News Media Section

Data mining Historical Newspaper Metadata

Old news teaches history

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 the state-of-the-art OLR (optical layout recognition) technique in one of the largest heritage press digitization projects in Europe (www. europeana-newspapers.eu, 2012-2015) was used in a data mining experiment. Data analysis was applied to descriptive metadata (number of pages, articles, words, illustrations, ads…) derived from a subset of the Europeana Newspapers collection. The METS/ALTO XML data from a 850K page subset of six XIXth-XXth century French newspaper titles from the collection 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 digital libraries (digitization programs management, curation and mediation of newspaper collections) as well as for the digital humanities. Equipped with basic tools widely used in libraries (XSL, spreadsheet, charts generator), we show that simple descriptive metadata can give insights into the history of the press and into history itself.

Keywords: data mining; data visualisation; heritage newspapers; metadata; OCR/OLR.

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 encourages 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 (auteur, lieu, date, etc.) are built on top of catalogs (OPAC) [3]; libraries managment can be backed by the analysis of attendance and reading data [4]; a history of newspaper front pages can be written on data extracted from digital libraries [5],[6].

E.g., does it make sense to data mine descriptive metadata of the daily newspapers digitized during the Europeana Newspapers project [7]? We attempt to answer this hypothesis by first presenting the process of creating a set of descriptive metadata (§II); then some methods of analysis and interpretation of data (§III); and finally data quality issues (§ IV).

2 CREATING NEW DATA

2.1 The Europeana Newspapers dataset

Six national and regional newspapers (1814-1945, 880,000 pages, 150,000 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, classification of content types) using the METS/ALTO formats [8].

2.2 From the digital documents to the derived data

From each digital document is derived a set of bibliographical and descriptive metadata relating to content (date of publication, number of pages, articles, words, illustrations, etc.). Shell and XSLT scripts called with Xalan-Java (Fig. 1) 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 4,500,000 atomic metadata values.

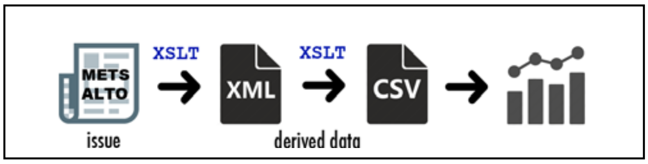


Fig. 1. Derived data production process

This process produces XML, JSON and CSV data, the latter then imported into a spreadsheet template to produce other data (e.g. means), or to compile data over a time period (year), and finally to generate interactive web graphs (with Highcharts).

More effective technical solutions exist: ETL frameworks, APIs to access content [6],[9], high level languages with SAX/DOM API, XML databases, statistical environments like R, etc. but a choice was made to only apply common library formats and tools (XML, XSLT, spreadsheet) and to reuse an in-house toolbox dedicated to the statistical analysis of OCR contents.

3 WHEN THE DATA TALK

3.1 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. 2)? What is the potential impact on OLR processing costs?

