DATA MINING HISTORICAL NEWSPAPERS METADATA

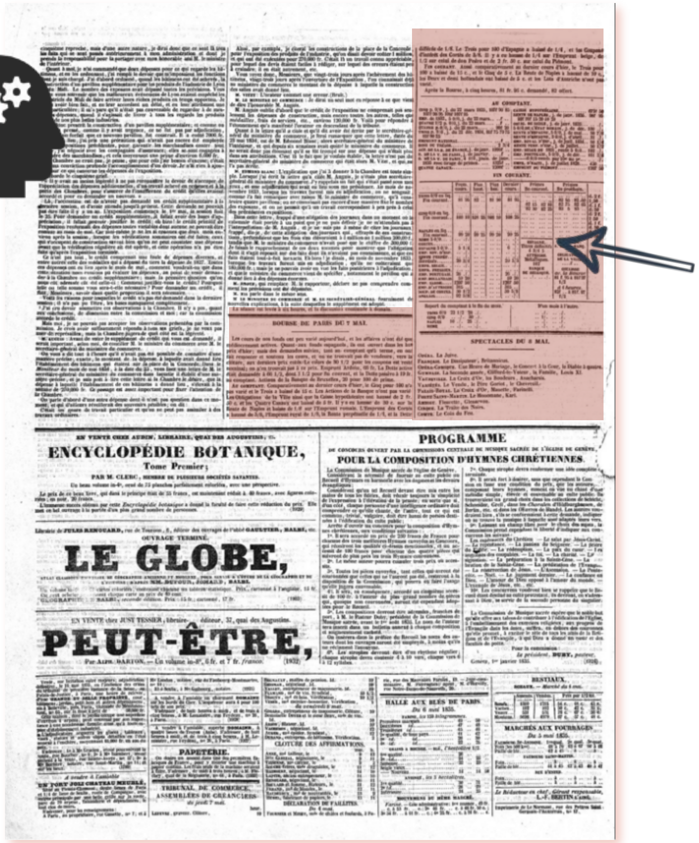
Old News Teaches History

Jean-Philippe Moreux
Bibliothèque national de France,
Digitization dpt

**IFLA News Media Section,
Hamburg, April 2016**

A True Story (@ BnF) about the Researchers' Needs

- How can we help a historian working on Stock Market quotes creation and development in French newspapers? (1800-1870)



BOURSE DE PARIS DU 7 MAI

SPECTACLES DU 8 MAI

ENCYCLOPÉDIE BOTANIQUE,

Tome Premier;

LE GLOBE,

PEUT-ÊTRE,

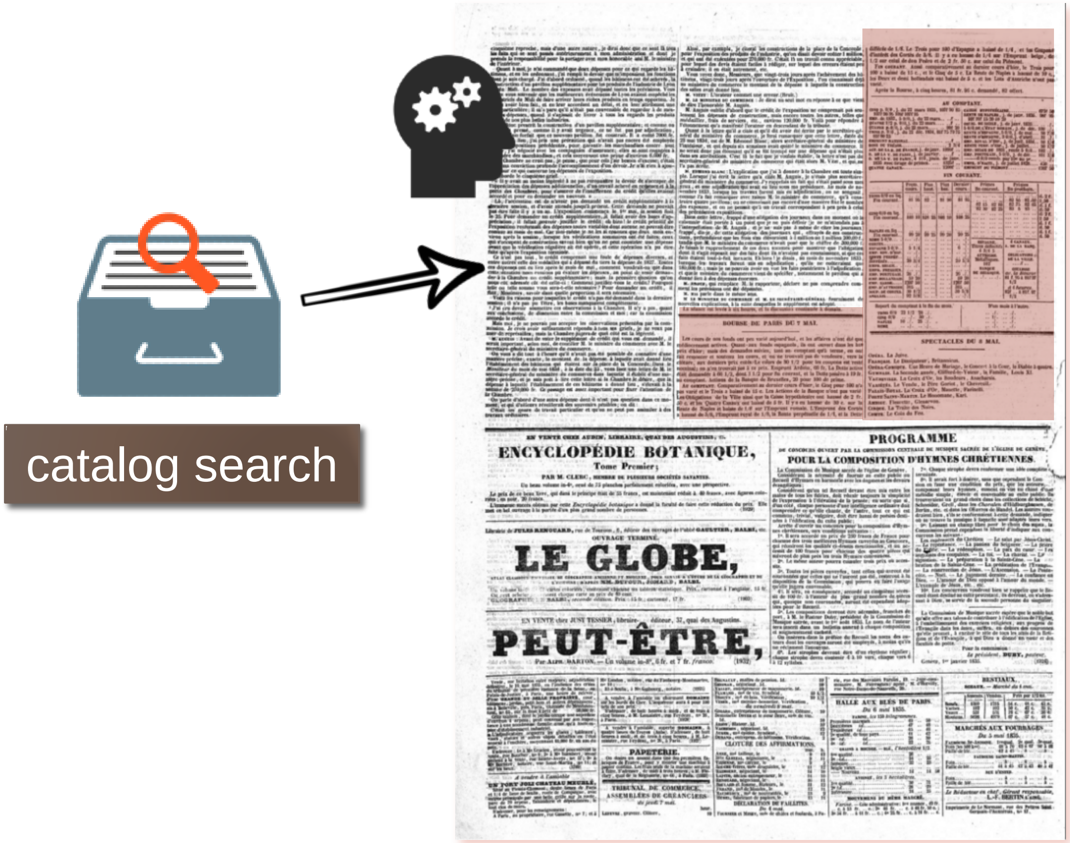
PROGRAMME

POUR LA COMPOSITION D'HYMNES CHRÉTIENNES

here

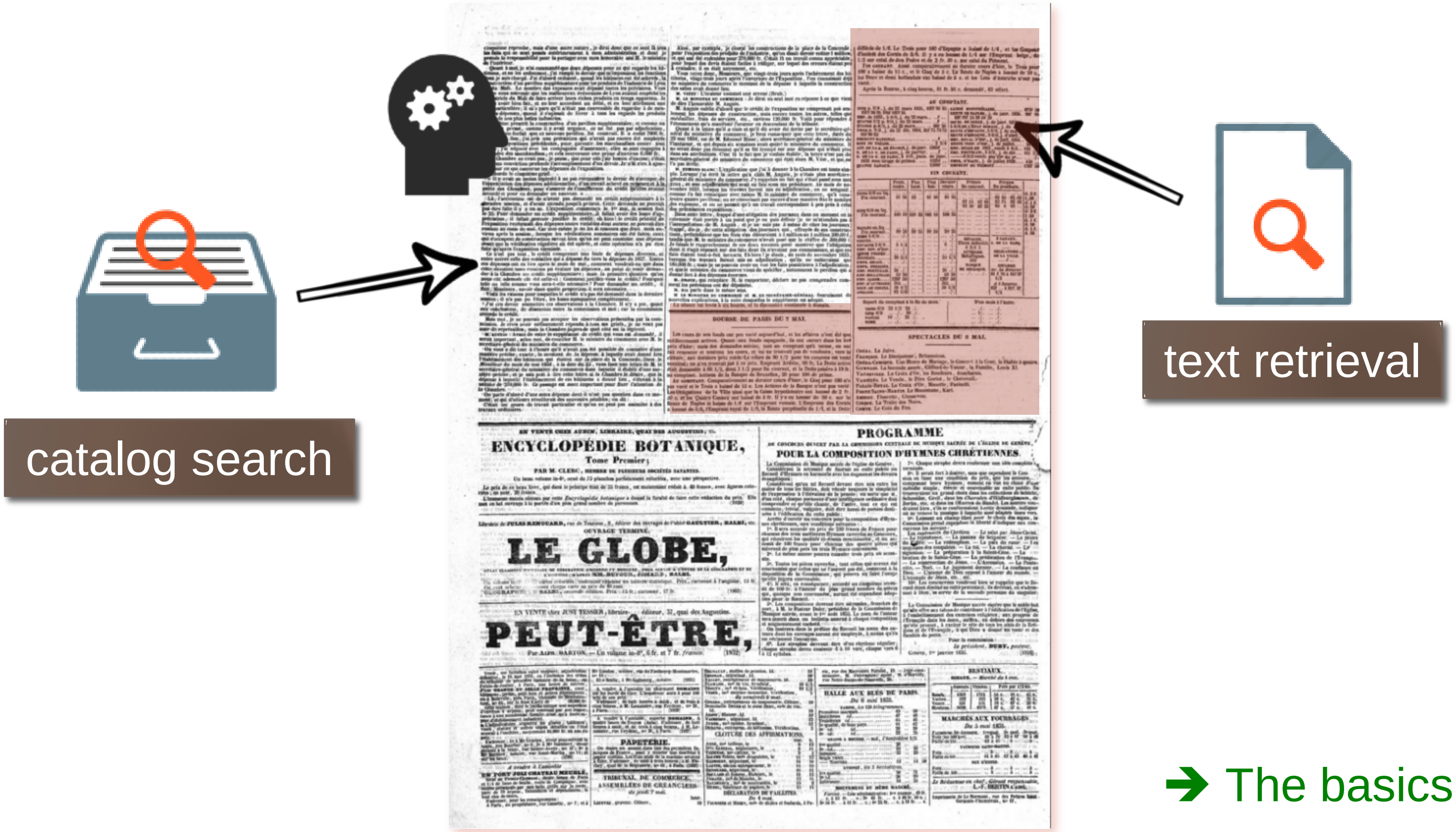
A True Story about the Researchers' Needs

- Obviously, he had to query the digital library catalog.



