INNOVATIVE APPROACHES OF HISTORICAL NEWSPAPERS: DATA MINING, DATA VISUALIZATION, SEMANTIC ENRICHMENT

Facilitating Access for various Profiles of Users

Jean-Philippe Moreux, Caroline Kageneck
Bibliothèque national de France

IFLA News Media Section
Lexington, August 2016

A True Story about the Researchers’ Needs

- How can we help historian working on the creation and the development of Stock Market’s section in French newspapers? (1800-1870)

PROGRAMME
ENCYCLOPÉDIE BOTANIQUE,
LE GLOBE,
PEUT-ÊTRE,

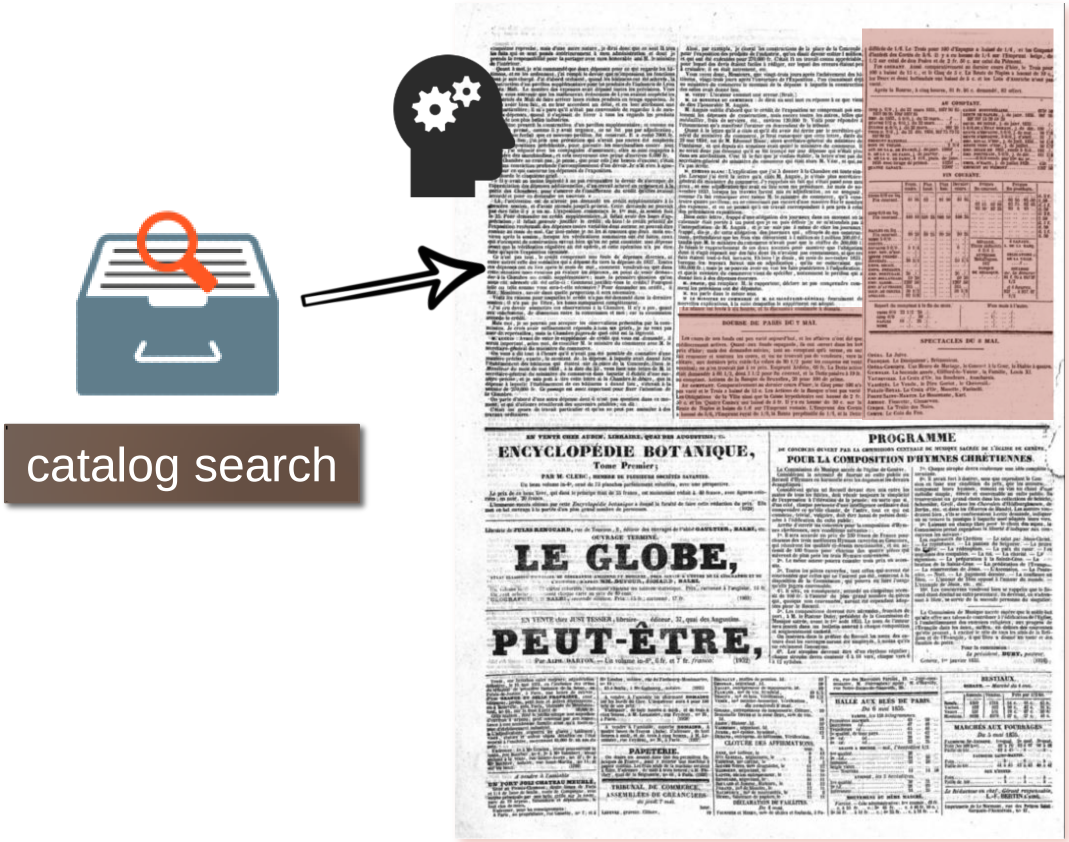
Here...

INSTITUT DENTAIRE
CRÉDIT
SOMNOL
OBÉSITÉ
POURQUOI ?
LES PAQUETS "PAC"
recueillent les BISCUITS PERNOT au sortir du FOUR
et les transmettent intacts au Consommateur
AU BUCHERON
PARIS - 8 & 10 RUE DE RIVOLI - PARIS
MEUBLES - LITERIE
TAPIS
TAPISSERIE - MÉNAGE

and here

A True Story about the Researchers’ Needs

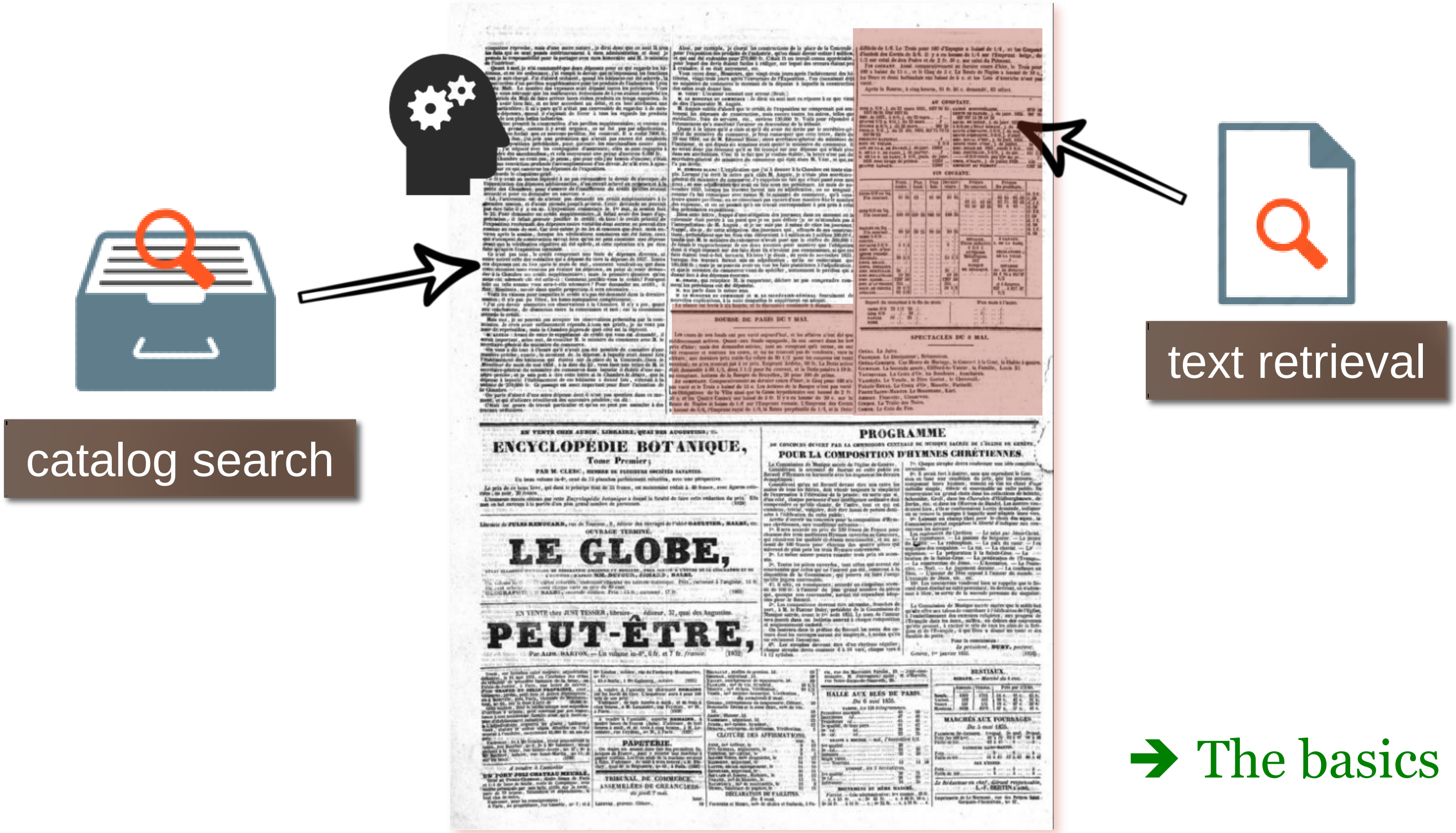
- Obviously, he has to request the digital library catalog.



catalog search

A True Story about the Researchers' Needs

- And he needs a text retrieval functionality.



catalog search

text retrieval

➔ The basics

A True Story about the Needs of Researchers

- But is it enough? Shouldn't we do better?

catalog search

ENCYCLOPÉDIE BOTANIQUE,

LE GLOBE,

PEUT-ÊTRE,

PROGRAMME

POUR LA COMPOSITION D'HYMNES CHRÉTIENNES

text retrieval

+ Corpora Builder

+ Predefined qualitative and easy-to-use corpora

+ Advanced query on document Structure and Layout (to spot Stock Market section)

How to Satisfy Researchers' (and other DLs' Users) Needs?

Let's try to address this issue and let's focus on two questions. Should we:

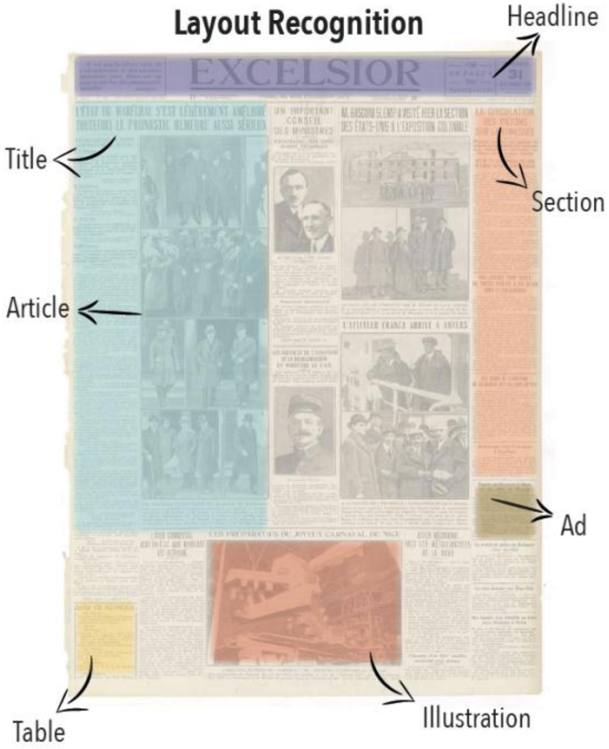
- Give end-users access to quantitative metadata describing documents structure and layout?
- Feed our DLs with semantically enriched documents?