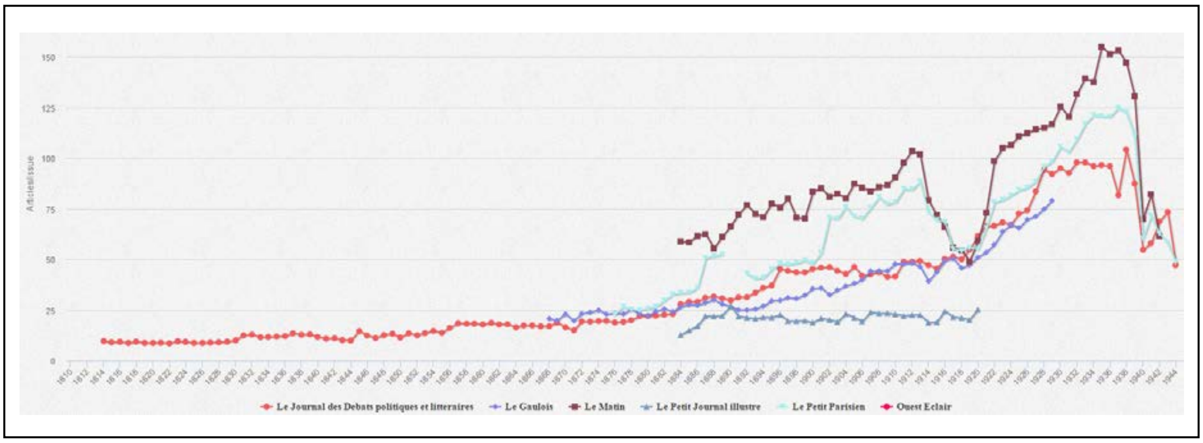


Fig. 2. Average number of articles per issue

- *Text correction*: What is the average density in words of these newspapers (Fig. 3)? What text correction efforts will be required?

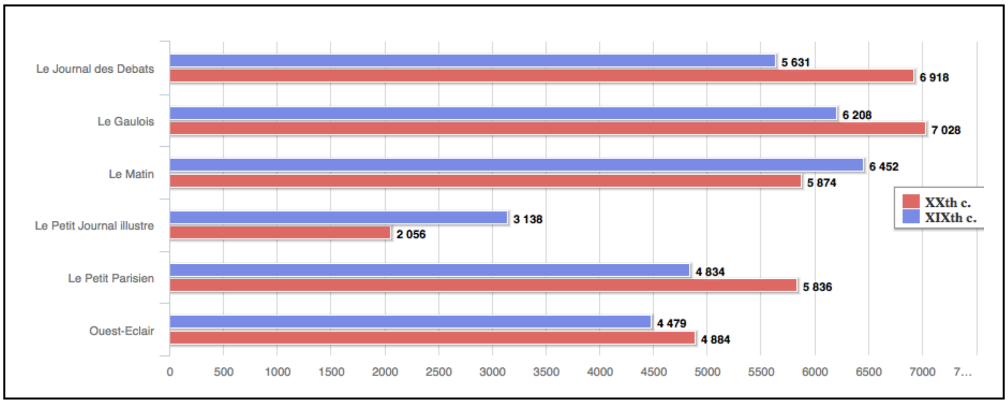


Fig. 3. Average number of words per page

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

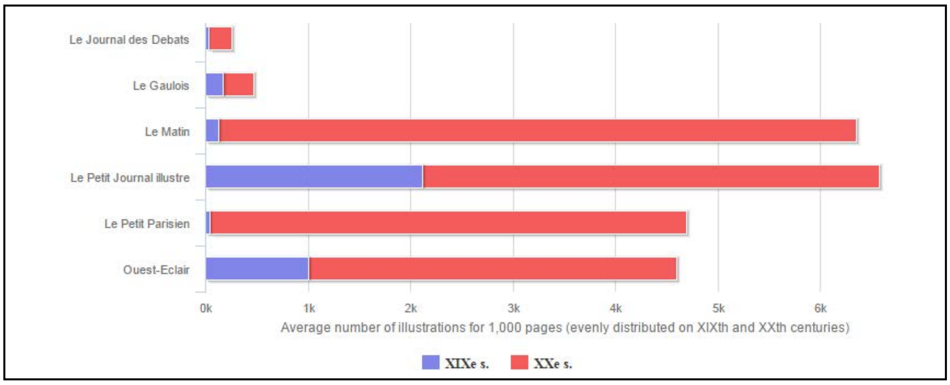


Fig 4. Average number of illustrations for 1,000 pages

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. 2: 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. 2: 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… (e.g. *Le Matin* collection contains 194,000 illustrations and 1.9 million articles, with a maximum on June 13, 1936: 224 articles).

3.2 Discovering knowledge through visualization

Data visualization allows us to discover meaning and information hidden in large volumes of data, but also facilitates rediscovery and reappropriation of the digital documents described by these data.