A True Story about the Researchers’ Needs

- Moreover, he needed a text retrieval functionality.



catalog search

text retrieval

➔ The basics

A True Story about the Needs of Researchers

- But is it enough? Could we do better?

catalog search

text retrieval

+ Corpora **builder**

+ Predefined qualitative & easy-to-use corpora

+ **Advanced query** on document structure and layout (to spot Stock Market regions)

The True Story (cont’d): unhappy Ending

“Stock Market quotes in French Newspapers (1801-1870)”
PhD in Communication and Information Science (P.-C. Langlais)

- **The creation of his corpus was very painful**:
  1. The historian had to script the DL to extract OCR and metadata from multiple newspaper titles.
  2. Then he had to refine/structure his text corpora.

More than 100 Python scripts were needed!

☹

Historians generally prefer to focuse on research, not on writing scripts…

How to Satisfy Scientists' Needs?

Let's try to address this question, regarding the heritage daily corpus enriched during the Europeana Newspapers project:

- Feed the DL with enriched digital documents?
- Give end-users access to quantitative metadata describing documents structure and layout?
- Give end-users an ad hoc corpora builder functionality?

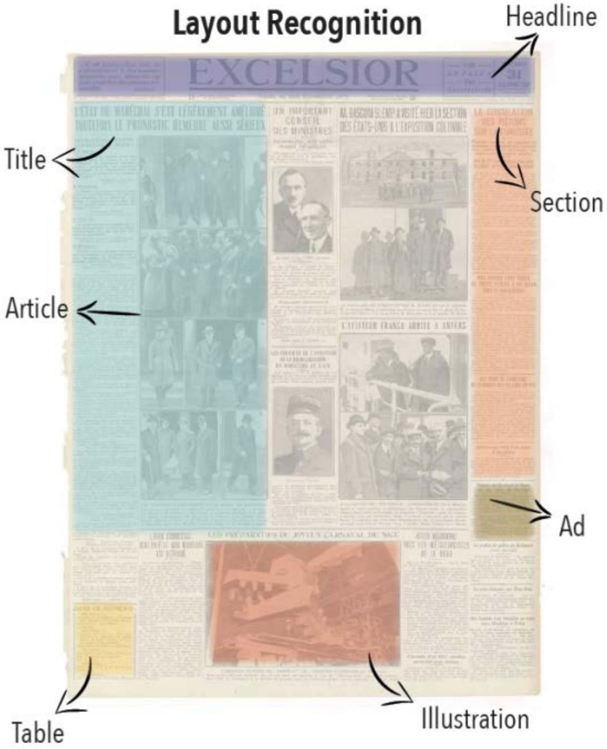
Plan

1. The Europeana Newspapers test bed
2. Building a quantitative metadata dataset
3. Data mining and data visualization use-cases

Enriching Digital Documents

- **Europeana Newspaper project** has enriched and aggregated millions of heritage newspapers pages with advanced refinement techniques like Optical Layout Recognition and Named Entities Recognition.

Europeana Newspapers project (2012-2015): 11,5M OCR'ed pages, 2M OLR'ed pages from 14 European libraries



What is OLR?

- Identification of structural elements, including separation of articles and sections.
- Classification of types of content (ads, offers, obituaries…)

Document Analysis Technique like OLR Produce Quantitative Metadata

The good new is OCR and OLR files are full of interesting objects tagged 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, etc.
  - identification of sections (groups of articles)
  - content types classification (ads, judicial review, stock market…)

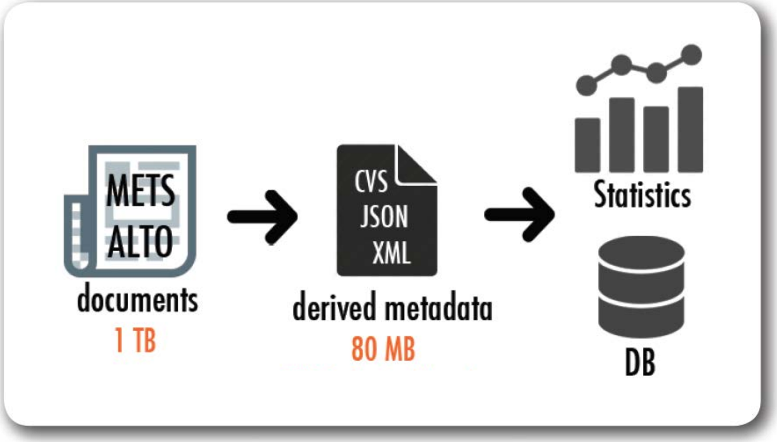
Huge amount of valuable data for historians!

How to Build such Datasets?

- We have to count the number of objects in each page of the collection. Straightforward with XSLT, Java, Python, Perl, etc.
- We have to package and deliver these datasets to end-users.

Europeana Newspapers project / BnF: 880,000 OLR'ed pages from BnF newspapers collection, 6 titles, 1814-1944

{BnF Bibliothèque nationale de France



Pros:

- Give to users light derived datasets, not TB of XML files!
- It's not rocket science.
- It's fast (2-3 h/title with an optimized NoXML parsing script)

No Cons!

Who are the End-Users of the BnF Dataset?

- The EN-BnF dataset includes 5.5 M of values (150K issues, 880K p.)
- 7 metadata at issue level, 5 at page level
- XML, JSON or CSV formats

Researchers (Digital Humanities, History of Press, Information Science)

Digital Curators & Mediators: insights on the collections

Digitization Program Managers: statistics on digitized content

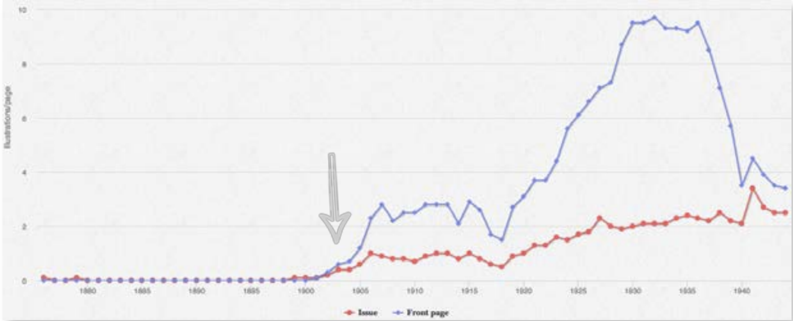
tools

Discovering Knowledge through Visualization

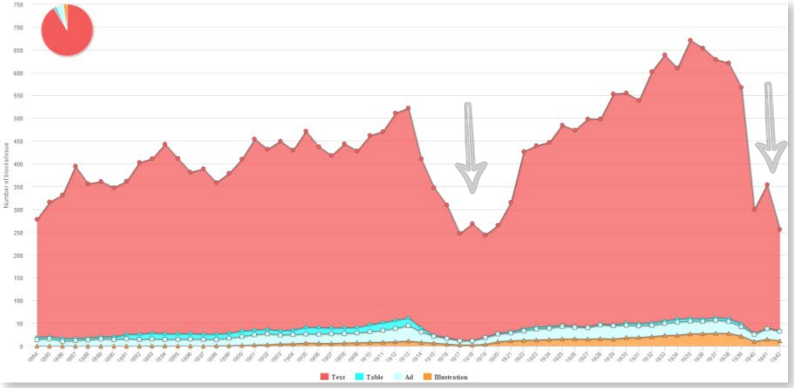
Data visualization allows researchers to discover meaning and information hidden in large volumes of data

tools

- **History of press/illustration:** Dataviz demonstrates the growing importance of illustration (blue: front page, red: inside pages).



- **History of press/activity:** Dataviz of types of content shows the impact of the Great War on the economical activity and assesses the period of return to pre-war level activity (roughly 10 years).