Plan

1. The Europeana Newspapers Test Bed: Data Mining Quantitative Metadata of Historical Newspapers
2. The RetroNews portal: Semantic Enrichment of Newspapers at Large Scale
3. Conclusion

EN: 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.



What is OLR?

Identification of structural elements, like articles and sections.

Classification of types of content (ads, offers, novel, stock market…)

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

Document Analysis Technique like OLR Produce Quantitative Metadata

The good news is that OCR and OLR files are full of interesting objects marked up in 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 quantitative metadata on high level informational objects:
  - Articles, sections (group of articles), titles, etc.
  - Content types (ads, judicial review, stock market…)

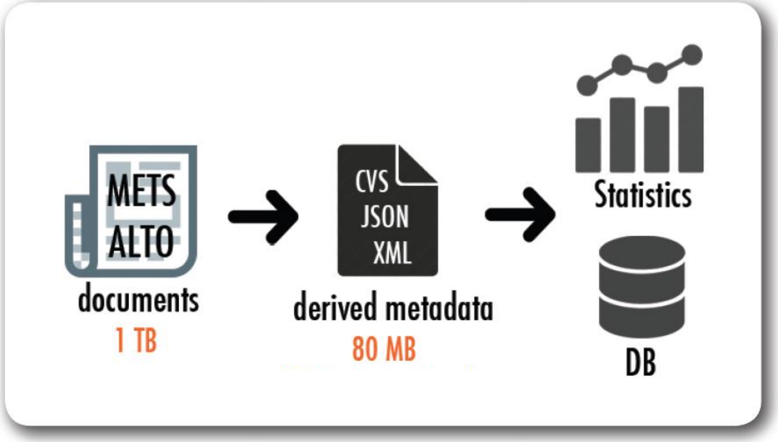
Huge amount of valuable data for historians!

How to Build such Datasets?

- We have to count the number of objects in each page of the collection (7 metadata at issue level, 5 at page level).
- We need to package and deliver these datasets to end-users (XML, JSON and CSV formats).

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

{BnF Bibliothèque nationale de France



Pros:

- Rich dataset: 5,5 M of metadata
- Give to users light derived datasets, not TB of XML files!
- No copyright issue (no text, no image)
- It's not rocket science. It's fast (optimized NoXML parsing script)

No Cons!

Who are the End-Users of such Metadata?

Various profiles of users will benefit from this quantitative metadata, inside and outside the digital library.

General Public

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

Digital Curators & Mediators

Digitization Program Managers

Who are the End-Users of such Metadata?

Various profiles of users will benefit from this quantitative metadata, inside and outside the digital library.

DL Users

General Public

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

Digital Curators & Mediators

Digitization Program Managers

Who are the End-Users of such Metadata?

Various profiles of users will benefit from this quantitative metadata, inside and outside the digital library.

General Public

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

DL Professionals

Digital Curators & Mediators

Digitization Program Managers

Digitization or IT People might be Interested by those Metadata

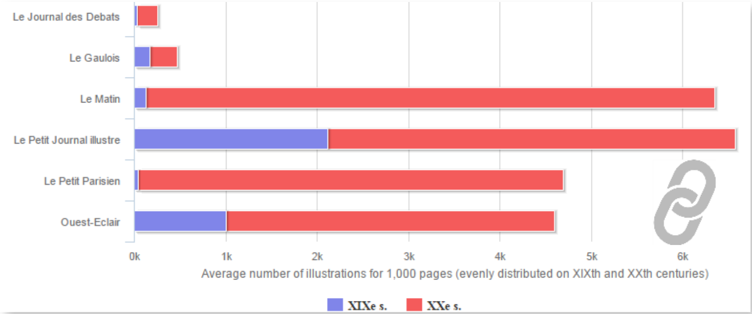
tools

Statistical information on digitized content for project managers and IT teams.

- **Automatic correction of noisy OCR:** What is the average density in words of the press collection ?
  - *x* digital documents/day received by the DL from its service providers, *y* pages/document, *z* words/page
  - ➢ *n* words per day to process
  - ➢ **computing power to provide for a real-time post-processing of the OCR**

➔ Better knowledge of the collection

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



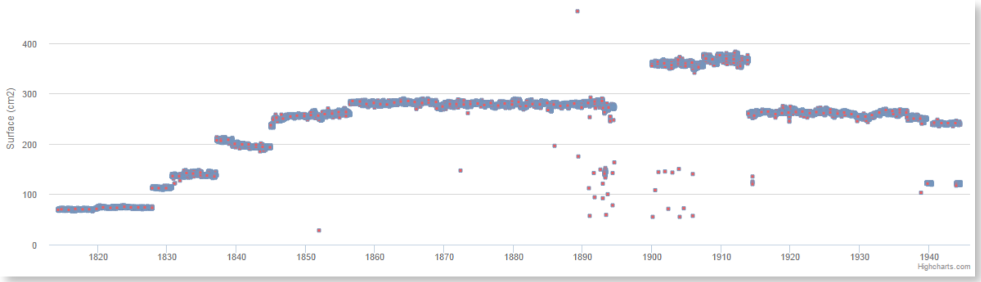
© Highcharts

Discovering Knowledge through Visualization

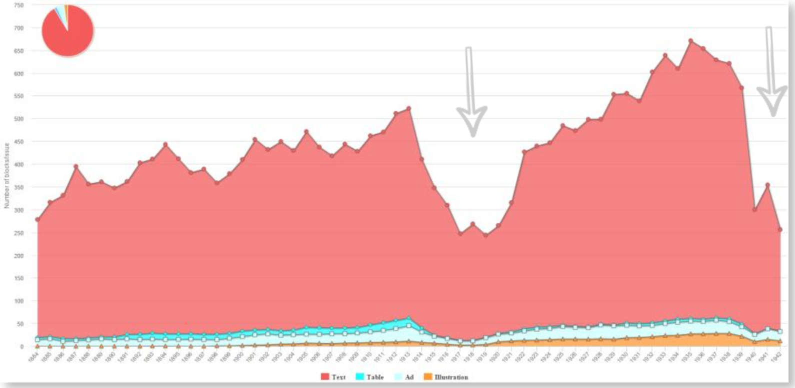
Data visualization allows researchers to discover meaning and information hidden in large volumes of data

tools

- **History of press/page format:** Digital archeology of papermaking and printing. Source: page dimension@OCR



- **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).
  Source: types of content@OLR



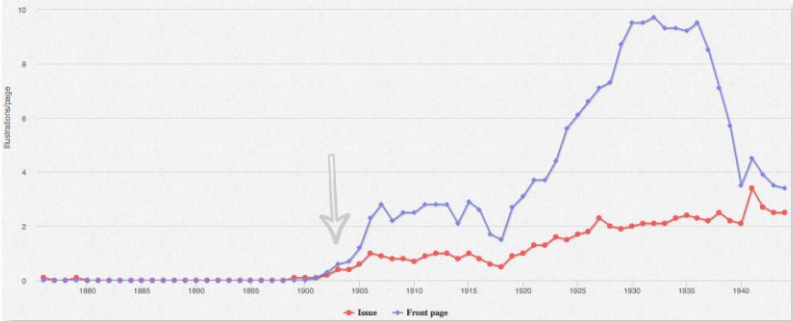
© Highcharts

Discovering Knowledge through Visualization

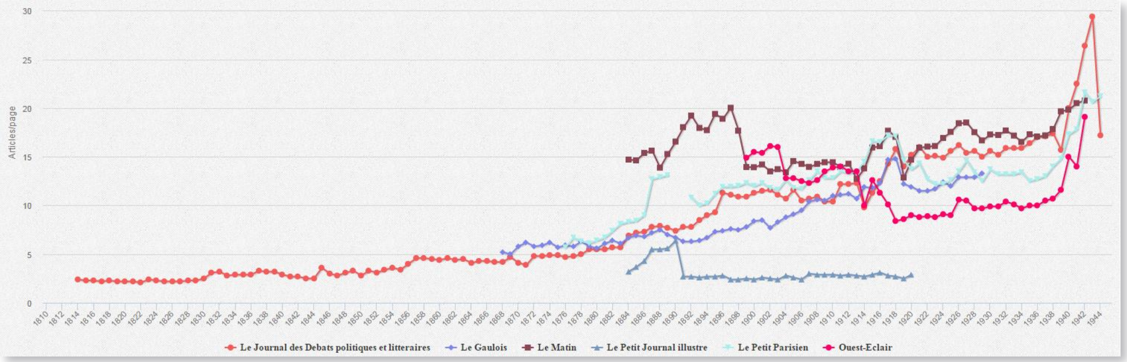
Data visualization allows <u>researchers</u> to discover meaning and information hidden in large volumes of data

tools

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



- **History of press/layout:** Visualization of the articles density per page reveals the shift from XVIIth "gazettes" to modern dailies and their complex layout. Source: articles count@OCR



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 (Gallica) to extract text 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…

The True Story (cont’d): Could we Have Helped him?