3.2.1 History of 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.

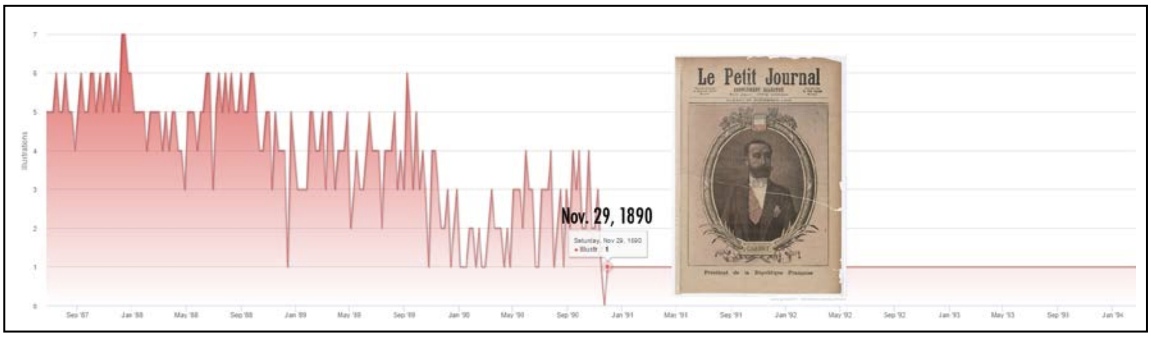


Fig. 5. Average number of illustrations on front page (*Le Petit Journal illustré*)

Figure 6 also shows 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.

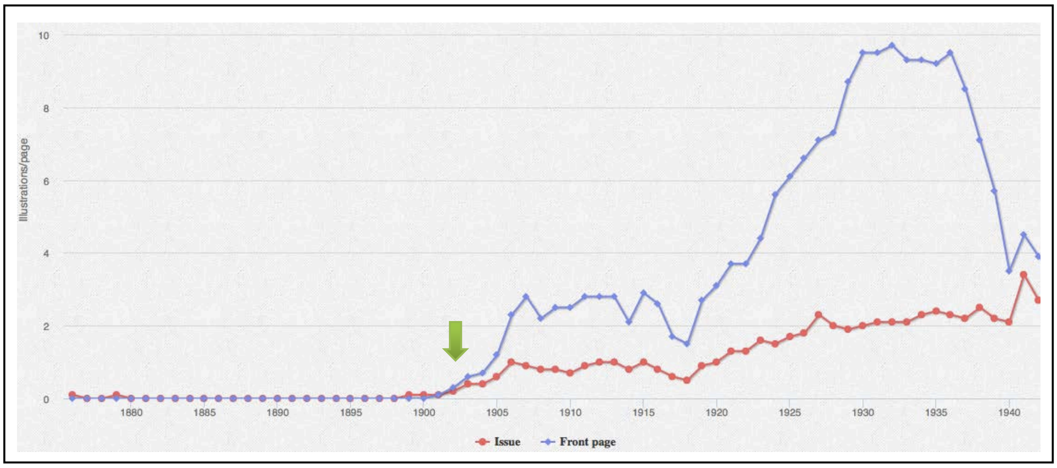


Fig. 6. [Average number of illustrations per page]()
(*Le Petit Parisien*)

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 of the press and assesses the period of return to pre-war level activity (roughly 10 years).