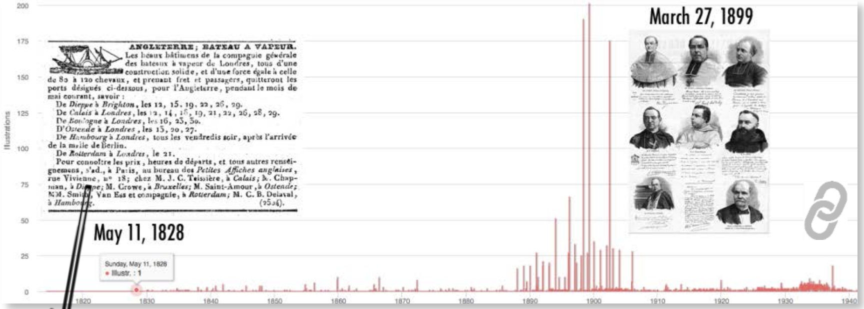
©Highcharts

Engaging new Audiences with Dataviz

Data visualization facilitates rediscovery and reappropriation of heritage documents (by the general public)

tools

- Data visualization of illustrations density can reveal trends or outliers, like highly illustrated issues (illustr. suppl.) or the first published illustration in a title.

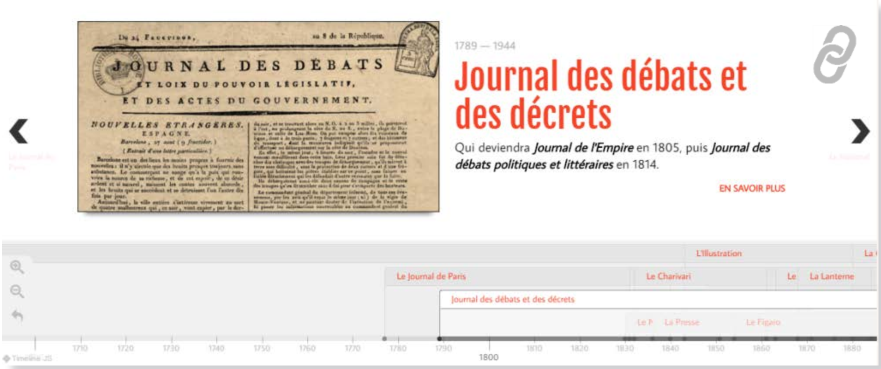


ANGLETERRE; BATEAU A VAPEUR. Les beaux bâtimens de la compagnie générale des bateaux à vapeur de Londres, tous d'une construction solide, et d'une force égale à celle de 80 à 120 chevaux, et prenant fret et passagers, quitteront les ports désignés ci-dessous, pour l'Angleterre, pendant le mois de mai 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 15, 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*. (2534).

Facts extracted thanks to dataviz can then enrich other digital artefacts like **timelines.**



1789 — 1944

Journal des débats et des décrets

Qui deviendra *Journal de l'Empire* en 1805, puis *Journal des débats politiques et littéraires* en 1814.

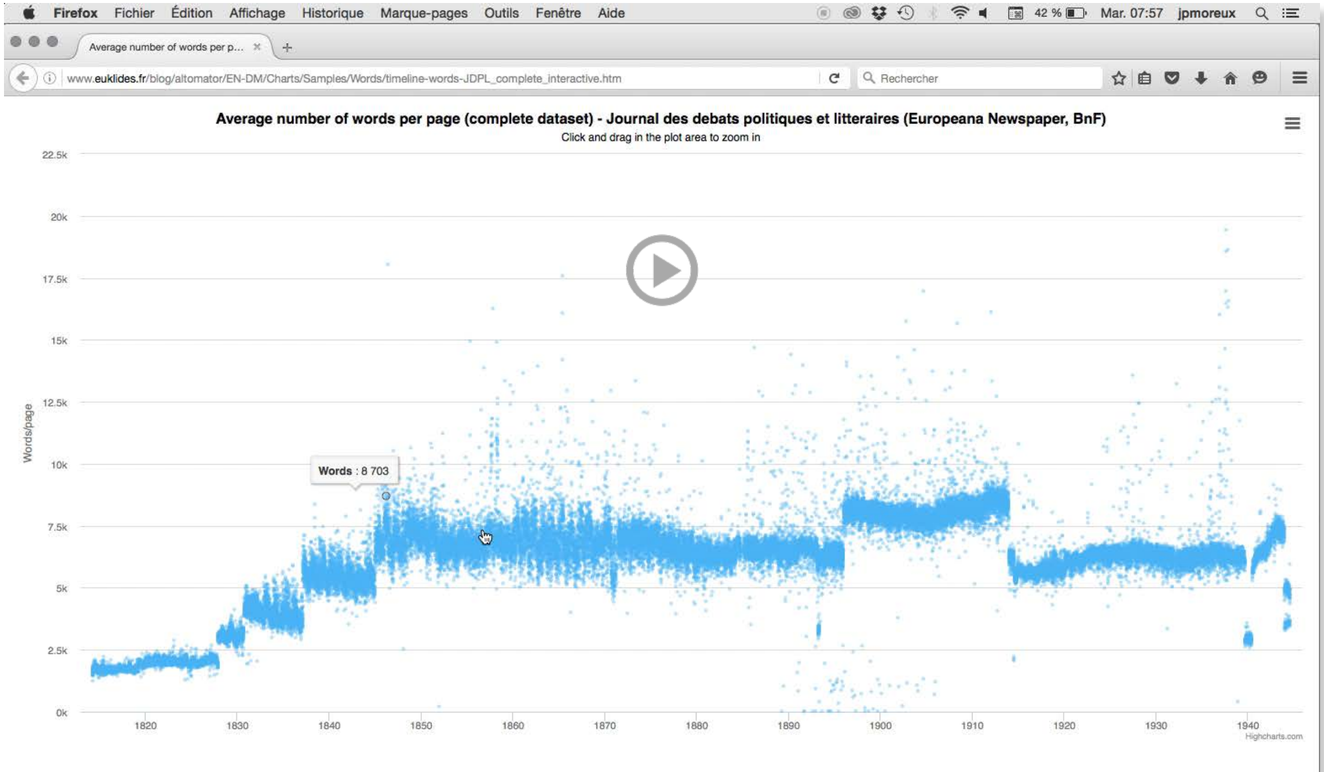
EN SAVOIR PLUS

©www.RetroNews.fr

Engaging new Audiences with Dataviz

Interactive chart of the word density reveals breaks due to changes in layout & paper format, outlier issues…

tools


Firefox
Fichier
Édition
Affichage
Historique
Marque-pages
Outils
Fenêtre
Aide
42 %
Mar. 07:57
jpmoreux
Average number of words per p...
www.euklides.fr/blog/altomator/EN-DM/Charts/Samples/Words/timeline-words-JDPL_complete_interactive.htm
Rechercher
Average number of words per page (complete dataset) - Journal des debats politiques et litteraires (Europeana Newspaper, BnF)
Click and drag in the plot area to zoom in
Words/page
22.5k
20k
17.5k
15k
12.5k
10k
7.5k
5k
2.5k
0k
Words : 8 703
1820
1830
1840
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
Highcharts.com

Journal des débats politiques et littéraires, 1814-1944, 45,334 issues displayed

➔Go beyond keyword spotting and page flip!

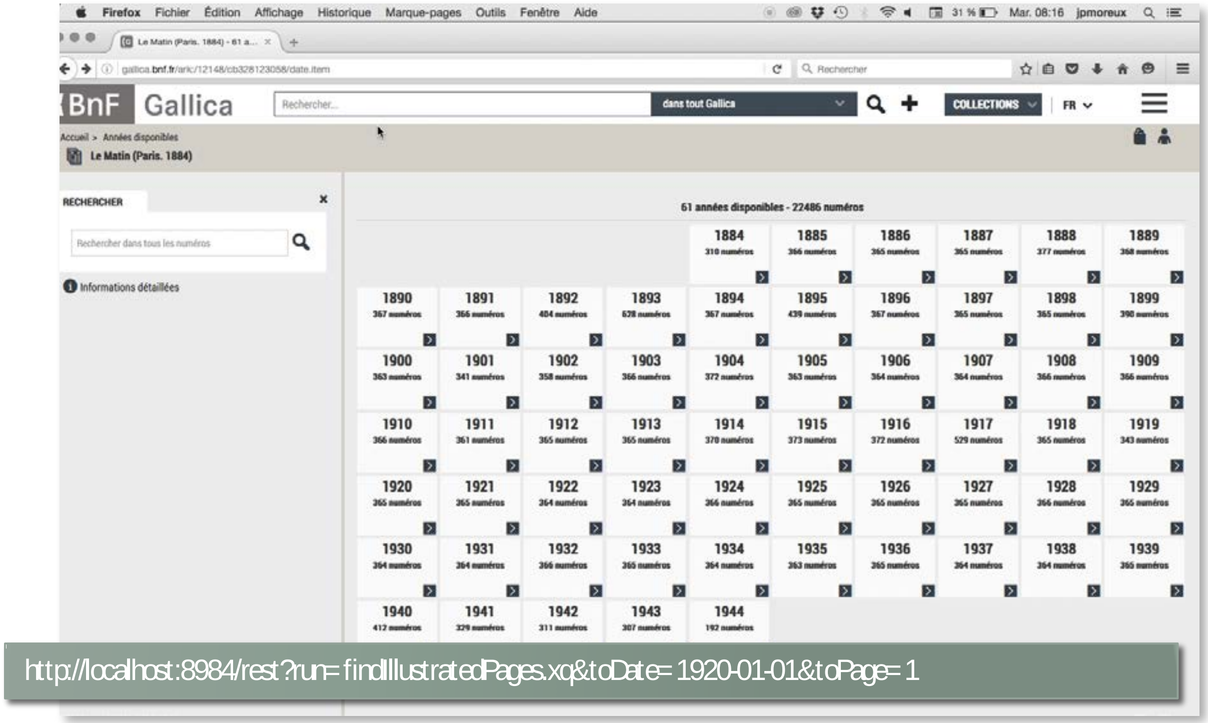
➔Some users would like to play with those charts!