OLR facilitates the corpus creation task ☺

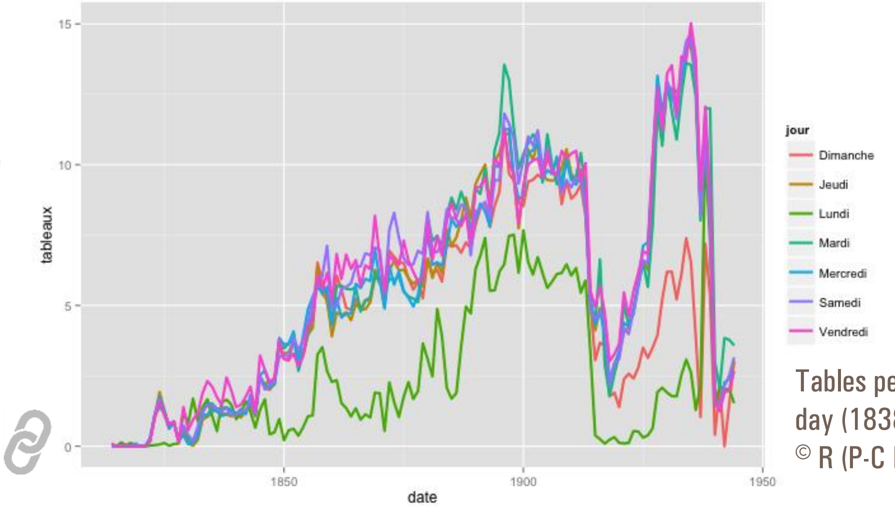
➔ Content Types classification, Section identification

Types of content are tagged, like Stock Market

The quantitative dataset is of a great help ☺

tools

“Tables” in newspapers are predominantly used in Stock Market section ➔ **instant** use of this metadata!



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

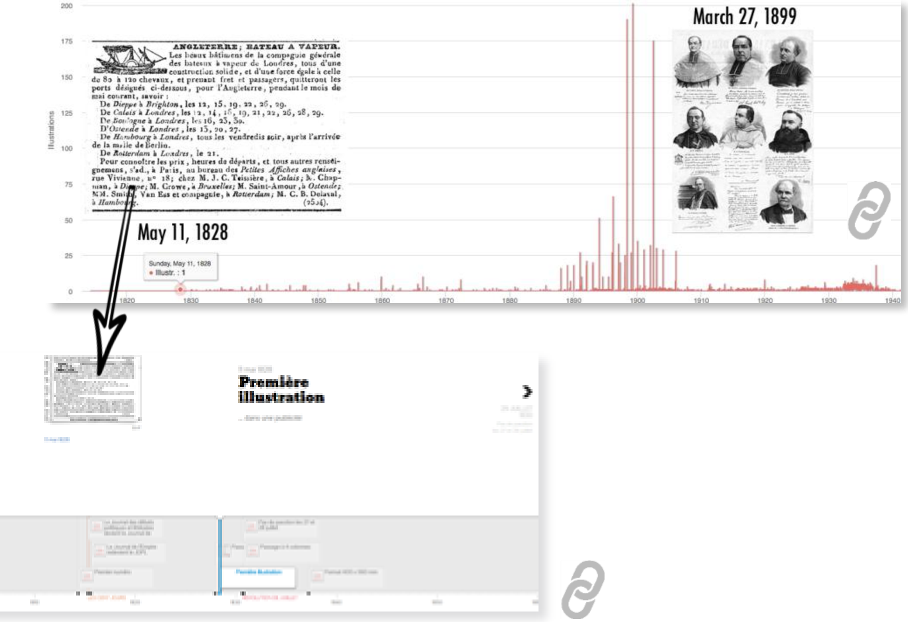
➔ Easy to make graphs (from the 80 Mb high level dataset, compared to 1Tb of raw XML OCR)

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.



March 27, 1899

May 11, 1828

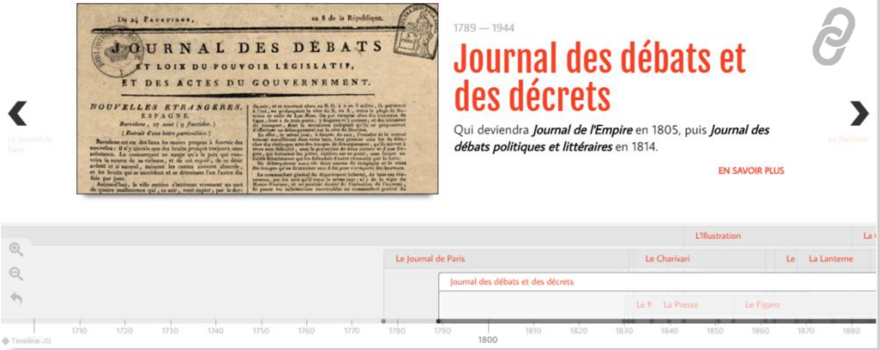
ANGLETERRE; BATEAU A VAPEUR. Les beaux bâtimens de la compagnie générale des bateaux à vapeur de Londres, tous d'une construction solide, et d'une force égale à celle de 80 à 120 chevaux, et prenant fret et passagers, quitteront les ports désignés ci-dessous, pour l'Angleterre, pendant le mois de mai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

© https://timeline.knightlab.com

Facts extracted thanks to dataviz can then enrich other digital objects 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

Du 24 Fructidor, an 8 de la République.

JOURNAL DES DÉBATS ET LOIX DU POUVOIR LÉGISLATIF, ET DES ACTES DU GOUVERNEMENT.

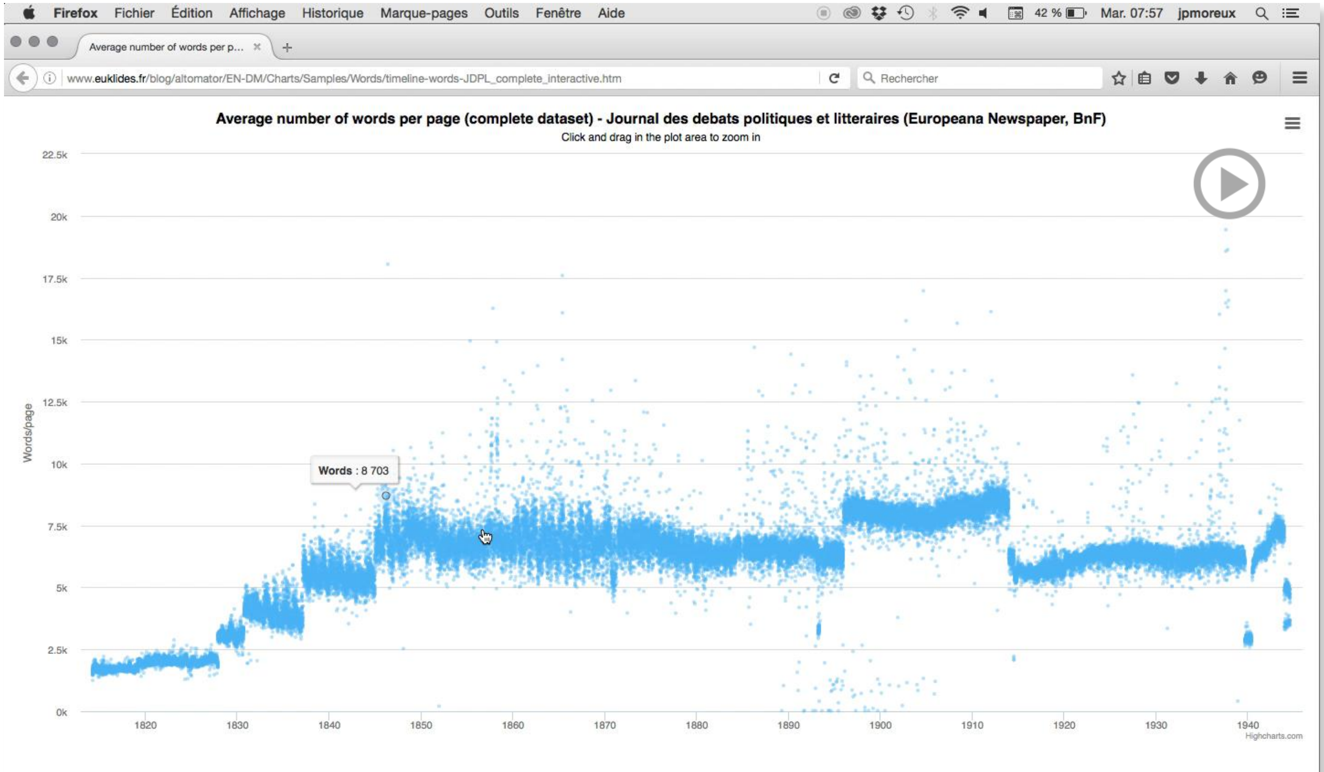
NOUVELLES ETRANGERES. ESPAGNE. *Barcelone, 27 août (9 fructidor.)* (*Extrait d'une lettre particulière.*)

© 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 debats politiques et litteraires, 1814-1944, 45,334 issues displayed

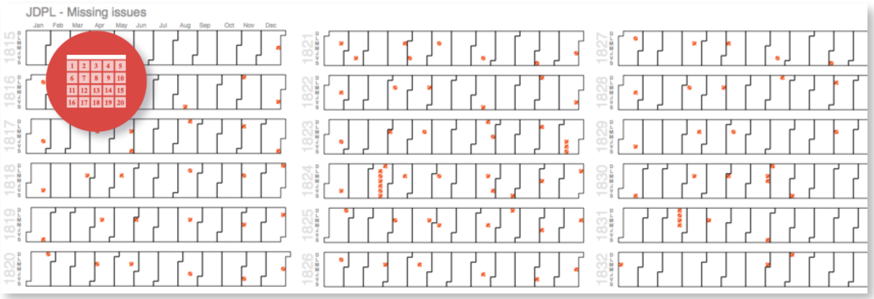
➔ Go beyond keyword spotting and page flip!

