveness of the analysis can be achieved with dedicated tools or processes (ETL frameworks, APIs to access content [6],[9], XML or no-SQL databases, statistical environments like R…). BaseX (basex.org) is one of these simple and elegant solutions to agglomerate all the individual metadata files in a unique database and to query it using XPath/Xquery languages. As part of a digital mediation action devoted to a press title or to the complete data set, a basic FLWOR query will identify all "graphical" pages, that is to say both those poor in words (here based from an average) and including at least one illustration:

```
<result>
{let $textDensity := 0.25
 let $threshold := avg(//page/nbString) * $textDensity
 for $page in //page[blockIllustration>1]
 where  $page/nbString < $threshold
  let $tokens := fn:tokenize($page/../../metad/date, "\.")
  return
    <illustration>
      <date>{$tokens[3]}-{$tokens[2]}-{$tokens[1]}</date>
      <pageNo>{$page/fn:count(preceding-sibling::page) + 1}</pageNo>
    </illustration>}
</result>
```

This query retrieves hundred of pages from the whole data set (comics, portraits, press cartoon, maps, ads…), which would have been extremely laborious to manually identify.

Fig. 13. Samples from the results: *Ouest-Éclair*, *Le Petit Journal*, *Le Petit Parisien*, *Le Matin*, *Le Gaulois*, *Le JDPL*

BnF digital curators and mediators have expressed their interest in this approach and a XQuery HTTP API have been set up on the BaseX database, which helps them to identify graphical "nuggets" through the BnF newspapers collection. Fig. 14 shows an iconographical research related to the murder of Gaston Calmette: relevant illustrations are easily retrieved.

Fig. 14. Image search API results with *date* and *front page* criteria

Similar queries can be written to dig into the data and find specific types of content previously identified with dataviz, e.g. the pages censored during the Great War (see Fig. 12), which have a slightly smaller word counts than the pages average. This method leads to a 45% recall rate and a 68% precision rate (based on a ground truth carried on the *JDPL* front pages for 1915). Obviously a medium performance, showing the limits of a statistical approach when applied to a word based metric biased by layout singularities (titles, ads, etc.) and proved to be ineffective on illustrated pages. However a successful method if completeness is not required, like in a mediation context (see the resulting Gallica blog post) or for fuzzy search on length of documents in terms of word counts (see [13]).

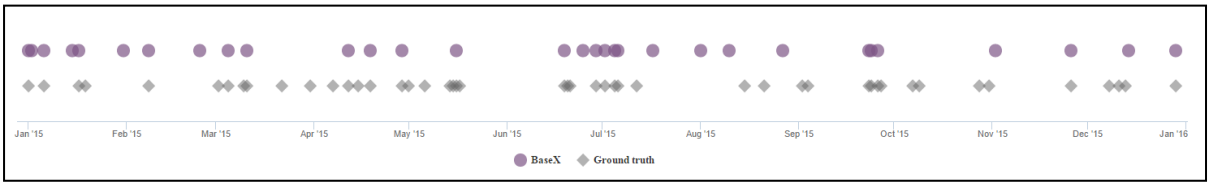


Fig. 15. Censored issues: query results & GT (*Journal des débats politiques et littéraires*, 1915)

2.2.5 Advanced Search Modes for Newspapers

DLs' items are not anonymous text. They all have specific form factors, they are all part of the long history of publishing. Consequently, feeding the DL search engine with layout and structural metadata could allow users to perform advanced mixed queries taking into account this fact.

Fig. 16 lists a couple of queries leveraging some of the rich informational objects which have been previously discussed (illustrated article, article title, article including tables) and the other available information sources (catalog, OCRed text).