Requesting the Dataset

Those datasets can be requested with dedicated tools (statistical environments, NoSQL or XML databases...)

- **Images search solution used by Gallica Mediation Service:** a XQuery HTTP API identifies “graphical” pages, that is to say both those poor in words and including illustrations.

tools


Firefox Fichier Édition Affichage Historique Marque-pages Outils Fenêtre Aide
Mar. 08:16 jpmoreux
Le Matin (Paris. 1884) - 61 a...
gallica.bnf.fr/ark:/12148/cb328123058/date.item
Rechercher
BnF Gallica
Rechercher...
dans tout Gallica
COLLECTIONS
FR
Accueil > Années disponibles
Le Matin (Paris. 1884)
RECHERCHER
Rechercher dans tous les numéros
Informations détaillées

61 années disponibles - 22486 numéros

				1884 310 numéros	1885 366 numéros	1886 365 numéros	1887 365 numéros	1888 377 numéros	1889 368 numéros
1890 367 numéros	1891 366 numéros	1892 404 numéros	1893 628 numéros	1894 367 numéros	1895 439 numéros	1896 367 numéros	1897 365 numéros	1898 365 numéros	1899 390 numéros
1900 363 numéros	1901 341 numéros	1902 358 numéros	1903 366 numéros	1904 372 numéros	1905 363 numéros	1906 364 numéros	1907 364 numéros	1908 366 numéros	1909 366 numéros
1910 366 numéros	1911 361 numéros	1912 365 numéros	1913 365 numéros	1914 370 numéros	1915 373 numéros	1916 372 numéros	1917 529 numéros	1918 365 numéros	1919 343 numéros
1920 365 numéros	1921 365 numéros	1922 364 numéros	1923 364 numéros	1924 366 numéros	1925 365 numéros	1926 365 numéros	1927 365 numéros	1928 366 numéros	1929 365 numéros
1930 364 numéros	1931 364 numéros	1932 366 numéros	1933 365 numéros	1934 364 numéros	1935 363 numéros	1936 365 numéros	1937 364 numéros	1938 364 numéros	1939 365 numéros
1940 412 numéros	1941 329 numéros	1942 311 numéros	1943 307 numéros	1944 192 numéros					

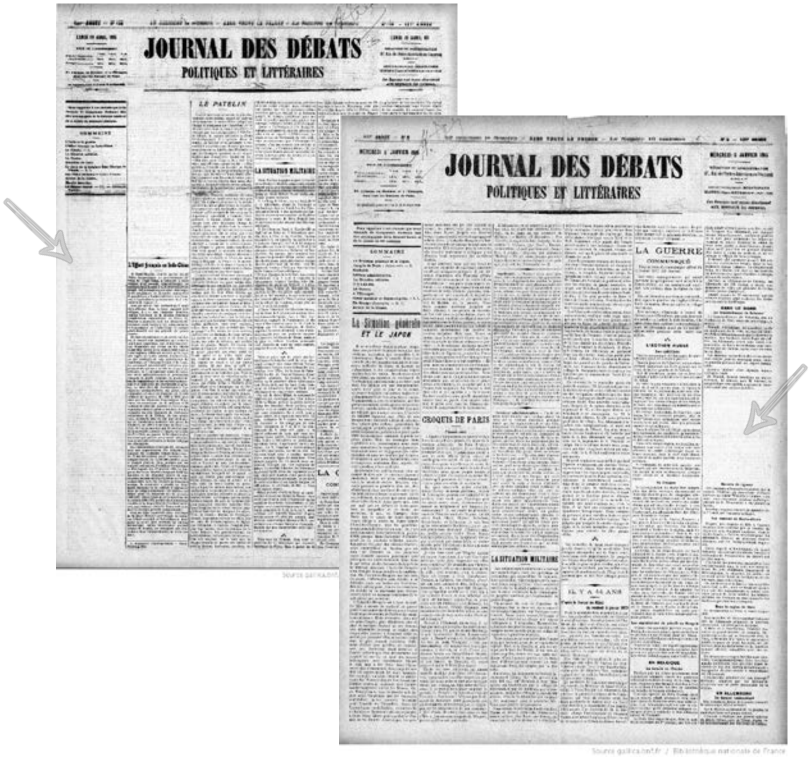
http://localhost:8984/rest?run=findIllustratedPages.xq&toDate=1920-01-01&toPage=1

➔"As a digital mediator, seeking for illustrations in our 12M p. collection is a nightmare…"

Requesting the Dataset

- **Looking for WW1 censored front pages with BaseX:** XQueries can be written to dig into the data and find specific types of content, e.g. the front pages censored during the Great war, which have a slightly smaller words count than the front pages average.

tools



JOURNAL DES DÉBATS
POLITIQUES ET LITTÉRAIRES

JOURNAL DES DÉBATS
POLITIQUES ET LITTÉRAIRES

Source gallica.bnf.fr / Bibliothèque nationale de France

JOURNAL DES DÉBATS
POLITIQUES ET LITTÉRAIRES

Source gallica.bnf.fr / Bibliothèque nationale de France

Is it effective?

- Recall rate: 45%
- Precision rate: 68%

(Based on a ground truth carried on the Journal des Débats front pages for 1915)

➔ Limits of a statistical approach when applied to a word based metric biased by layout singularities. Good enough for mediation.

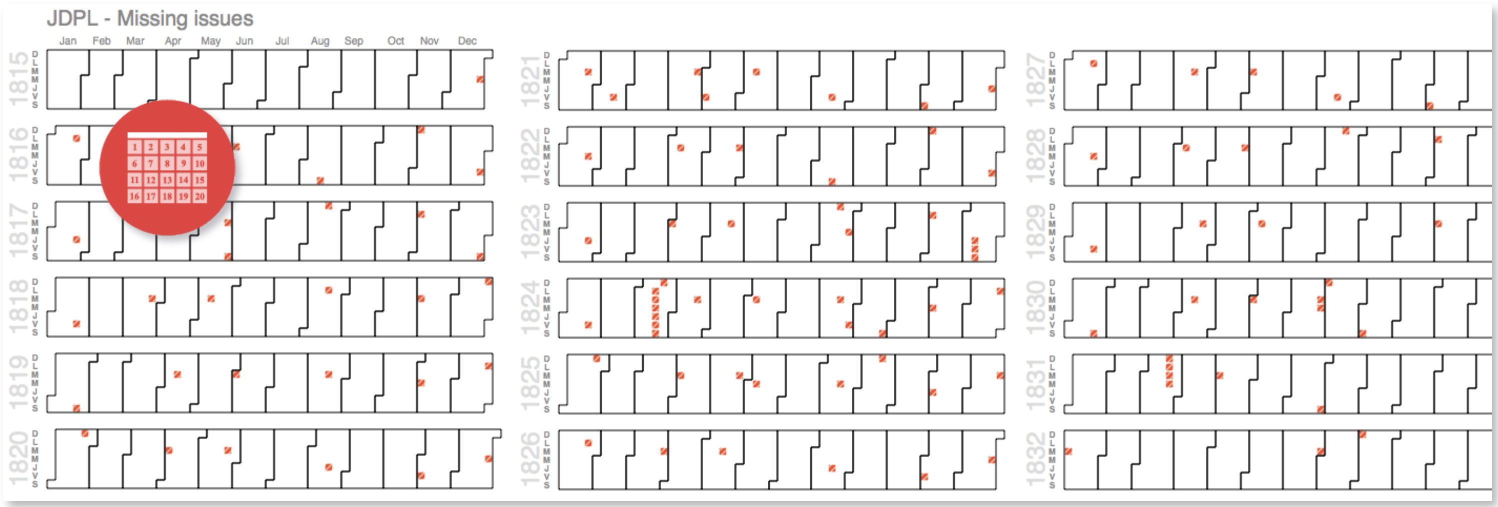
Gallica blog post

Are my Data Representative?