➔ Some users would like to play with those charts!

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.

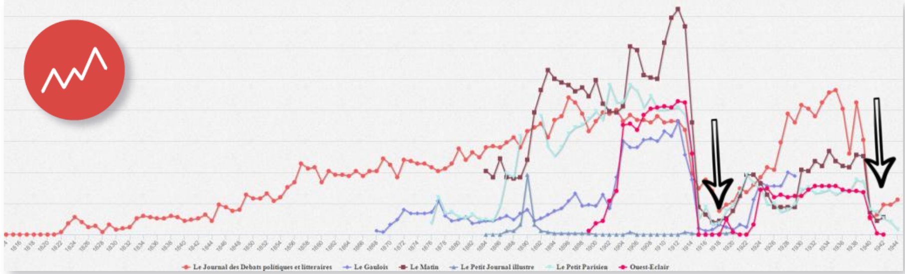
- A calendar display of a title data shows rare missing issues, which suggests that the digital collection is representative.



© Google Charts API

➔ "Inform end users about data quality!

- 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.



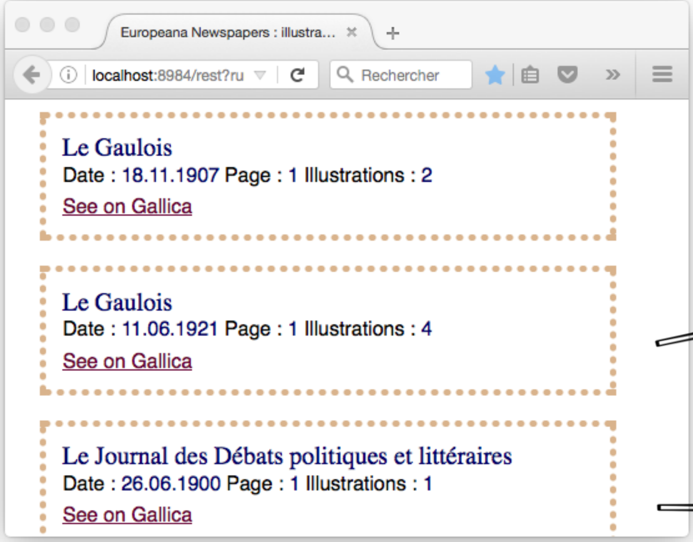
© Highcharts

Querying the Dataset

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

tools

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



Europeana Newspapers : illustra...

localhost:8984/rest?ru

Rechercher

Le Gaulois
Date : 18.11.1907 Page : 1 Illustrations : 2
See on Gallica

Le Gaulois
Date : 11.06.1921 Page : 1 Illustrations : 4
See on Gallica

Le Journal des Débats politiques et littéraires
Date : 26.06.1900 Page : 1 Illustrations : 1
See on Gallica

➔ "As a digital mediator, seeking for illustrations in our *x*M pages newspaper collection is a nightmare..."

➔ Let’s remember that newspapers are full of illustrations not described in our catalogs!

http://localhost:8984/rest?run= findIllustratedPages.xq&fromDate=1900-01-01&toPage=1

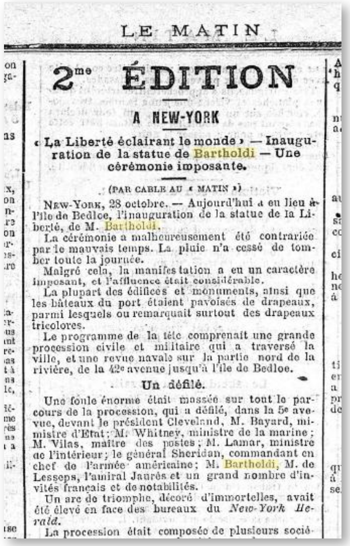
Advanced Search Mode for Newspapers

tools

- Feeding the Search Engine with layout and structural metadata allows DL users to perform **advanced mixed queries**:

? **articles** from 1886 where
title OR **subtitle** contains
"bartholdi" OR "statue" & "liberté"

? **articles** with **table**
in *Le Matin* where **title** contains
"metal prices" and **body** contains
"gold"



LE MATIN

2me ÉDITION

A NEW-YORK

« La Liberté éclairant le monde » — Inauguration de la statue de Bartholdi — Une cérémonie imposante.

(PAR CABLE AU « MATIN »)

New-York, 28 octobre. — Aujourd'hui a eu lieu à l'île de Bedloe, l'inauguration de la statue de la Liberté, de M. Bartholdi.

La cérémonie a malheureusement été contrariée par le mauvais temps. La pluie n'a cessé de tomber toute la journée.

Malgré cela, la manifestation a eu un caractère imposant, et l'affluence était considérable.

La plupart des édifices et monuments, ainsi que les bâteaux du port étaient pavoisés de drapeaux, parmi lesquels ou remarquait surtout des drapeaux tricolores.

Le programme de la fête comprenait une grande procession civile et militaire qui a traversé la ville, et une revue navale sur la partie nord de la rivière, de la 42e avenue jusqu'à l'île de Bedloe.

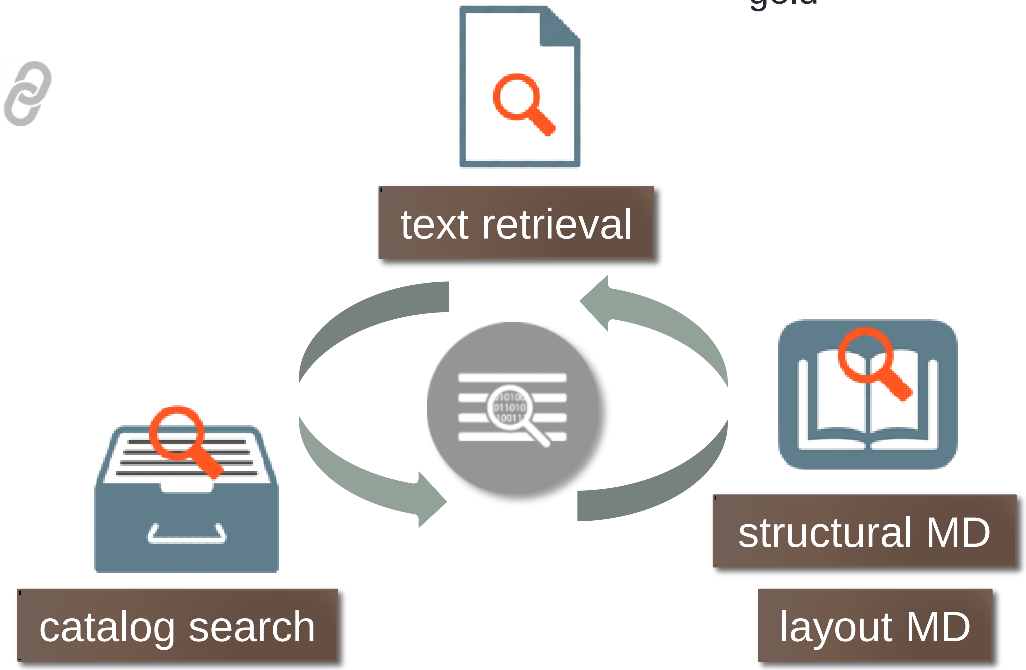
Un défilé.

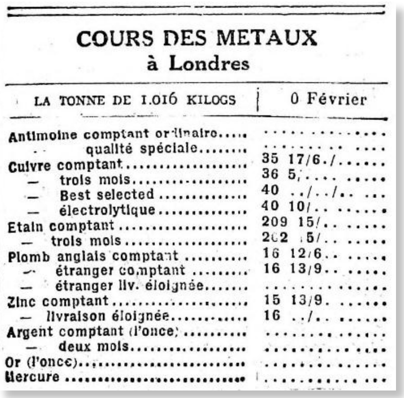
Une foule énorme était massée sur tout le parcours de la procession, qui a défilé, dans la 5e avevue, devant le président Cleveland, M. Bayard, ministre d'Etat; M. Whitney, ministre de la marine; M. Vilas, maître des postes; M. Lamar, ministre de l'intérieur; le général Sheridan, commandant en chef de l'armée américaine; M. Bartholdi, M. de Lesseps, l'amiral Jaurès et un grand nombre d'invités français et de notabilités.

Un arc de triomphe, décoré d'immortelles, avait été élevé en face des bureaux du *New-York Herald*.