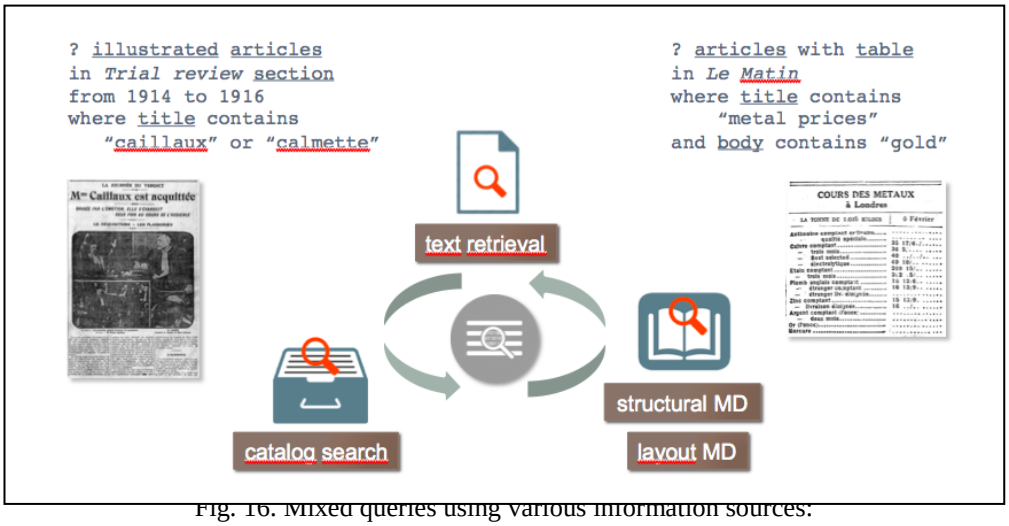


Fig. 16. Mixed queries using various information sources: famous Caillaux trial (left); metal prices quotes (right)

Trove (http://trove.nla.gov.au) is an emblematic example of this approach. Fig. 17 shows Trove advanced search form dedicated to structural and layout information search criteria (category of content, word count, illustration count).

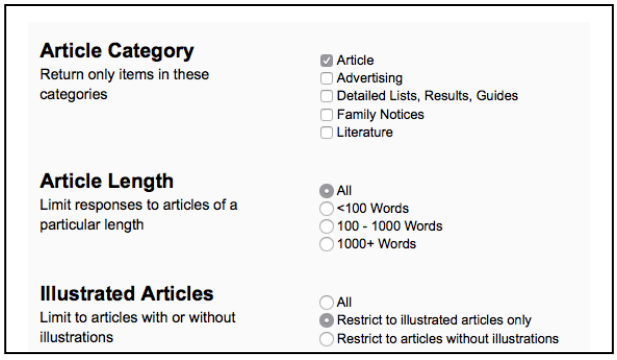
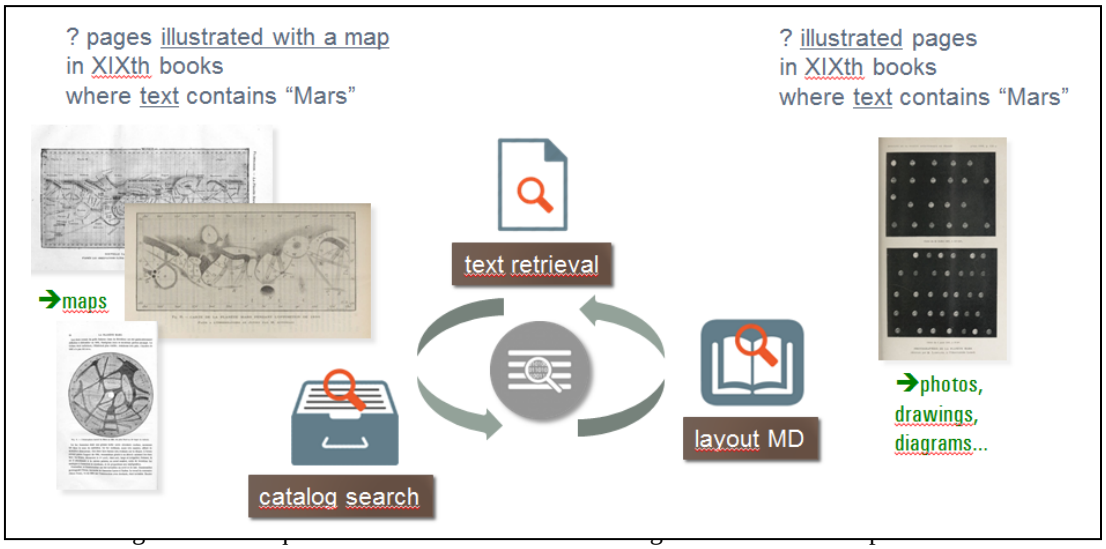


Fig. 17. Trove advanced search form (extract)

And what about books? (one could ask). Books' OCR also contains meaningfull quantitative information [13]: word, table, map, ornament, drop cap… Using the same approach, users could perform complex illustration retrieval tasks on illustrated content (maps, pictures).