The quality of datasets affects the validity of the analysis and interpretation. Irregular data in nature or discontinuous in time may introduce bias. ➔ QA should be conducted.

Data vizualisation can contribute to quality control (and information of end-users)

- A compact calendar display of a title shows rare missing issues, which suggests that the digital collection is representative.



©Google Charts API

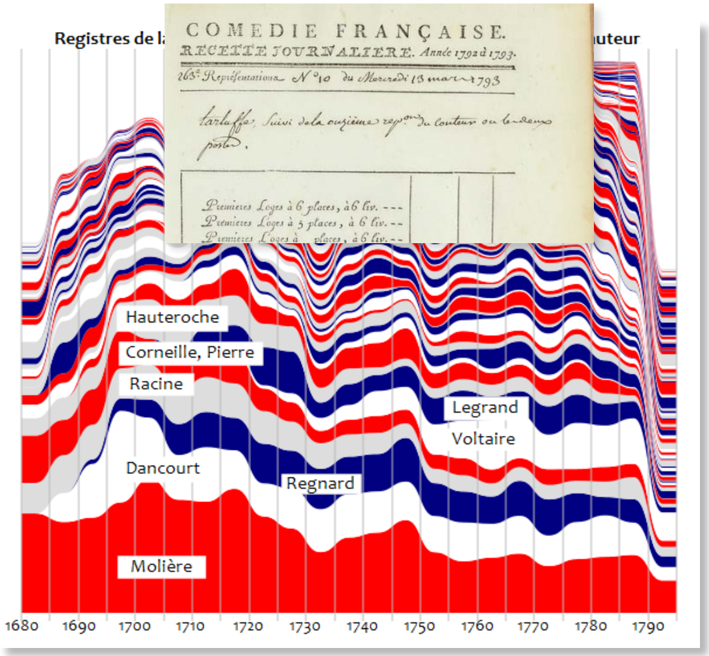
Perspectives

- Apply the same data mining process to the other Europeana Newspapers OLR’ed datasets to produce more datasets. Apply on the on-going BnF newspapers digitization program.
- Automatically build the quantitative metadata datasets.
- Experiment on other types of materials with a temporal dimension (e.g. long life magazines or revues, early printed books).
- @BnF: Assess the opportunity of setting up a data mining framework targeting DH researchers (“Corpus” BnF research project, 2016-2018): Corpora builder? API? OCR dumps? Derived datasets? Remote processing?...

Conclusion

- Quantitative metadata are relevant for all DLs' users: scientists, general public, institution employees
- Only basic data mining & dataviz methods and tools are needed.
- OLR enrichment provides a rich source of information for researchers.
Such data, possibly crossed with the OCRed text, usually provide a fertile ground for research hypotheses.
- Quantitative metadata is sometimes enough to satisfy users. Example of a "pure" quantitative metadata DH project ➔

"The Comédie-Française Registers" project:
From 1680 until 1791, only one theater troupe in Paris was allowed to perform the plays of Molière, Corneille, Racine, Voltaire, Beaumarchais, etc. This troupe played the works of these authors over 34,000 times and kept detailed records of their box office receipts for every single one of those performances. (Partners: Paris-Sorbonne, Harvard University, MIT)



COMEDIE FRANÇAISE.
RECETTE JOURNALIERE. Année 1792 à 1793.
263.e Repréſentation N° 10 du Mercredi 13 mars 1793

tartuffe, Suivi de la onzième rep.on du Conteur ou les deux postes.

Premieres Loges à 6 places, à 6 liv. ---
Premieres Loges à 5 places, à 6 liv. --
Premieres Loges à places, à 6 liv. --

©http://cfregisters.org/fr (Frédéric Glorieux)

Advanced Search in Newspapers?

- Feeding the search engine with layout and structural metadata will allow users to perform **advanced mixed queries**:

```
? illustrated articles
in Trial review section
from 1914 to 1916
where title contains
   “caillaux” or “calmette”
```

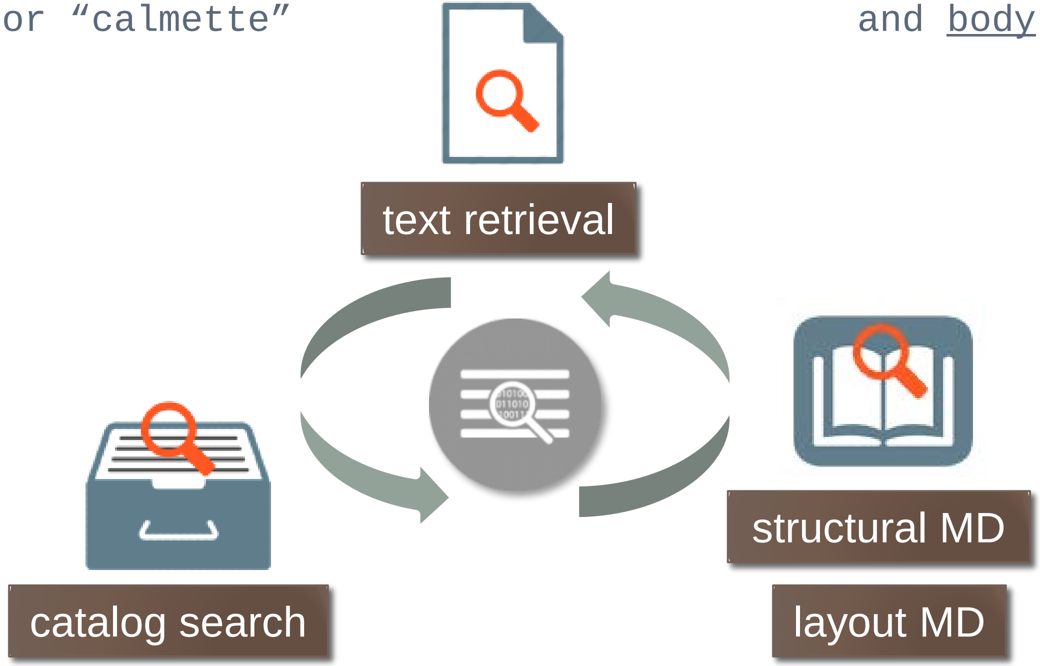
```
? articles with table
in Le Matin
where title contains
    “metal prices”
and body contains “gold”
```

LA JOURNÉE DU VERDICT

Mme Caillaux est acquittée

BRISÉE PAR L'ÉMOTION, ELLE S'ÉVANOUIT
DEUX FOIS AU COURS DE L'AUDIENCE

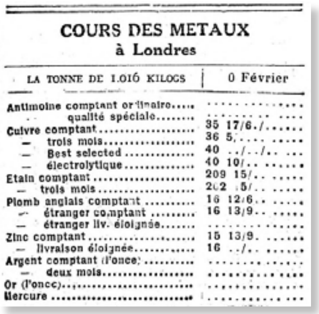
LE RÉQUISITOIRE — LES PLAIDOIRIES


text retrieval

catalog search

structural MD

layout MD



COURS DES METAUX
à Londres

LA TONNE DE 1.016 KILOGS	0 Février
Antimoine comptant ordinaire	
— qualité spéciale	
Cuivre comptant	35 17/6./
— trois mois	36 5,
— Best selected	40 ../../..
— électrolytique	40 10/..
Etain comptant	209 15/..
— trois mois	202 5/..
Plomb anglais comptant	16 12/6..
— étranger comptant	16 13/9..
— étranger liv. éloignée	
Zinc comptant	15 13/9.
— livraison éloignée	16 ../..
Argent comptant (l'once)	
— deux mois	
Or (l'once)	
Mercure	

Advanced Search in Newspapers?