La procession était composée de plusieurs socié-

Le Matin, Oct. 29th 1886





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 15/.. .....
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 Mode for Newspapers

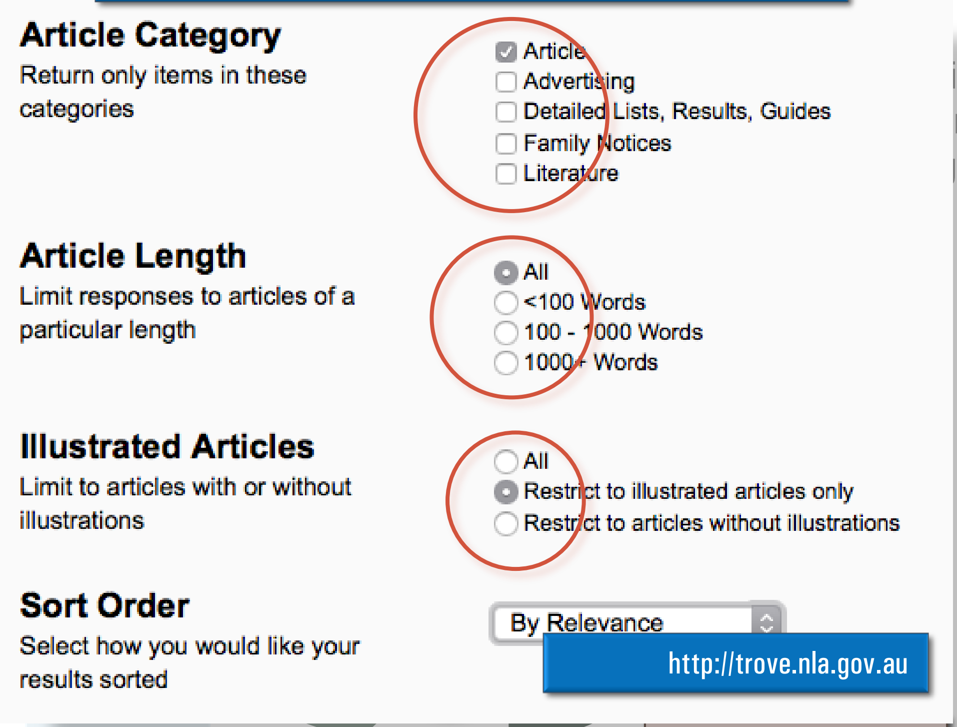
- **Trove** (@Australia) example:

? **articles** from 1886 **title** OR **subtitle** con
“statue & liberté”

in *Le Matin*
metal prices”
old”

tools

➔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

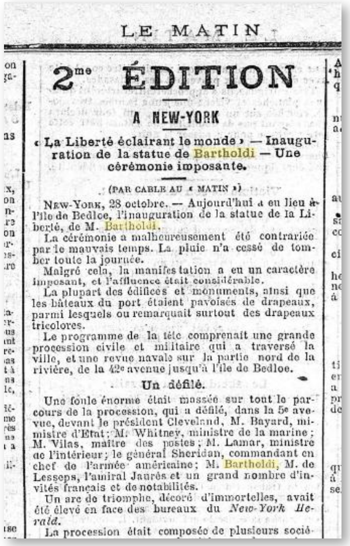
Illustrated Articles
Limit to articles with or without illustrations

○ 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



LE MATIN

2me ÉDITION

A NEW-YORK

« La Liberté éclairant le monde » — Inauguration de la statue de Bartholdi — Une cérémonie imposante.

(PAR CABLE AU « MATIN »)

NEW-YORK, 28 octobre. — Aujourd'hui a eu lieu à l'île de Bedloe, l'inauguration de la statue de la Liberté, de M. Bartholdi.

La cérémonie a malheureusement été contrariée par le mauvais temps. La pluie n'a cessé de tomber toute la journée.

Malgré cela, la manifestation a eu un caractère imposant, et l'affluence était considérable.

La plupart des édifices et monuments, ainsi que les bâteaux du port étaient pavoisés de drapeaux, parmi lesquels on remarquait surtout des drapeaux tricolores.

Le programme de la fête comprenait une grande procession civile et militaire qui a traversé la ville, et une revue navale sur la partie nord de la rivière, de la 42e avenue jusqu'à l'île de Bedloe.

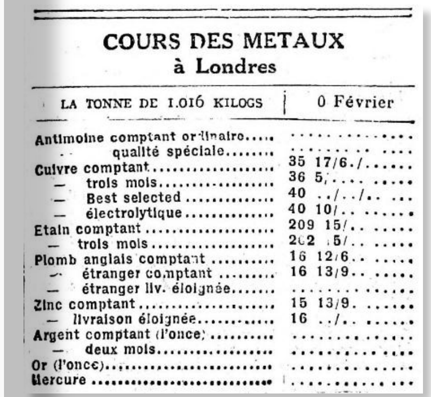
Un défilé.

Une foule énorme était massée sur tout le parcours de la procession, qui a défilé, dans la 5e avenue, devant le président Cleveland, M. Bayard, ministre d'Etat; M. Whitney, ministre de la marine; M. Vilas, maître des postes; M. Lamar, ministre de l'intérieur; le général Sheridan, commandant en chef de l'armée américaine; M. Bartholdi, M. de Lesseps, l'amiral Jaurés et un grand nombre d'invités français et de notabilités.

Un arc de triomphe, décoré d'immortelles, avait été élevé en face des bureaux du *New-York Herald*.

La procession était composée de plusieurs socié-

Le Matin, Oct. 29th 1886



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 15/.. .....
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	....

structural MD

catalog search

layout MD

Advanced Search Mode: Looking for Images

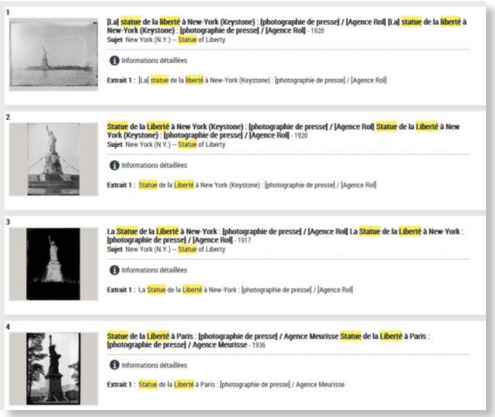
- Newspapers are full of illustrations. But where? **OCR tells us.**
- What are those illustrations about? **OCR gives us the text** around the illustrations, sometimes the illustrations' caption.

tools

?**illustrations** after 1885 where **caption** contains "statue&liberté"

- **Gallica** search on bibliographic metadata only: **15** docs (photos, drawings) in a 400k images collection

- **OCR/OLR helped API search**: **11** hits in the (small) EN test bed (850k pages)



1

[La] statue de la liberté à New-York (Keystone) : [photographie de presse] / [Agence Rol] [La] statue de la liberté à New-York (Keystone) : [photographie de presse] / [Agence Rol] - 1920
Sujet New York (N.Y.) -- Statue of Liberty

Informations détaillées

Extrait 1 : [La] statue de la liberté à New-York (Keystone) : [photographie de presse] / [Agence Rol]

2

Statue de la Liberté à New York (Keystone) : [photographie de presse] / [Agence Rol] Statue de la Liberté à New York (Keystone) : [photographie de presse] / [Agence Rol] - 1920
Sujet New York (N.Y.) -- Statue of Liberty

Informations détaillées

Extrait 1 : Statue de la Liberté à New York (Keystone) : [photographie de presse] / [Agence Rol]

3

La Statue de la Liberté à New-York : [photographie de presse] / [Agence Rol] La Statue de la Liberté à New-York : [photographie de presse] / [Agence Rol] - 1917
Sujet New York (N.Y.) -- Statue of Liberty

Informations détaillées

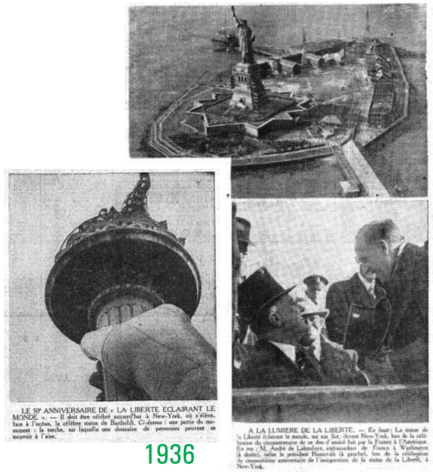
Extrait 1 : La Statue de la Liberté à New-York : [photographie de presse] / [Agence Rol]

4

Statue de la Liberté à Paris : [photographie de presse] / Agence Meurisse Statue de la Liberté à Paris : [photographie de presse] / Agence Meurisse - 1936

Informations détaillées

Extrait 1 : Statue de la Liberté à Paris : [photographie de presse] / Agence Meurisse



LE 50e ANNIVERSAIRE DE « LA LIBERTE ECLAIRANT LE MONDE ». — Il doit être célébré aujourd'hui à New-York, où s'élève, face à l'océan, la célèbre statue de Bartholdi. *Ci-dessus* : une partie du monument : la torche, sur laquelle une douzaine de personnes peuvent se mouvoir à l'aise.

A LA LUMIERE DE LA LIBERTE. — *En haut* : La statue de la Liberté éclairant le monde, sur son îlot, devant New-York, lors de la célébration du cinquantenaire de ce don d'amitié fait par la France à l'Amérique. *En bas* : M. André de Laboulaye, ambassadeur de France à Washington (à droite), salue le président Roosevelt (à gauche), lors de la célébration du cinquantième anniversaire de l'inauguration de la statue de la Liberté, à New-York.

1936

Wide World Photos.

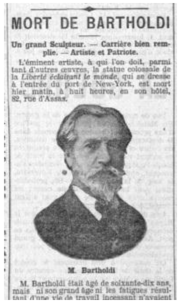
LA STATUE DE LA LIBERTE
QUI DOMINE L'ENTREE DU PORT DE NEW-YORK

1931

LA STATUE DE LA LIBERTÉ DU PONT DE GRENELLE FAIT DÉSORMAIS FACE A SA RÉPLIQUE DE NEW-YORK

Il semblait illogique que la statue de la Liberté qui s'élève à la pointe de l'île des Cygnes tournât le dos aux visiteurs arrivant de l'ouest de Paris. Pour remédier à cet état de choses, on a fait exécuter au monument un quart de tour sur lui-même

1937



MORT DE BARTHOLDI

Un grand Sculpteur. — Carrière bien remplie. — Artiste et Patriote.