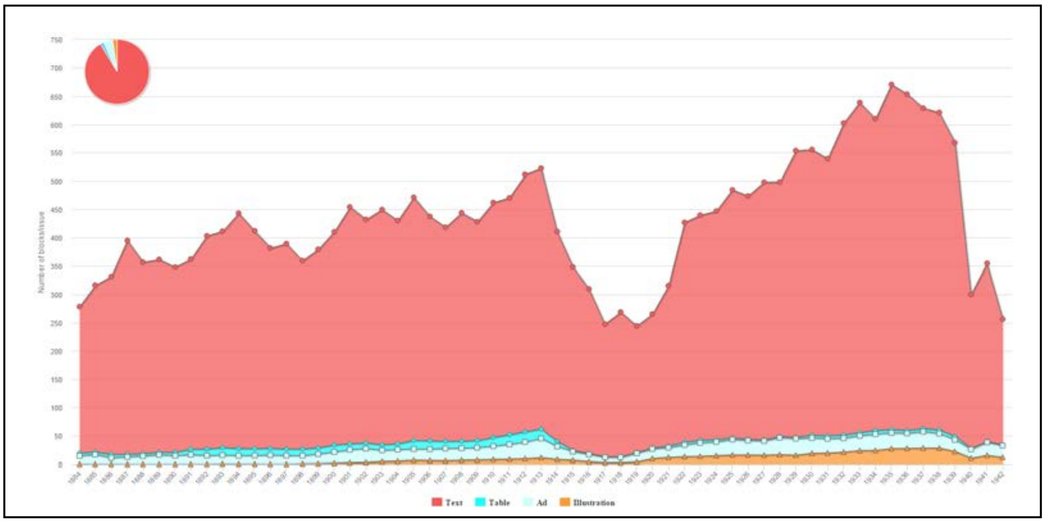


Fig. 7. [Types of content per issue]() (*Le Matin*)

Layout*:* Form factors and layout of dailies have varied considerably over time. Fig. 2 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.

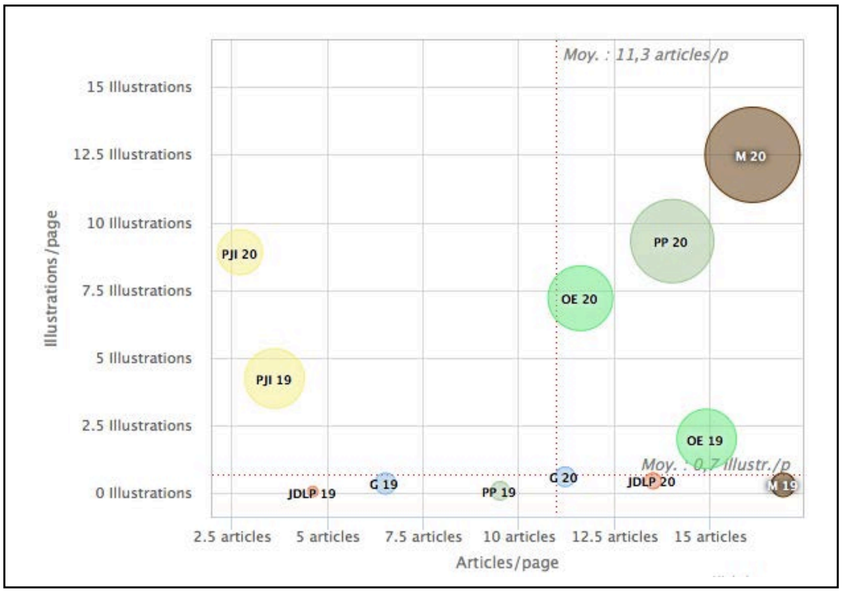


Fig. 8. Newspaper modernity classification

3.2.2 *History of newspaper titles*

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

Layout: Visualization of the word density per page over the complete dataset of the *Journal des débats politiques et littéraires* (1824-1944) reveals significant breaks (Fig. 9). This phenomenon is 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 the political context (Aug. 4, 1914: move to 2 p. and 3 col. then back to 6 col. on Aug. 8). Aberrant values can also reveal treasures, likes this 24 words/page issue (May 2, 1899, Paris Universal Exposition's map) or examples of censorship during the Great War (e.g. 22 May, 1915).