- Feeding the search engine with layout and structural metadata will allow users to perform **advanced mixed queries**:

? illustrated articles
in *Judicial revie*
from 1914 to 1916
where title conta
"caillaux" or

articles with table
Matin
title contains
tal prices"
ly contains "gold"

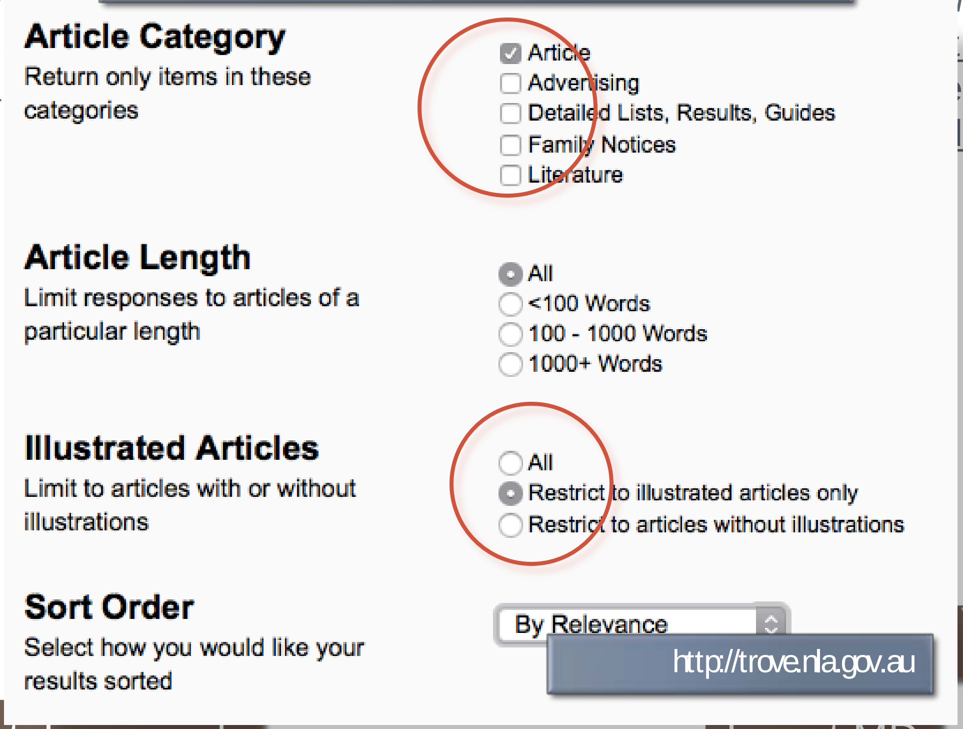
LA JOURNÉE DU VERDICT

Mme Caillaux est acquittée

BRISÉE PAR L'ÉMOTION, ELLE S'ÉVANOUIT DEUX FOIS AU COURS DE L'AUDIENCE

LE RÉQUISITOIRE — LES PLAIDOIRIES

→Trove Advanced Search



Article Category
Return only items in these categories

☑ Article
□ Advertising
□ Detailed Lists, Results, Guides
□ Family Notices
□ Literature

Article Length
Limit responses to articles of a particular length

● All
○ <100 Words
○ 100 - 1000 Words
○ 1000+ Words

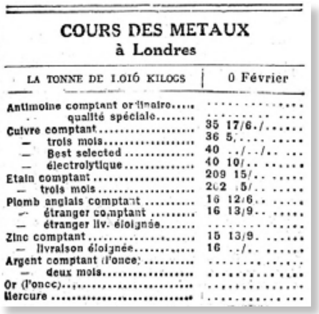
Illustrated Articles
Limit to articles with or without illustrations

○ All
● Restrict to illustrated articles only
○ Restrict to articles without illustrations

Sort Order
Select how you would like your results sorted

By Relevance

http://trove.nla.gov.au



COURS DES METAUX
à Londres

LA TONNE DE 1.016 KILOGS	0 Février
Antimoine comptant ordinaire.....	....
— qualité spéciale........	....
Culvre comptant....................	35 17/6./.......
— trois mois....................	36 5,.... .....
— Best selected	40 ../../.. ...
— électrolytique.............	40 10/.. ......
Etain comptant	209 15/.. .....
— trois mois.....................	202 5/.. .....
Plomb anglais comptant	16 12/6.. .....
— étranger comptant	16 13/9.. .....
— étranger liv. éloignée.......	....
Zinc comptant........................	15 13/9. ......
— livraison éloignée............	16 ../.. ......
Argent comptant (l'once)...........	....
— deux mois....................	....
Or (l'once)..............................	....
Mercure	....

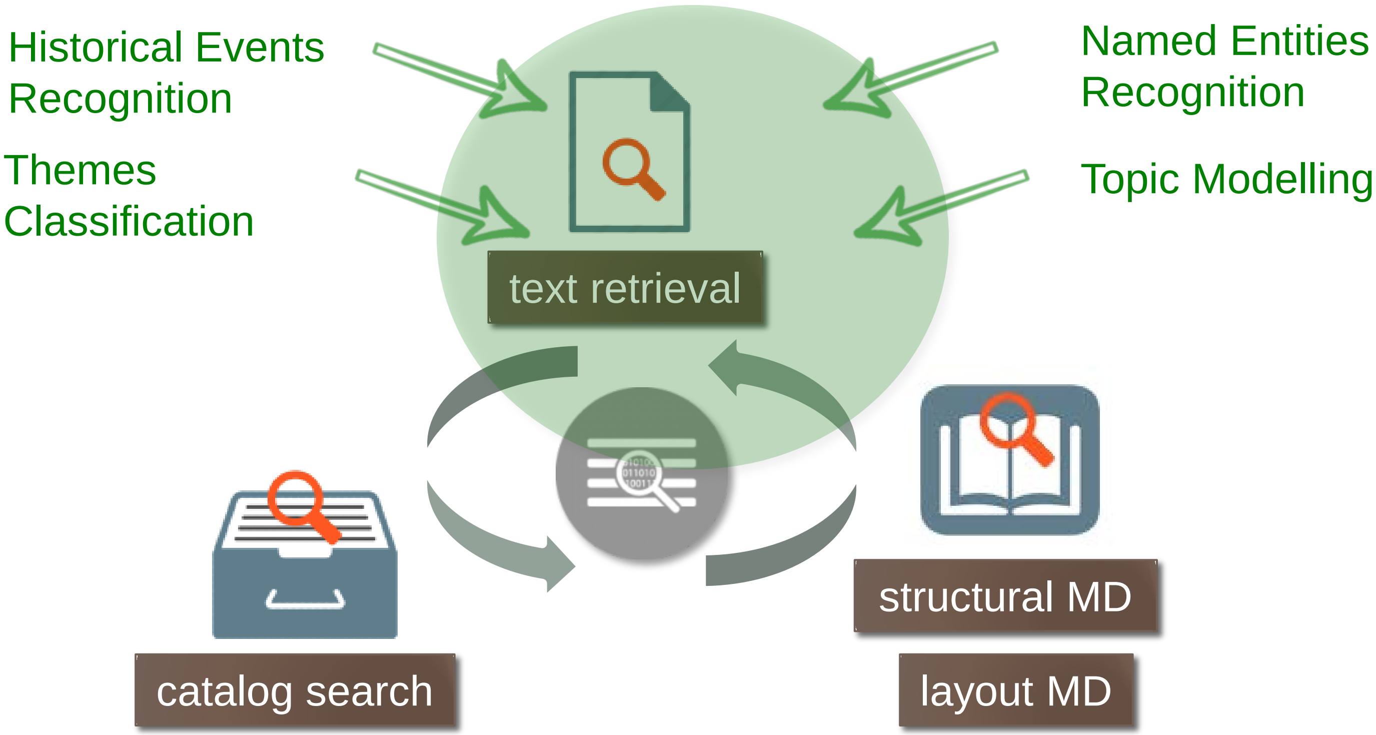
catalog search

layout MD

Advanced Search in Newspapers?

- Adding a pinch of **semantic** flavor to get closer to natural language query:

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**"