2.3 Data QA

The quality of derived data affects the validity of the analysis and interpretation [4],[14]. Irregular data in nature or discontinuous in time may introduce bias. A qualitative assessment should be conducted prior to any interpretative analysis.

Newspapers are characterized by the relative homogeneity of their shape over time, which induces consistency and constant granularity of the derived metadata (issue, page, article...). Moreover, its large size and the option to apply the analysis to the entire data set and not a subset of it also guarantees its representativity [15].

The data itself can sometimes contribute to their own QA. A synthetic calendar display of available data for a title (*JDPL*, Fig. 19) shows rare missing issues, which suggests that the digital collection is representative of the reality [16].

Or, before starting a study on stock market section ([11], §2.2.2) based on the content typed "table", one can empirically validate this hypothesis by the sudden inflections recorded in 1914 and 1939 for all titles (Fig. 20), being known and established the historical fact of the temporary halt of trading during the two World Wars.

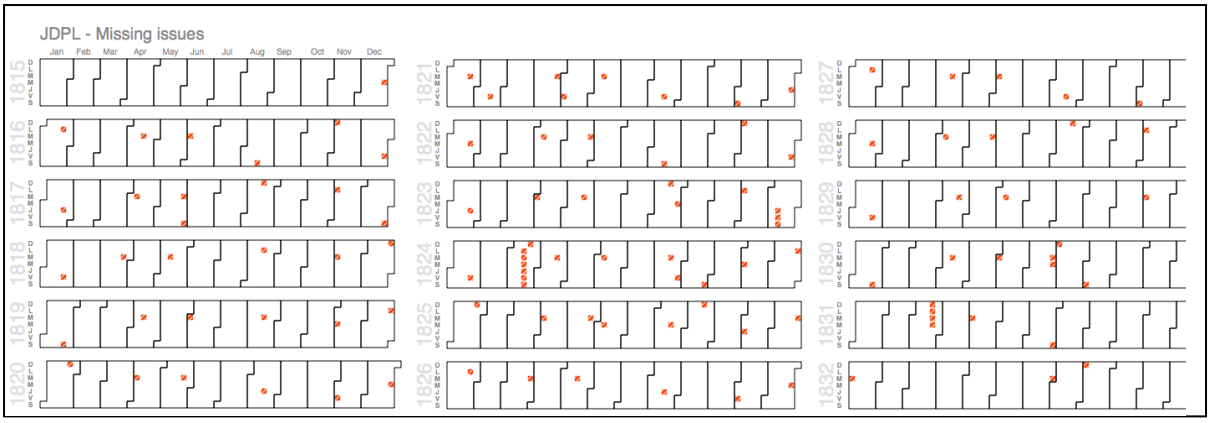


Fig. 19. JDPL missing issues (1814-1944)

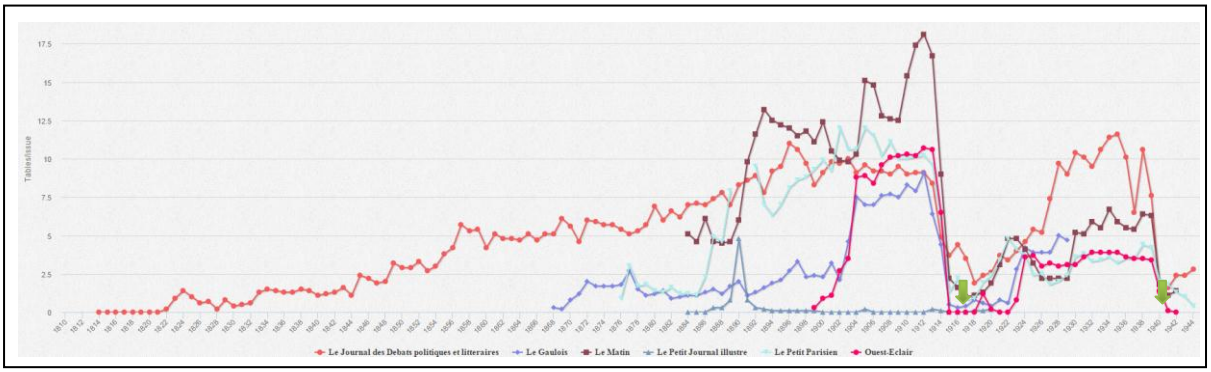


Fig. 20. Average number of tables per issue

Furthermore, care must be taken to inform users of the data set characteristics: production method, known deficiencies, through or under-representation issues, etc.