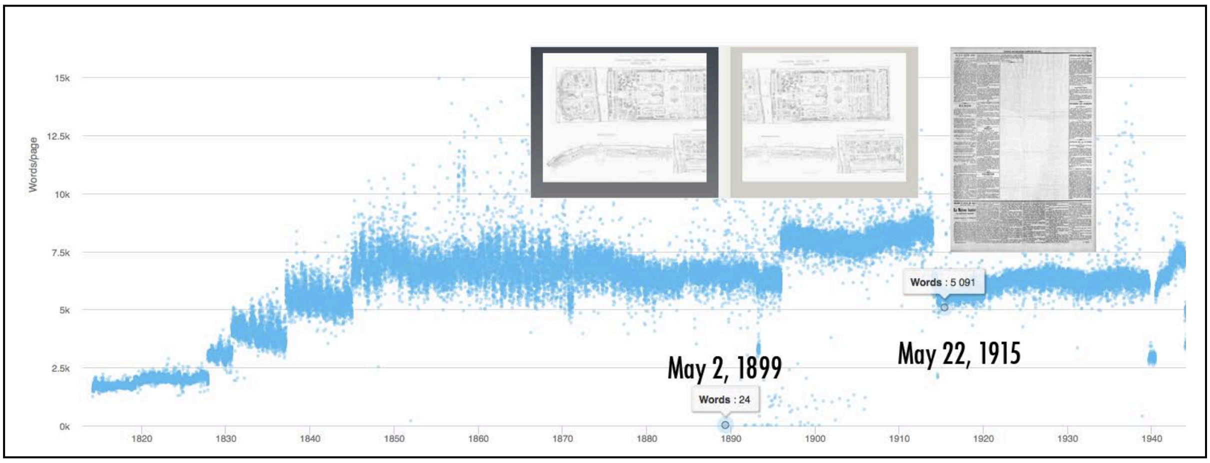
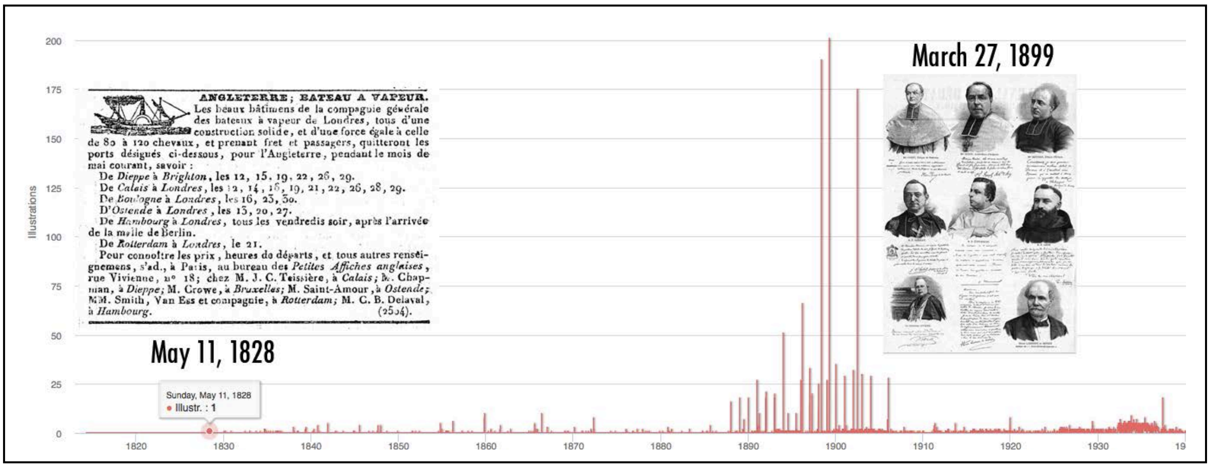


Fig. 9. Average number of words per page
(*JDPL*, complete dataset)

Supplements: Data visualization of illustration density can reveal outstanding values like these highly illustrated issues of the *Journal des Débats politiques et littéraires* (Fig. 10), which prove to be illustrated supplements (March 27, 1899, 201 illustrations). This chart also reveals micro-facts such as the first published illustration in this title (in an ad, May 11, 1828).