Historical Events Recognition

Themes Classification

Named Entities Recognition

Topic Modelling

text retrieval

catalog search

structural MD

layout MD

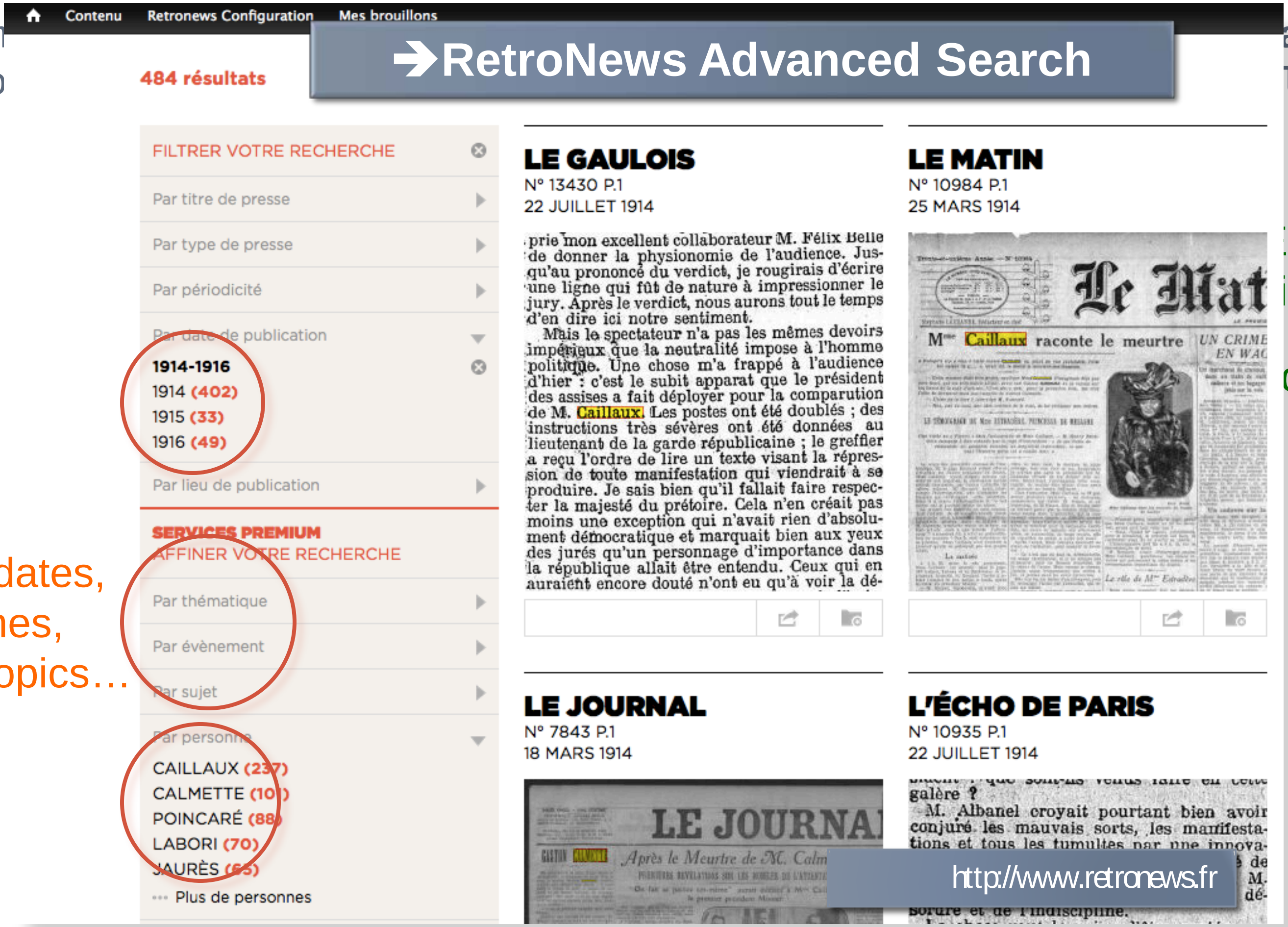
Advanced Search in Newspapers?

- Adding a slice of **semantic** flavor to get closer to natural language query:

I'm
fro
or

al topic
laux”

ntities
ion
delling



🏠 Contenu Retronews Configuration Mes brouillons

➔RetroNews Advanced Search

484 résultats

FILTRER VOTRE RECHERCHE ⊗

Par titre de presse ▸

Par type de presse ▸

Par périodicité ▸

Par date de publication ▾

1914-1916 ⊗

1914 (402)

1915 (33)

1916 (49)

Par lieu de publication ▸

SERVICES PREMIUM

AFFINER VOTRE RECHERCHE

Par thématique ▸

Par évènement ▸

Par sujet ▸

Par personne ▾

CAILLAUX (237)

CALMETTE (101)

POINCARÉ (88)

LABORI (70)

JAURÈS (65)

••• Plus de personnes

LE GAULOIS

N° 13430 P.1

22 JUILLET 1914

prie mon excellent collaborateur M. Félix Belle de donner la physionomie de l'audience. Jusqu'au prononcé du verdict, je rougirais d'écrire une ligne qui fût de nature à impressionner le jury. Après le verdict, nous aurons tout le temps d'en dire ici notre sentiment.

Mais le spectateur n'a pas les mêmes devoirs impérieux que la neutralité impose à l'homme politique. Une chose m'a frappé à l'audience d'hier : c'est le subit apparat que le président des assises a fait déployer pour la comparution de M. Caillaux. Les postes ont été doublés ; des instructions très sévères ont été données au lieutenant de la garde républicaine ; le greffier a reçu l'ordre de lire un texte visant la répression de toute manifestation qui viendrait à se produire. Je sais bien qu'il fallait faire respecter la majesté du prétoire. Cela n'en créait pas moins une exception qui n'avait rien d'absolument démocratique et marquait bien aux yeux des jurés qu'un personnage d'importance dans la république allait être entendu. Ceux qui en auraient encore douté n'ont eu qu'à voir la dé-

LE MATIN

N° 10984 P.1

25 MARS 1914

Le Matin

Mme Caillaux raconte le meurtre

UN CRIME EN WAG

LE JOURNAL

N° 7843 P.1

18 MARS 1914

LE JOURNAL

Après le Meurtre de M. Calm

L'ÉCHO DE PARIS

N° 10935 P.1

22 JUILLET 1914

galère ?

M. Albanel croyait pourtant bien avoir conjuré les mauvais sorts, les manifestations et tous les tumultes par une innova-

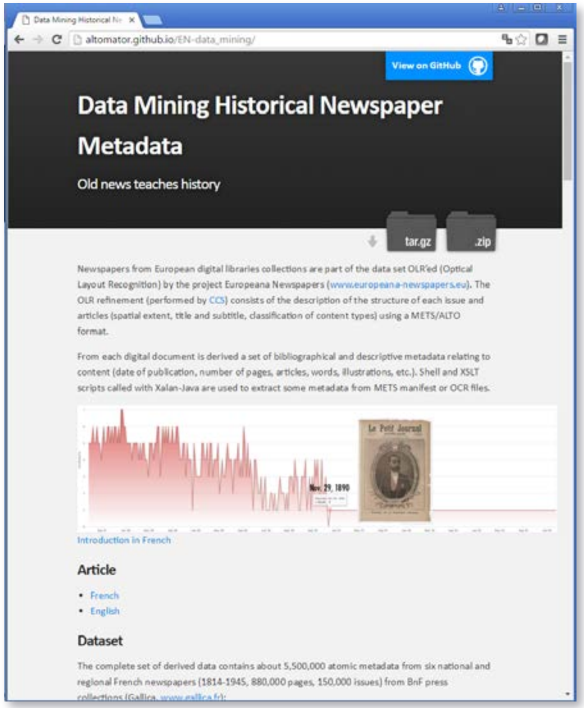
sordre et de l'indiscipline.

http://www.retronews.fr

Faceted search: dates, NE, themes, events, topics…

Thank you for your attention!

- Dataset (CSV, XML, JSON) and charts are publicly available. Just play with it! (no language barrier: not a single word of French inside)


Data Mining Historical Ne
altomator.github.io/EN-data_mining/
View on GitHub
tar.gz
.zip
Le Petit Journal
Nov. 29, 1890

Data Mining Historical Newspaper Metadata

Old news teaches history

Newspapers from European digital libraries collections are part of the data set OLR'ed (Optical Layout Recognition) by the project Europeana Newspapers (www.europeana-newspapers.eu). The OLR refinement (performed by CCS) consists of the description of the structure of each issue and articles (spatial extent, title and subtitle, classification of content types) using a METS/ALTO format.

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 are used to extract some metadata from METS manifest or OCR files.

Introduction in French

Article

- French
- English

Dataset

The complete set of derived data contains about 5,500,000 atomic metadata from six national and regional French newspapers (1814-1945, 880,000 pages, 150,000 issues) from BnF press collections (Gallica, www.gallica.fr):

http://altomator.github.io/EN-data_mining

europeana
newspapers

Thanks to all
the EN partners!

ICT PSP

Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

universität innsbruck

CCS

This project runs from February 2012 to February 2015. It is led by the Staatsbibliothek zu Berlin and co-funded by the European Commission under the Competitiveness and Innovation framework Programme. http://ec.europa.eu/ict_psp

The True Story (cont'd)

- Could we have helped thim?

tools

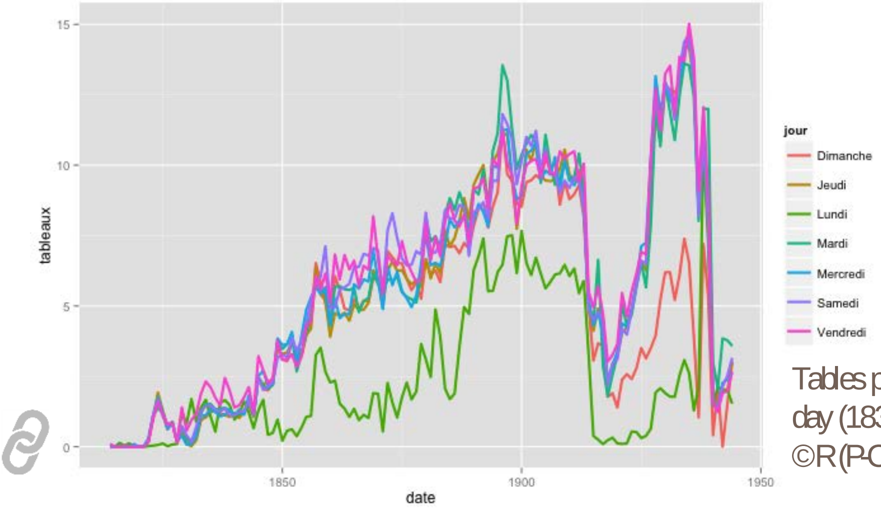
OLR facilitates the corpus creation task ☺

➔ Content Types classification, Section identification

The quantitative dataset is of a great help ☺

"Tables" in newspapers are predominantly used in Stock Market quotes ➔ instant use of this metadata!