L'éminent artiste, à qui l'on doit, parmi tant d'autres œuvres, la statue colossale de la *Liberté éclairant le monde*, qui se dresse à l'entrée du port de New-York, est mort hier matin, à huit heures, en son hôtel, 82, rue d'Assas.

M. Bartholdi

M. Bartholdi était âgé de soixante-dix ans, mais ni son grand âge ni les fatigues résultant d'une vie de travail incessant n'avaient

1904

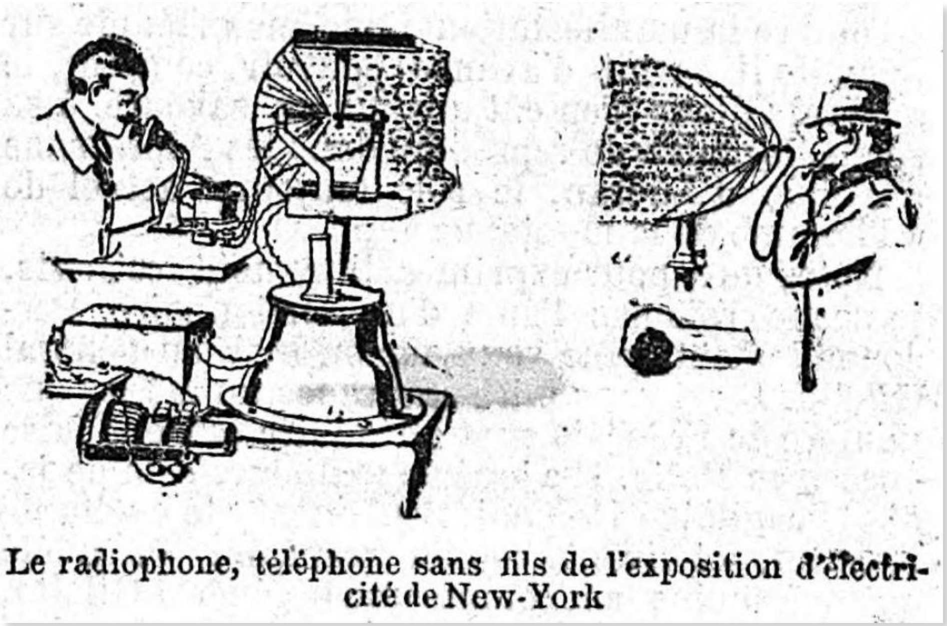
```
http://localhost:8984/rest?run=findCaptionedIllustrations.xq&fromDate=1886-01-01&keyword=statue.*liberte
```

Advanced Search Mode: Looking for Images

- After 1890, newspapers are a rich iconographic source regarding historical events but also everyday life, popular culture... We need to give our users the tools to use it!

tools

Fireman (NY police), *Le Matin*, 1932



Le radiophone, téléphone sans fils de l'exposition d'électricité de New-York

Wireless Phone (Graham Bell), New York Electricity Exposition, *Le Gaulois*, July. 12th 1899

Jacqueline Cochrane (jet pilot), *Le Matin*, 1938

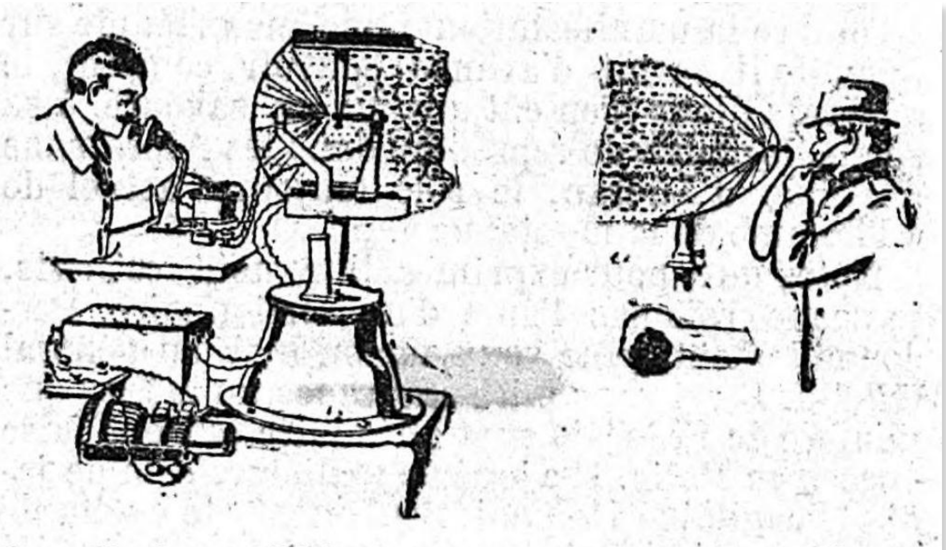
➔ And illustrated articles are more likely to hold important content…

Advanced Search Mode: Looking for Images

- After 1890, newspapers are a rich iconographic source regarding historical events but also everyday life, popular culture... We need to give our users the tools to use it!

tools

Jean-Philippe



First IFLA remote presentation at the BnF (1899)

Caroline

Retronews.fr's Aims

Retronews covers 3 centuries of French Newspapers (1631-1945).

A dynamic process is on to search, find and share a better understanding of these times and bring connections to nowadays.

We aim at:

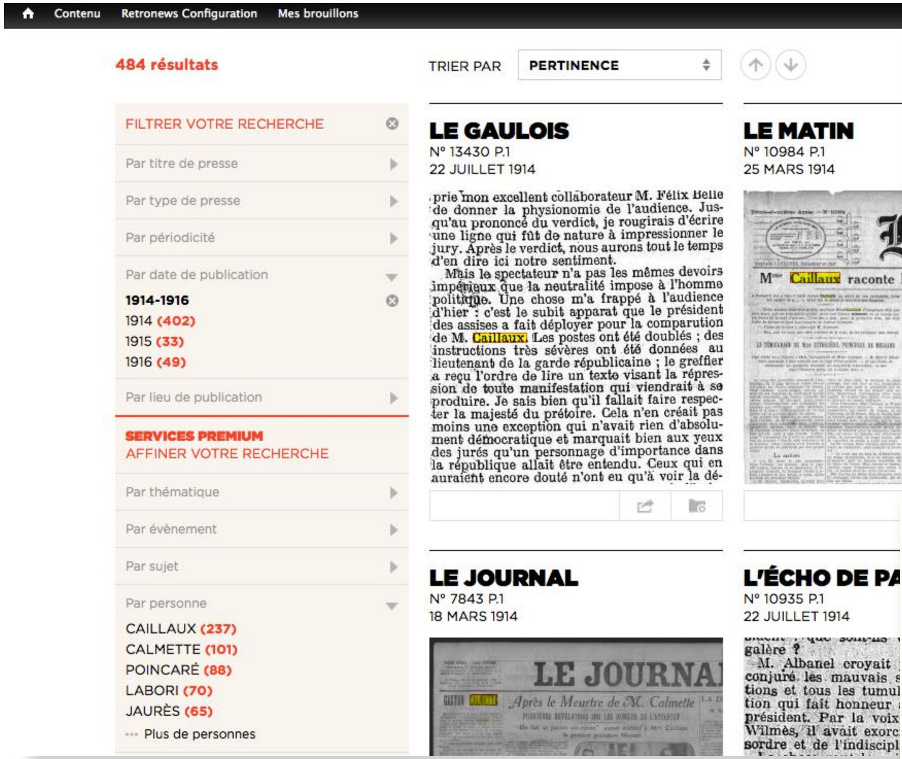
- ✓ Improving the access to the newspapers with the latest online functionalities.
- ✓ Helping to find and better understand many topics.

In term of volume, this means today 3M pages; 15M articles are available on the web site. The double amount will be available in 2018.
(At the BnF, 5 % of the press collections are digitized.)

Specific Key Points

- The documents are enriched and structured with thematics, topics, named entities... This means the largest directory book of people.
- This brings both the possibility to go deeper into a research and, to enlarge the research when it is in a dead-end.
- Users also enjoy specific tools to work with the content and be able to insert it into their academic works (with saved queries, personal selection, extraction of content, etc.).
- The portal is editorialized with chronicles: daily "Echos", thematic folders, the timelines of the newspapers in France.