ANGLETERRE; BATEAU A VAPEUR. Les beaux bâtimens de la compagnie générale des bateaux à vapeur de Londres, tous d'une construction solide, et d'une force égale à celle de 80 à 120 chevaux, et prenant fret et passagers, quitteront les ports désignés ci-dessous, pour l'Angleterre, pendant le mois de mai courant, savoir :

De *Dieppe* à *Brighton*, les 12, 15, 19, 22, 26, 29.
De *Calais* à *Londres*, les 12, 14, 16, 19, 21, 22, 26, 28, 29.
De *Boulogne* à *Londres*, les 16, 23, 30.
D'*Ostende* à *Londres*, les 13, 20, 27.
De *Hambourg* à *Londres*, tous les vendredis soir, après l'arrivée de la malle de Berlin.
De *Rotterdam* à *Londres*, le 21.

Pour connoître les prix, heures de départs, et tous autres renseignemens, s'ad., à Paris, au bureau des *Petites Affiches anglaises*, rue Vivienne, n° 18; chez M. J. C. Teissière, à *Calais*; M. Chapman, à *Dieppe*; M. Crowe, à *Bruxelles*; M. Saint-Amour, à *Ostende*; MM. Smith, Van Ess et compagnie, à *Rotterdam*; M. C. B. Delaval, à *Hambourg*. (2504).

Fig. 10. Number of illustrations per issue
(*JDPL*, complete dataset)

3.3 Requesting the dataset

As said earlier, improving the effectiveness of the analysis can be achieved with dedicated tools or processes. 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. As part of a digital mediation action devoted to a press title or to the complete dataset, a basic FLWOR query will identify all “graphical” pages, that is to say both those poor in words and including at least one illustration:

```
<result>
  {for $pages in //analyseAlto/contents/page
  where  $pages/illustrationBlocks>=1  and
        $pages/nbString<100
  return
  <page>
    <title>{data($pages/../../metad/title)}</title>
    <date>{data($pages/../../metad/date)}</date>
    <illustrations>{data($pages/illustrationBlocks)}
      </illustrations>
    <file>{data($pages/file)}</file>
  </page>}
</result>
```

This query retrieves hundred of pages from the whole dataset (Fig. 11: comics, portraits, press cartoon, maps, ads…), what it would have been extremely laborious to manually identify.

Fig. 11. Samples from the result list: *Ouest-Éclair*, *Le Petit Journal*, *Le Petit Parisien*, *Le Matin*, Le Gaulois, *Le JDPL*

Similar queries can be written to dig into the data and find specific types of content previously identified with dataviz, e.g. the front pages censored during the Great war, which have a smaller word count than the average of the issue:

```
<result>
{for $annee in ("1914","1915","1916","1917","1918")
  return
    let $average := avg(//metad[matches(//date,$year) and
       (matches(//title,"Le Journal"))]/../content/page[1]/
         [blockIllustration=0]/nbString)
    for $pages in //analyseAlto[(matches(//title,"Le
        Journal ")) and (matches(//date,$year))]/content/page[1]
   where $pages/blockIllustration=0 and
         (1.03*$pages/nbString) < $average

    return ...
```

Parameterized with a 3% threshold, 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, illustrations, ads, etc.). However a successful method if completeness is not required.

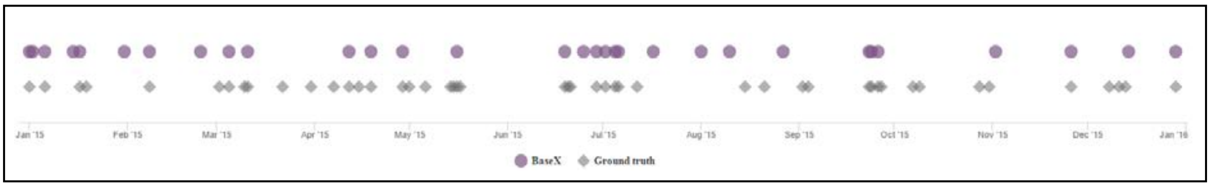


Fig. 12. Censored issues (*Journal des débats politiques et littéraires*, 1915)

4 DATA QA

The quality of derived data affects the validity of the analysis and interpretation [4],[13]. Irregular data in nature or discontinuous in time may introduce bias. A qualitative assessment should be conducted prior to any interpretative analysis.