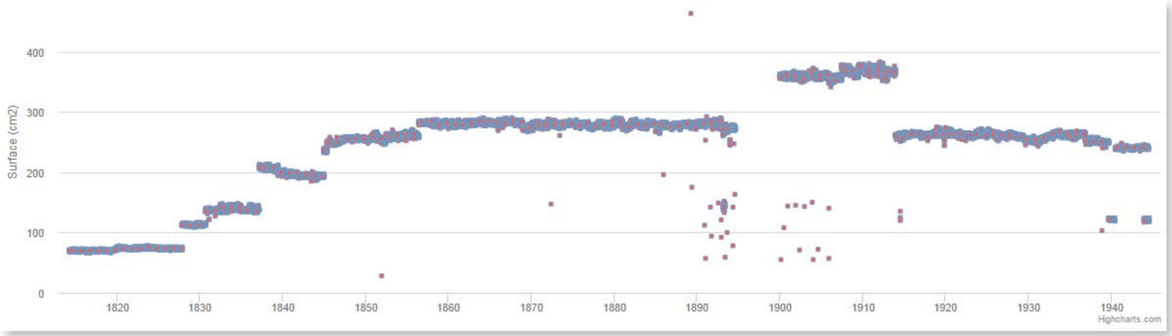
Types of content are tagged



Tables per week day (1838-1870) ©R (P-C Langlais)

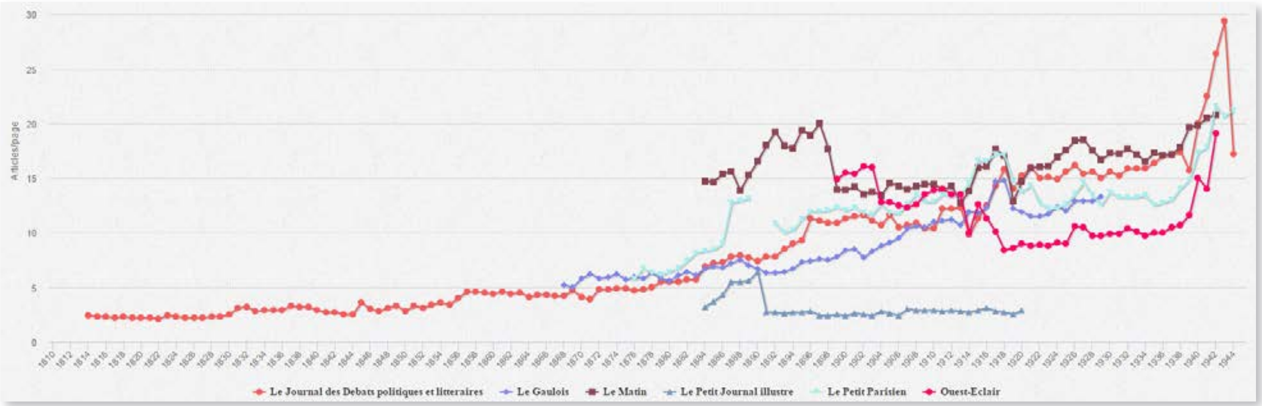
Discovering Knowledge through Visualization

- **History of press/page format:** Digital archeology of papermaking and printing.



tools

- **History of press/layout:** Visualization of the articles density per page reveals the shift from XVIIth “gazettes” to modern daily.



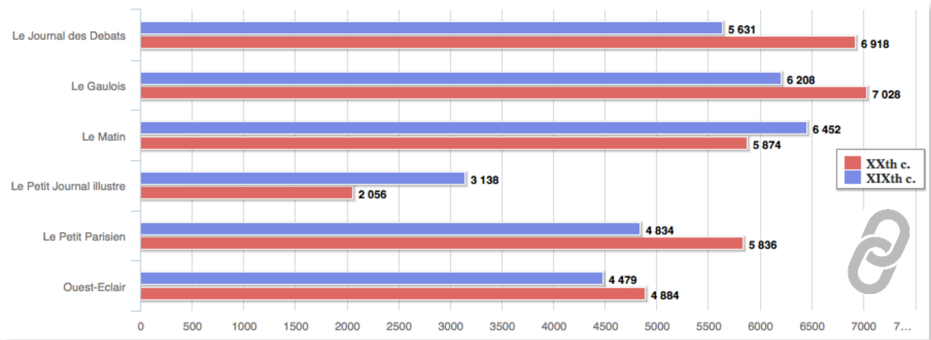
©Highcharts

Other Users might be Interested by those Metadata: Digitization dpt

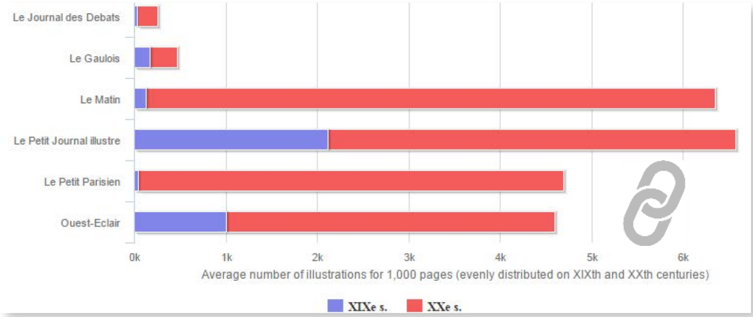
tools

Statistical information on digitized content for project managers.

- **OCR Crowdsourcing project:** What is the average density in words of these documents? What text correction efforts will be required?



- **Image bank**: What titles contain illustrations? What is the total number of images one can expect?



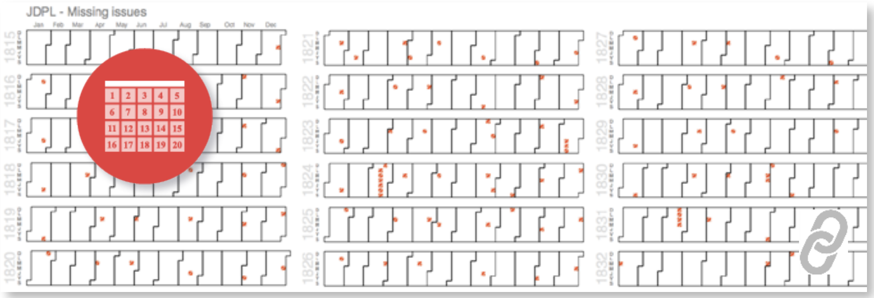
©Hghcharts

Are my Data Representative?

The quality of datasets affects the validity of the analysis and interpretation. Irregular data in nature or discontinuous in time may introduce bias. ➔A qualitative assessment should be conducted.

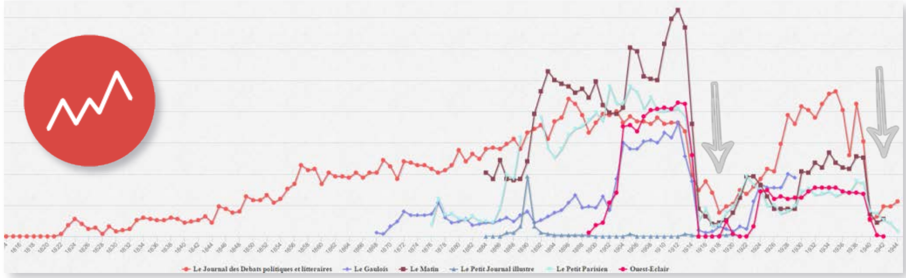
Data vizualisation can contribute to quality control (and information of end-users)

- A calendar display of a title data shows rare missing issues, which suggests that the digital collection is representative.



©Google Charts API

- Stock Market quotes study based on the content tagged “table”: one can empirically validate this hypothesis by the sudden inflections recorded in 1914 and 1939 for all titles, being known and established the historical fact of the virtual halt of trading during the two World Wars



©Highcharts