3 THE RETRONEWS PORTAL

Retronews (www.retronews.fr) is a BnF public/private partnership web portal project, launched in 2016. The overall objective of the project is to facilitate access to information contained in a corpus of three centuries of heritage press.

3.1 Creating new Data

Retronews foundations are based on four concepts (named entities, themes, events, topics) from which the text corpus is enriched with semantic annotations:

- *Named entities* (the NE's 3 categories are: person, place, organization): The NE recognition is driven by linguistic grammar-based techniques and authorities' records (BnF, VIAF, dictionaries of famous people from the XVIIth up to the XIXth c.)
- *Themes* (14 top level themes, 231 second level themes): Derived from the IPTC classification and refurbed for the heritage press. A lexica has been created for each theme, and the text corpus indexed relatively to these lexica.
- *Events* (147): Closed list of historical events defined by the editorial team, each event associated with a lexica (Wikipedia).

- *Topics* (≅20,000): The topic modelling uses Wikipedia articles titles and a list of the most frequent queries expressed by users on Gallica press content.

3.2 When the Data Talk

3.2.1 Advanced Search Modes for Newspapers

Adding a pinch of semantic flavor helps to get closer to natural language queries. If we reformulate the previous example (§2.2.5, "Caillaux trial"), taking advantage of the semantic features available in Retronews (Fig. 21) like named entities recognition or topic modelling, we now gain a richer expressivity:

```
I'm looking for illustrated articles on front page in "Trial" topic
from 1914 to 1916 which contain NE.person "Henriette Caillaux"
or "Gaston Calmette"
```

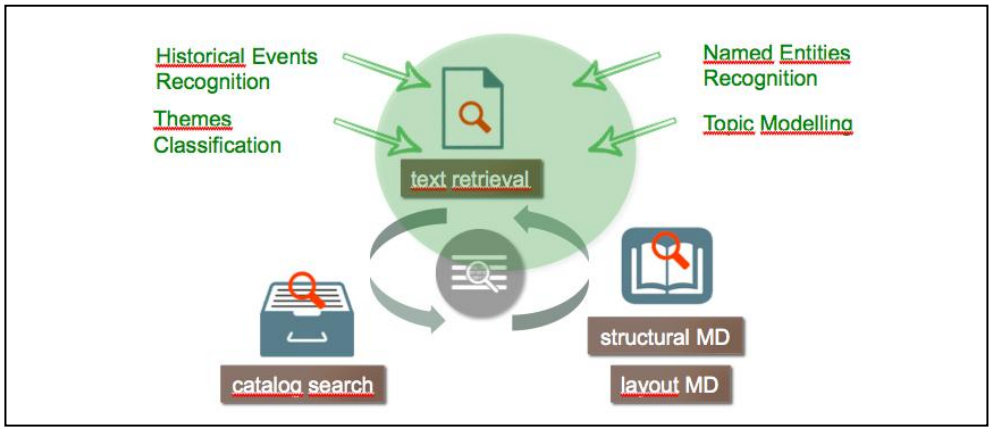


Fig. 21. Advanced query system on a semantically refined corpus

Retronews has implemented a faceted search functionality to allow end-user to express queries on the semantic concepts already mentioned as well as on a layout criteria ("front page only"). Fig. 22 shows the "Caillaux trial" query expressed with these features and a page extracted from the result list, showing how the semantic classification data are rendered: NE occurrences colored on the image itself; semantic classification data (themes and NE) listed under the page.

3.2.2 Engaging new Audience with Data Vizualisation

The semantic annotated corpus are legitimate candidates for innovative visualization modes. The next version of Retronews will implement classic text mining tools (terms frequency, concordance, proximity, comparison of speech, named entities graph, etc.).

Data visualization artefacts like timelines are also perfectly suited for historical daily. Fig. 23 shows the Retronews newspapers timeline produced by digital mediators for editorial purposes.

Thanks to data mining and data visualization techniques, extracted facts can also enrich individual daily timelines. Fig. 24 shows the *Journal des Débats politiques et littéraires* timeline, mixing the main events of the daily history (extracted from bibliographic data but also thanks to the data mining analysis) and historical events, some of the latest sometimes interfering with the first (e.g. changes of format or paging during wartime).