This press corpus is characterized by the relative homogeneity of its 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 [14].

The data itself can sometimes contribute to their own QA. A calendar display of available data for a title (*JDPL*, Fig. 13) shows rare missing issues, which suggests that the digital collection is representative of the reality [15].

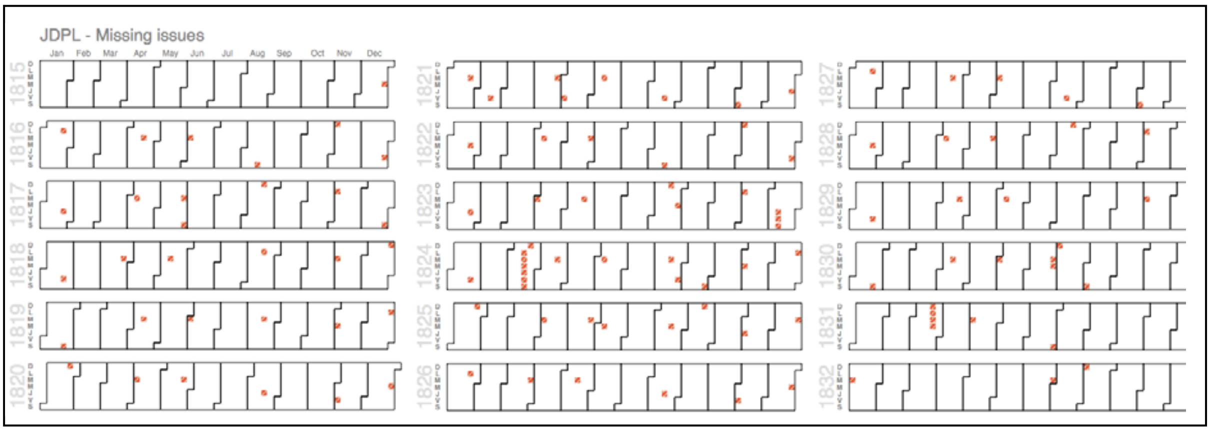


Fig. 13. JDPL missing issues (1814-1944)

Or, before starting a study on daily stock market quotes based on the content typed “table” [11], one can empirically validate this hypothesis by the sudden inflections recorded in 1914 and 1939 for all titles (Fig. 14), being known and established the historical fact of the virtual halt of trading during the two World wars.