RetroNews



484 résultats

TRIER PAR PERTINENCE

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

Mme Caillaux raconte le

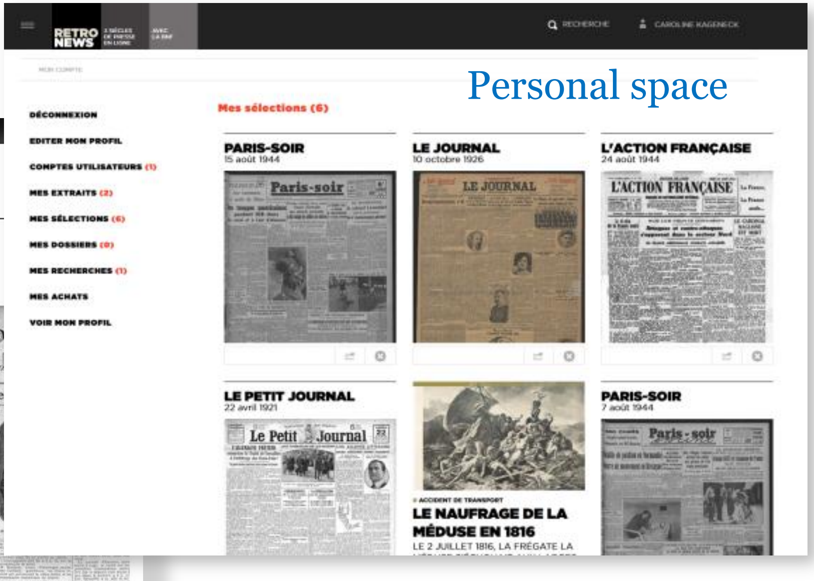
LE JOURNAL
N° 7843 P.1
18 MARS 1914

LE JOURNAL
Après le Meurtre de M. Calmette

L'ÉCHO DE PA
N° 10935 P.1
22 JUILLET 1914

galère ?
M. Albanel croyait conjuré les mauvais tions et tous les tumul tion qui fait honneur président. Par la voix Wilmès, il avait exorc sordre et de l'indiscipl

Facetted search on bibliographical and semantical criteria



Personal space

- DÉCONNEXION
- EDITER MON PROFIL
- COMPTES UTILISATEURS (1)
- MES EXTRAITS (2)
- MES SÉLECTIONS (6)
- MES DOSSIERS (0)
- MES RECHERCHES (1)
- MES ACHATS
- VOIR MON PROFIL

Mes sélections (6)

PARIS-SOIR
15 août 1944

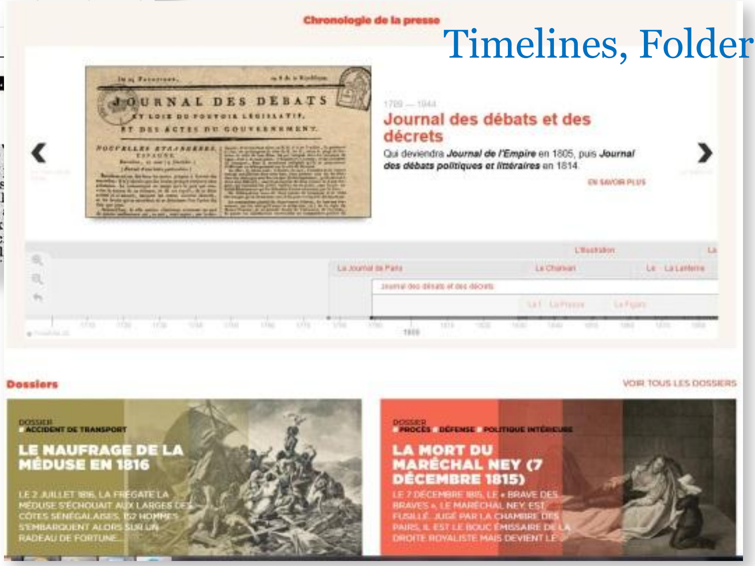
LE JOURNAL
10 octobre 1926

L'ACTION FRANÇAISE
24 août 1944

LE PETIT JOURNAL
22 avril 1921

ACCIDENT DE TRANSPORT
LE NAUFRAGE DE LA MÉDUSE EN 1816
LE 2 JUILLET 1816, LA FRÉGATE LA

PARIS-SOIR
7 août 1944



Chronologie de la presse

Timelines, Folders

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

Dossiers

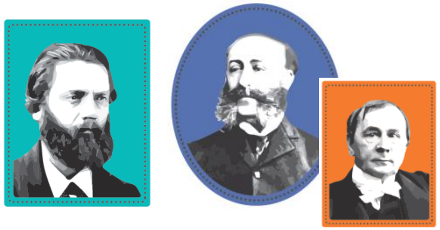
VOIR TOUS LES DOSSIERS

DOSSIER
ACCIDENT DE TRANSPORT
LE NAUFRAGE DE LA MÉDUSE EN 1816
LE 2 JUILLET 1816, LA FRÉGATE LA MÉDUSE S'ÉCHOUAIT AUX LARGES DES CÔTES SÉNÉGALAISES. 152 HOMMES S'EMBARQUENT ALORS SUR UN RADEAU DE FORTUNE...

DOSSIER
PROCÈS DÉFENSE POLITIQUE INTÉRIEURE
LA MORT DU MARÉCHAL NEY (7 DÉCEMBRE 1815)
LE 7 DÉCEMBRE 1815, LE « BRAVE DES BRAVES », LE MARÉCHAL NEY, EST FUSILLÉ. JUGÉ PAR LA CHAMBRE DES PAIRS, IL EST LE BOUC ÉMISSAIRE DE LA DROITE ROYALISTE MAIS DEVIENT LE...

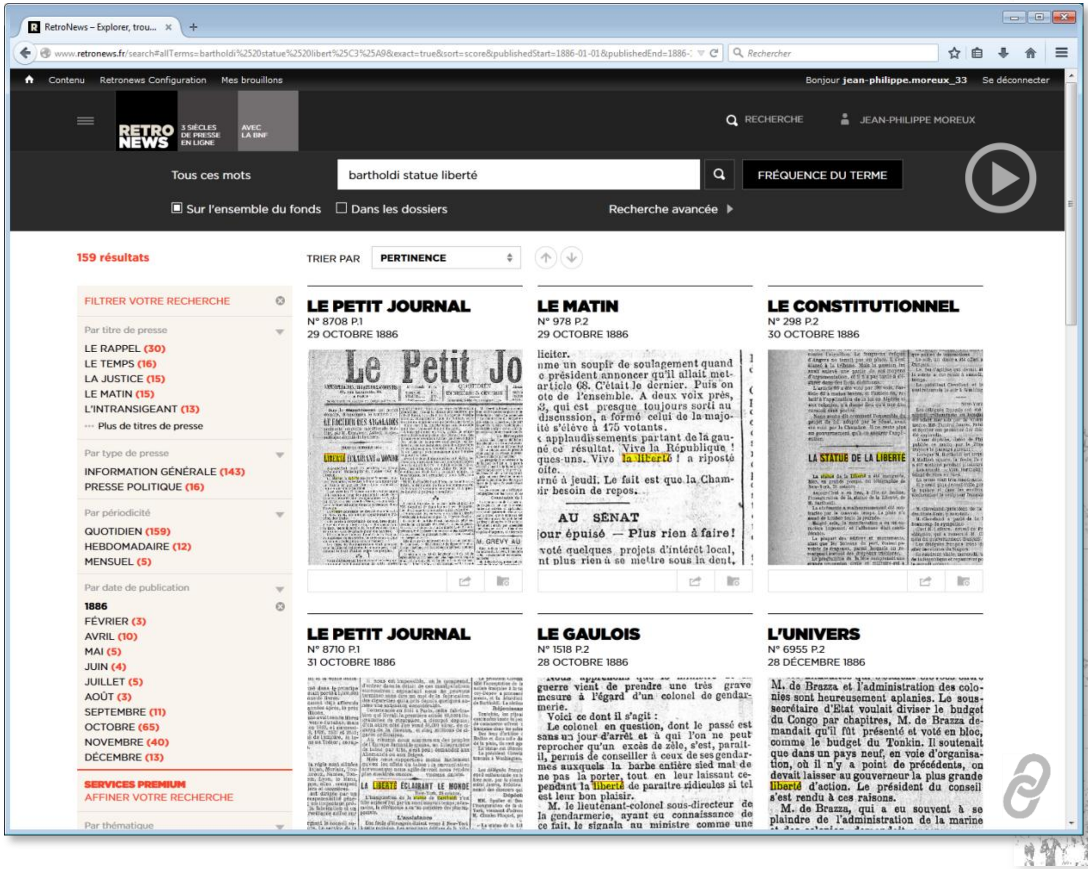
Enriching the Content with Semantic

RetroNews uses four concepts to enrich the text with semantic annotations:



- *Named entities* (3 categories: people, places, organizations): NE recognition is driven by linguistic grammar-based techniques and authorities files.
- *Themes* (14 top level themes, 231 second level): derived from the IPTC classification and refurbed for heritage press. A lexica has been created for each theme, and the text corpus is indexed according to these lexica.
- *Events* (147): closed list of historical events defined by experts of the editorial team, each event associated with a lexica and with dates.
- *Topics' modelling* uses Wikipedia articles' titles and a list of the most frequent queries expressed by users on the BnF's DL.

Advanced Query and Discovery



RetroNews – Explorer, trou...
www.retronews.fr/search#allTerms=bartholdi%2520statue%2520libert%25C3%25A9&exact=true&sort=score&publishedStart=1886-01-01&publishedEnd=1886-
Rechercher
Contenu
Retronews Configuration
Mes brouillons
Bonjour jean-philippe.moreux_33
Se déconnecter
RETRO NEWS
3 SIÈCLES DE PRESSE EN LIGNE
AVEC LA BNF
RECHERCHE
JEAN-PHILIPPE MOREUX
FRÉQUENCE DU TERME
Recherche avancée
TRIER PAR
FILTRER VOTRE RECHERCHE
SERVICES PREMIUM
AFFINER VOTRE RECHERCHE

Tous ces mots: bartholdi statue liberté

☑ Sur l'ensemble du fonds ☐ Dans les dossiers

159 résultats

TRIER PAR: PERTINENCE

Par titre de presse
- LE RAPPEL (30)
- LE TEMPS (16)
- LA JUSTICE (15)
- LE MATIN (15)
- L'INTRANSIGEANT (13)
- ··· Plus de titres de presse

Par type de presse
- INFORMATION GÉNÉRALE (143)
- PRESSE POLITIQUE (16)

Par périodicité
- QUOTIDIEN (159)
- HEBDOMADAIRE (12)
- MENSUEL (5)