Fig. 22. Retronews faceted search: date, theme, NE (person)

Fig 23. [Retronews timeline]

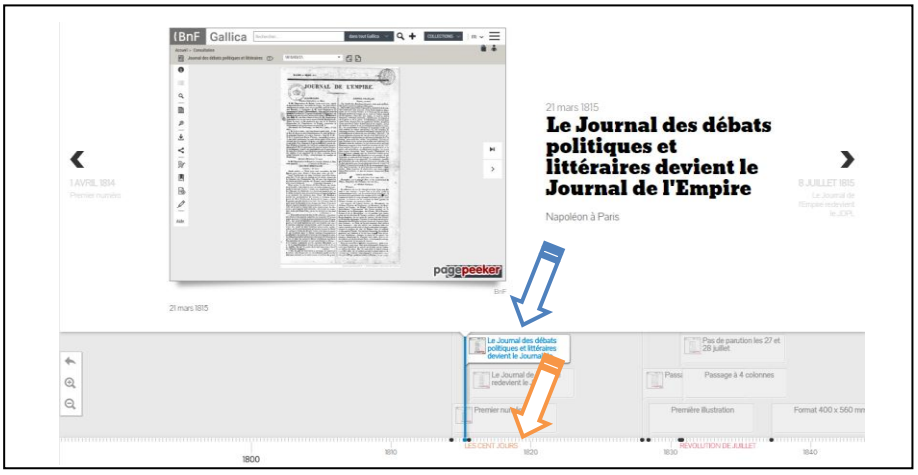


Fig 24. *Journal des Débats* timeline: daily events (blue) and historical events (orange)

3.3 Data QA

The different semantic annotation techniques used have posed particular challenges when applied to historical news content:

- *OCR Text Recognition*: The quality of text produced by the OCR has a direct influence on the results of semantic extraction tasks (NE, topic modelling, etc.). Now the corpora has an average quality due to its nature (degraded documents) or its mode of production (scan from microfilm).
- *OCR Segmentation*: Some semantic extraction tasks operate at the informational entity of the paragraph or article's level (topic, events, themes). But this level of structure may not exist (OCR without OLR) or be unreliable (OCR detects paragraphs but is mistaken).
- *Reference corpus and real corpus not synchronous*: Some contemporary information resources used to facilitate the semantic extraction tasks (Wikipedia, IPTC) applied with difficulty to a heritage corpus.

Moreover, the classic difficulties of semantic annotation have occurred, particularly noise (over-sensing), silence (sub-detection) and disambiguation. All of these difficulties have been taken into account:

- *NE Recognition*: Noise has been reduced by working on syntactic tagging, by inserting a rules engine taking into account the context and by filtering NEs using alignment with the authorities' records.
- *Themes Extraction*: Editorial work on the IPTC classification to make it relevant for heritage dailies; fine tuning of the sub-detected and over-detected themes; filtering of the number of detected themes per page. Some barriers (beyond the state of the art) persist, such as the theme detection of an illustrated front page (cf. Fig. 5).
- *Events Extraction*: A date filter has been add to reduce noise; the lexica have been manually enriched to cope with the anachronism issue (Wikipedia/heritage daily).
- *Topic Modelling*: The 20,000 entries list extracted from Wikipedia article titles has been manually cleaned up to remove contemporary topics.

Finally, the semantic annotations are not perfect (they can't be), neither the information retrieval functionalities (we cannot index everything with everything), but the overall benefit for the end-users is quite significant.

CONCLUSION

The first live case exposed (Europeana Newspapers data set) shows that even meaningless descriptive metadata can give new insights into the history of the press and into history itself through the use of basic data mining methods and tools. This surprising finding is explained by the target corpus, daily press, ideal subject for OLR structural enrichment and hence, the production of consistent metadata over a large period of time.

As a digital library, to create and disseminate such metadata (by download or by any other means: API, web app, etc.) gives researchers a field of study ready to use and easy to use: a corpus of 1TB of METS/ALTO files leads to a set of metadata weighing a few MB, in formats (CSV, JSON) suitable for statistical analysis.

Moreover, we showned that this method also enhances the information retrieval capacities of DLs' end-users and helps them to cope with amounts of information ever larger, from innovative perspectives.