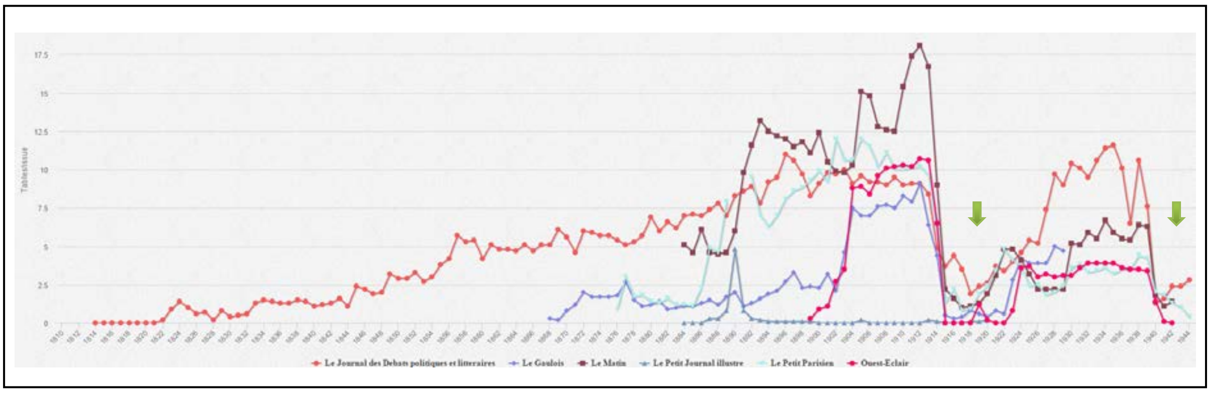


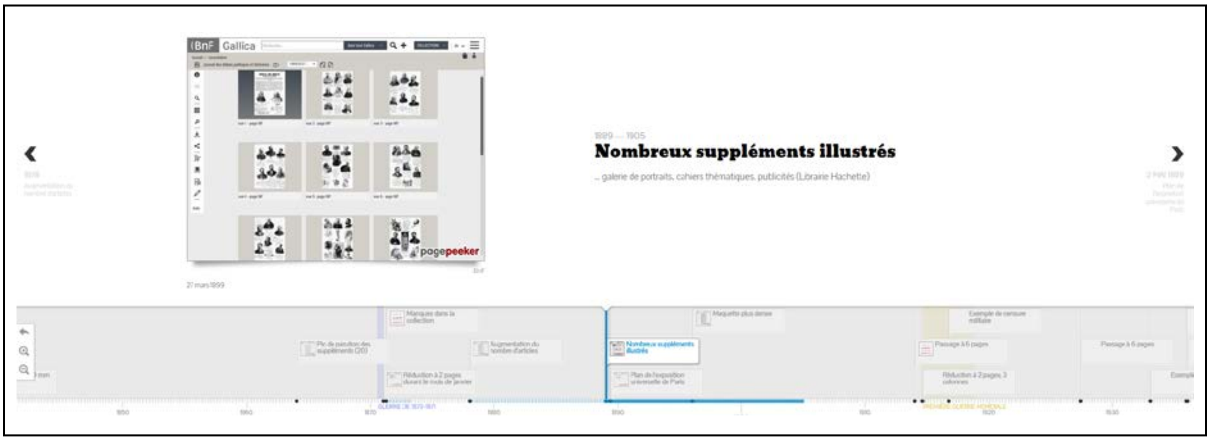
Fig. 14. Average number of tables per issue

CONCLUSION

This experiment showed that even meaningless descriptive metadata can give new insights into a collection 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.

Its results could be followed up in various ways:

- Apply the same data mining process to the other Europeana Newspapers OLR'ed datasets to expand the scope of the analysis to the entire European press and to the BnF press digitization program, which also uses OLR [16].
- 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 analysis to curators and digital mediators for editorialization purposes: introduction to the collection, timelines (fig. 15), and set up a BaseX toolkit (Web client/server architecture and JSON serialization of XML data) to empower them with data mining skills and methods.



1899 — 1905

Nombreux suppléments illustrés

... galerie de portraits, cahiers thématiques, publicités (Librairie Hachette)

Fig 15. Timeline example (*JDPL*)

- Provide the derived datasets to researchers in digital humanities, history of press, information science. Such data, possibly crossed with the OCRed text transcription, usually provide a fertile ground for research hypotheses.
- Assess the opportunity of setting up a data mining framework in the BnF to be feed with Gallica's collections.

The last two issues will be addressed during the BnF research project « Corpus » (2016-2019), which aims to study the data mining and text mining services a library can provide for researchers.

Acknowledgments

The author thanks all partners of the Europeana Newspapers project, and Frederick Zarndt for proofreading this article. Scripts, datasets and charts are 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. 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
14. Aiden, E., Michel, J.-B., *Uncharted: Big Data as a Lens on Human Culture*. New York: Riverhead Books, 2013
15. 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
16. Bibliothèque nationale de France, « Référentiel d’enrichissement du texte », 2015. http://www.bnf.fr/fr/professionnels/numerisation_boite _outils/ a.numerisation_referentiels_bnf.html