Par date de publication
- 1886
- FÉVRIER (3)
- AVRIL (10)
- MAI (5)
- JUIN (4)
- JUILLET (5)
- AOÛT (3)
- SEPTEMBRE (11)
- OCTOBRE (65)
- NOVEMBRE (40)
- DÉCEMBRE (13)

Par thématique

LE PETIT JOURNAL
N° 8708 P.1
29 OCTOBRE 1886

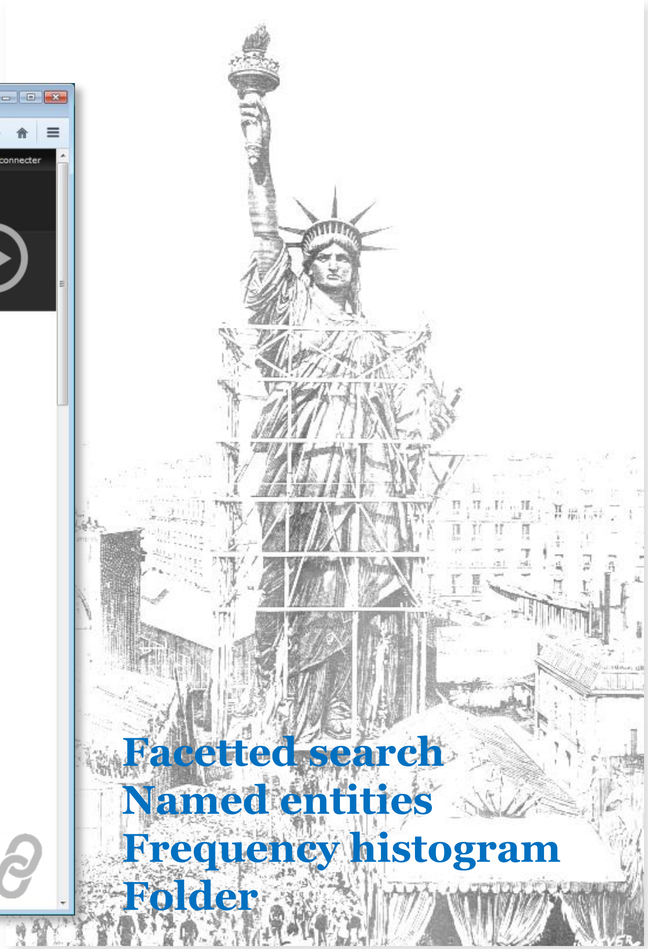
LE MATIN
N° 978 P.2
29 OCTOBRE 1886

LE CONSTITUTIONNEL
N° 298 P.2
30 OCTOBRE 1886

LE PETIT JOURNAL
N° 8710 P.1
31 OCTOBRE 1886

LE GAULOIS
N° 1518 P.2
28 OCTOBRE 1886

L'UNIVERS
N° 6955 P.2
28 DÉCEMBRE 1886



Facetted search
Named entities
Frequency histogram
Folder

The quality of the data

- For datavisualization, as well as for semantic enrichment, some qualities of the data are required to bring valuable results: Are the data representative enough and numerous enough?

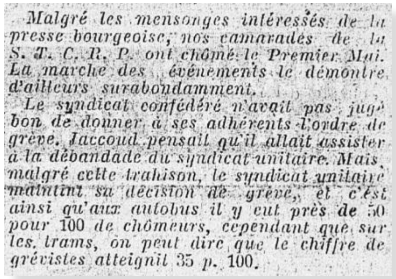
  EX. Is the time lengh of the newspapers long enough to bring meaningfull results?

- OCR quality of historical documents can be an obstacle to semantic enrichment processing.

  EX. Does it make sense to apply NER on bad to medium OCR quality?

- For some semantic enrichment tasks, article segmentation level is needed (OLR).

  EX. Topic modelling or events recognition at page level is almost useless.



Malgré les mensonges intéressés de la presse bourgeoise, nos camarades de la S. T. C. R. P. ont chômé le Premier Mai. La marche des événements le démontre d'ailleurs surabondamment.

Le syndicat confédéré n'avait pas jugé bon de donner à ses adhérents l'ordre de grève. Jaccoud pensait qu'il allait assister à la débandade du syndicat unitaire. Mais malgré cette trahison, le syndicat unitaire maintint sa décision de grève, et c'est ainsi qu'aux autobus il y eut près de 50 pour 100 de chômeurs, cependant que sur les trams, on peut dire que le chiffre de grévistes atteignit 35 p. 100.

The quality of the enrichment

- Semantic enrichment may need reference corpus, which are difficult to build for heritage content.

  EX. Using Wikipedia content for reference will introduce anachronism... Contemporary NE authorities won't adapt to historical newspapers. Etc.

- Do the semantic facettes embody the different aspects of the reality I focus on?

  EX. *The building of thematics such as* Women, Education *also need to go through the channel of* Religion *in the XIX*th *c. in France.*

- And finally, evaluation of semantic enrichment results is a challenging task.

  EX. How to evaluate topic modelling results? One needs a complex approach to build ground truth... and automation to compute rates.

Conclusion

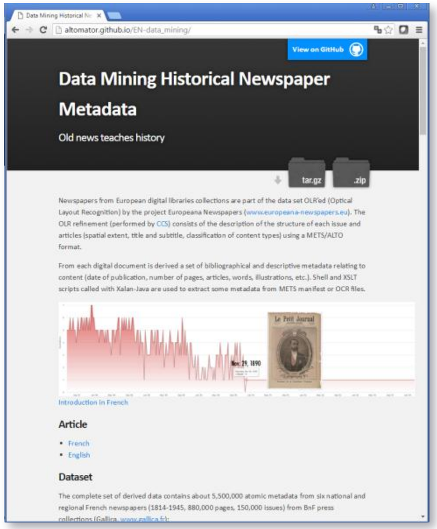
- Quantitative metadata are relevant for all DLs' users: scientists, general public, institutions' employees.
- OLR enrichment provides a rich source of information for researchers, through web portals or ready-to-use datasets. Such datasets, possibly crossed with the OCRed text, usually provide a fertile ground for research hypotheses. Only basic data mining & dataviz methods and tools are needed to exploit such datasets.
- Semantic enrichments helps DLs users to cope with amounts of information ever larger, from innovative perspectives.
- Quantitative metadata combined with OLR and semantic enrichments enhance the global information retrieval capacities of DLs' users.

RETRO NEWS 3 SIÈCLES DE PRESSE EN LIGNE AVEC LES COLLECTIONS DE LA BIBLIOTHÈQUE NATIONALE DE FRANCE

{BnF Bibliothèque nationale de France

Thank you for your attention!

- Datasets (CSV, XML, JSON) and charts are publicly available. Just play with it!



Data Mining Historical Ne ×

altomator.github.io/EN-data_mining/

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

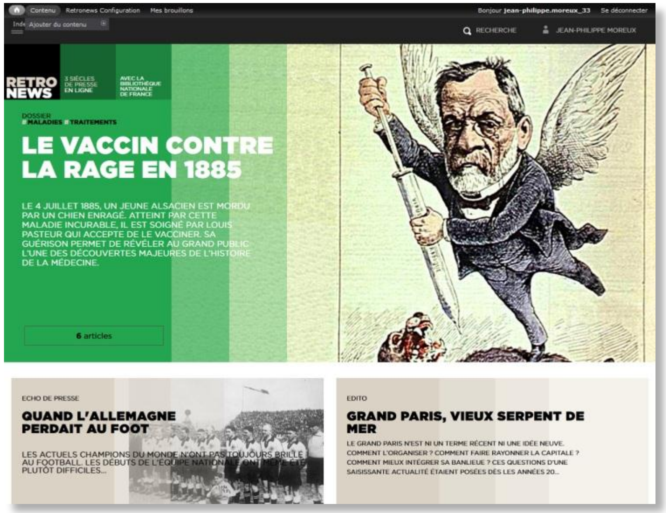
- French
- English

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

- Retronews is freemium. Try it!



Bonjour jean-philippe.moreux_33 Se déconnecter

JEAN-PHILIPPE MOREUX

DOSSIER
MALADIES # TRAITEMENTS

LE VACCIN CONTRE LA RAGE EN 1885

LE 4 JUILLET 1885, UN JEUNE ALSACIEN EST MORDU PAR UN CHIEN ENRAGÉ. ATTEINT PAR CETTE MALADIE INCURABLE, IL EST SOIGNÉ PAR LOUIS PASTEUR QUI ACCEPTE DE LE VACCINER. SA GUÉRISON PERMET DE RÉVÉLER AU GRAND PUBLIC L'UNE DES DÉCOUVERTES MAJEURES DE L'HISTOIRE DE LA MÉDECINE.

6 articles

ECHO DE PRESSE

QUAND L'ALLEMAGNE PERDAIT AU FOOT

LES ACTUELS CHAMPIONS DU MONDE N'ONT PAS TOUJOURS BRILLÉ AU FOOTBALL. LES DÉBUTS DE L'ÉQUIPE NATIONALE ONT MÊME ÉTÉ PLUTÔT DIFFICILES...

EDITO

GRAND PARIS, VIEUX SERPENT DE MER

LE GRAND PARIS N'EST NI UN TERME RÉCENT NI UNE IDÉE NEUVE. COMMENT L'ORGANISER ? COMMENT FAIRE RAYONNER LA CAPITALE ? COMMENT MIEUX INTÉGRER SA BANLIEUE ? CES QUESTIONS D'UNE SAISISSANTE ACTUALITÉ ÉTAIENT POSÉES DÈS LES ANNÉES 20...

http://www.retronews.fr/

europeana newspapers

Thanks to all the Europeana Newspapers partners!

Staatsbibliothek zu Berlin
Preußischer Kulturbesitz

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