Its results could be followed up in various ways:

- Apply the same data mining process to the other Europeana Newspapers OLR'ed data sets to expand the scope of the analysis to the entire European press and to the on-going BnF press digitization program, which also uses OLR [17].
- Experiment with other types of materials having the desired consistency characteristic and a temporal dimension (e.g. long life magazines or revues, early printed books).
- Provide the derived data sets to researchers. Such data, possibly crossed with the OCRed text transcription, usually provide a fertile ground for research hypotheses [18].
- Generalize the principle of derived data set for researchers to text data sets, based on the themes (Ad, Weather, Judicial sections…) marked up during the OLR process.
- Assess the opportunity of setting up a data mining framework in the BnF to be feed with Gallica's collections.

The last three issues will be addressed during the BnF research project "Corpus" (2016-2018), which aims to study the data mining and text mining services a library can provide for researchers.

The second live case (Retronews) demonstrates the potential of semantic refinements on information retrieval functionalities. Retronews faceted search considerabily increases the capacity of users to find facts and things in the newspapers' content.

We believe that these two use cases have amply demonstrated that DLs can benefit from the digital humanities methods and tools (data and text mining, dataviz, automatic language processing), and in turn provide better service to all users, including the DH community.

Acknowledgments

The author thanks all partners of the Europeana Newspapers project; the Retronews team; Frederick Zarndt and Caroline Kageneck for proofreading this article. Scripts, data sets and charts are freely available: http://altomator.github.io/EN-data_mining.

References

1. Cukier K., Mayer-Schönberger V., *Big Data: A Revolution That Will Transform How We Live, Work, and Think,* Eamon Dolan/Houghton Mifflin Harcourt, 2013.
2. Green R., Panzer M., "The Interplay of Big Data, WorldCat, and Dewey", in *Advances In Classification Research Online*, 24(1).
3. Teets M., Goldner M., "Libraries' Role in Curating and Exposing Big Data", *Future Internet* 2013, 5, 429-438.
4. Lapôtre, R. "Faire parler les données des bibliothèques : du Big Data à la visualisation de données – Let the data do the talking: from Big Data to Dataviz". Library Curator memorandum, ENSSIB, 2014. http://www.enssib.fr/bibliotheque-numerique/notices/65117-faire-parler-les-donnees-des-bibliotheques-du-big-data-a-la-visualisation-de-donnees
5. The Front Page, http://dhistory.org/frontpages.
6. Sherratt, T., "4 million articles later…", June 29, 2012. http://discontents.com.au/4-million-articles-later
7. www.europeana-newspapers.eu
8. Neudecker, C., Wilms L., KB National Library of the Netherlands, "Europeana Newspapers, A Gateway to European Newspapers Online", FLA Newspapers/GENLOC PreConference Satellite Meeting, Singapore, August 2013.
9. Beranger, F., "Big Data – Collecte et valorisation de masses de données", *Livre blanc Smile*, 2015. http://www.smile.fr/Livres-blancs/Erp-et-decisionnel/Big-data
10. Joffredo, L. "La fabrication de la presse". http://expositions.bnf.fr/ presse/arret/07-2.htm.
11. Langlais, P.-C., "La formation de la chronique boursière dans la presse quotidienne française (1801-1870). Métamorphoses textuelles d'un journalisme de données – The Stock exchange section in the French daily (1801-1870)". Thèse de doctorat en science de l'information et de la communication, CELSA Université Paris-Sorbonne, 2015
12. Feyel, G., *La Presse en France des origines à 1944. Histoire politique et matérielle*, Ellipses, 2007
13. Lease Morgan, E., "Use and understand: the inclusion of services against texts in library catalogs and discovery systems", *Libray Hi Tech*, Vol 30 Iss 1 pp. 35-59.
14. Jeanneret, Y., « Complexité de la notion de trace. De la traque au tracé » In: Galinon-Mélénec Béatrice (dir.). *L'Homme trace. Perspectives anthropologiques des traces contemporaines*. CNRS Editions, Paris, 2011
15. Aiden, E., Michel, J.-B., *Uncharted: Big Data as a Lens on Human Culture*. New York: Riverhead Books, 2013
16. Dunning A., and Neudecker, C., "Representation and Absence in Digital Resources: The Case of Europeana Newspapers", Digital Humanities 2014, Lausanne, Switzerland. http://dharchive.org/paper/ DH2014/Paper-773.xml
17. Bibliothèque nationale de France, « Référentiel d'enrichissement du texte », 2015. http://www.bnf.fr/fr/professionnels/numerisation_boite _outils/ a.numerisation_referentiels_bnf.html
18. The Comédie-Française Registers Project, http://cfregisters.org/en