

Mining, Visualising and Analysing Historical Newspaper Data: the French National Library Experience

Jean-Philippe Moreux
Bibliothèque nationale de France,
Conversation dpt/Digitisation service

{BnF} Bibliothèque
nationale de France

Digital Approach towards serial publications,
Bruxelles, Tuesday 12 September 2017



Outline

Introduction

Making collections accessible for research

- Pre-processed datasets
- On-demand datasets
- APIs

Digital Scholarship Lab

Historical Newspapers

- ✓ First mass media
- ✓ Essential for the study of the XIX-XXth c.

Challenges

- ✓ Volume (BnF: 100 M pages?)
- ✓ Conservation issues

Research topics

- ✓ Digital history
- ✓ Information sciences
- ✓ Social studies
- ✓ Visual studies...

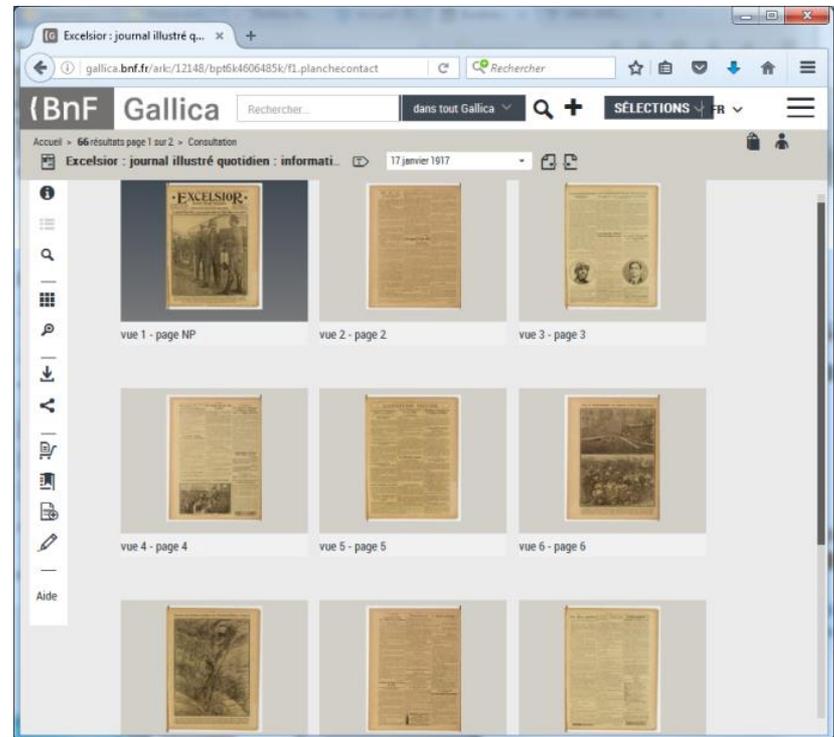


Historical Digitized Newspapers

Challenges

- ✓ **Complex** layout
- ✓ **Composite** contents
- ✓ **Noisy** OCR
- ✓ **Specific GUI** for (user friendly) browsing and searching
- ✓ Volume: 90%-99% **still to be digitized** in Europe?
- ✓ Digitisation **costs**

**... very popular
(70% of Gallica users)
and more & more for DH**



gallica.bnf.fr

Digital Scholarship and Newspapers

Challenges (they will not be addressed in the rest of the presentation!)

- ✓ Most of archives and DLs have not been designed for mining; they have different **access** modes
- ✓ **No centralised storage** (even in centralised countries like France)
- ✓ Complex digital objects
- ✓ Politics of digitisation are not neutral (from selection to digitisation techniques)
- ✓ Relative abundance but most sources are not digitized yet: incompleteness, representativity, “digital laziness”
- ✓ E-legal deposit of born-digital media: gaps, technical barriers (News apps)
- ✓ Copyrighted born-digital media: legal barriers, formats mess (XML, PDF, HTML...)



➔ **Complicate or make impossible text and data mining**

Historical Digitized Newspapers, DHs and DLs

As a digital library, what could we do for researchers?

1. Making collections **accessible** for research
2. Building the future: **digital scholarship lab**

Researchers should focus on research tasks, not on getting access to digital collections!



Making Collections accessible for Research

- ✓ Web access: requesting on catalog and OCRred text, browsing and close reading
- ✓ **Pre-processed datasets:** leveraging on our assets for fulfilling generic needs
- ✓ **On-demand datasets:** let users ask for what they really need
- ✓ **APIs:** let the machines work for us!



Pre-processed Datasets

Newspapers are **composite**. Spotting a theme, section... is a technical challenge (> state of the art)

E.g., how can we help a historian working on the **Stock Market column creation and development** in French newspapers? (1800-1870)

➔ **Article separation,
Layout recognition**



Here, and only here

Article Separation

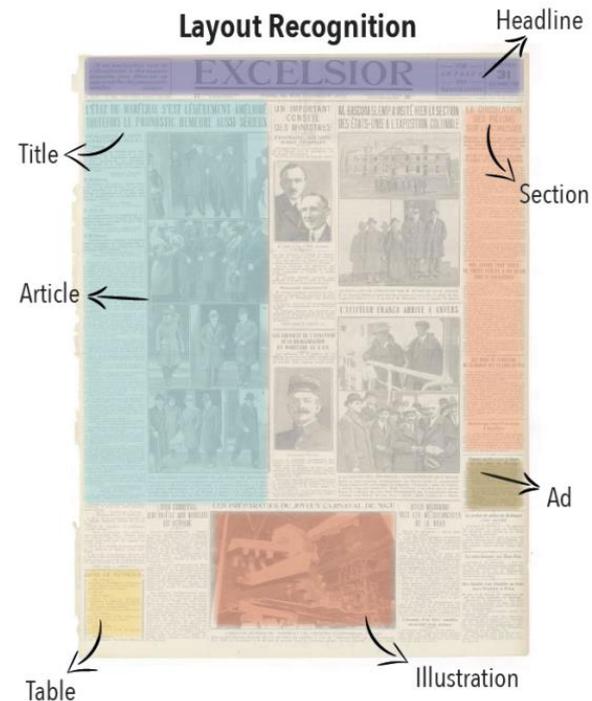
Europeana Newspaper project (2012-2015) has enriched 2M of heritage newspapers pages with Optical Layout Recognition (OLR)

BnF is running its newspaper digitisation program with OLR

- OLR is user friendly
- With OLR, you can build datasets for researchers
- ... but OLR is expensive



11.5M OCR'ed pages,
2M OLR'ed pages
from 14 European
libraries

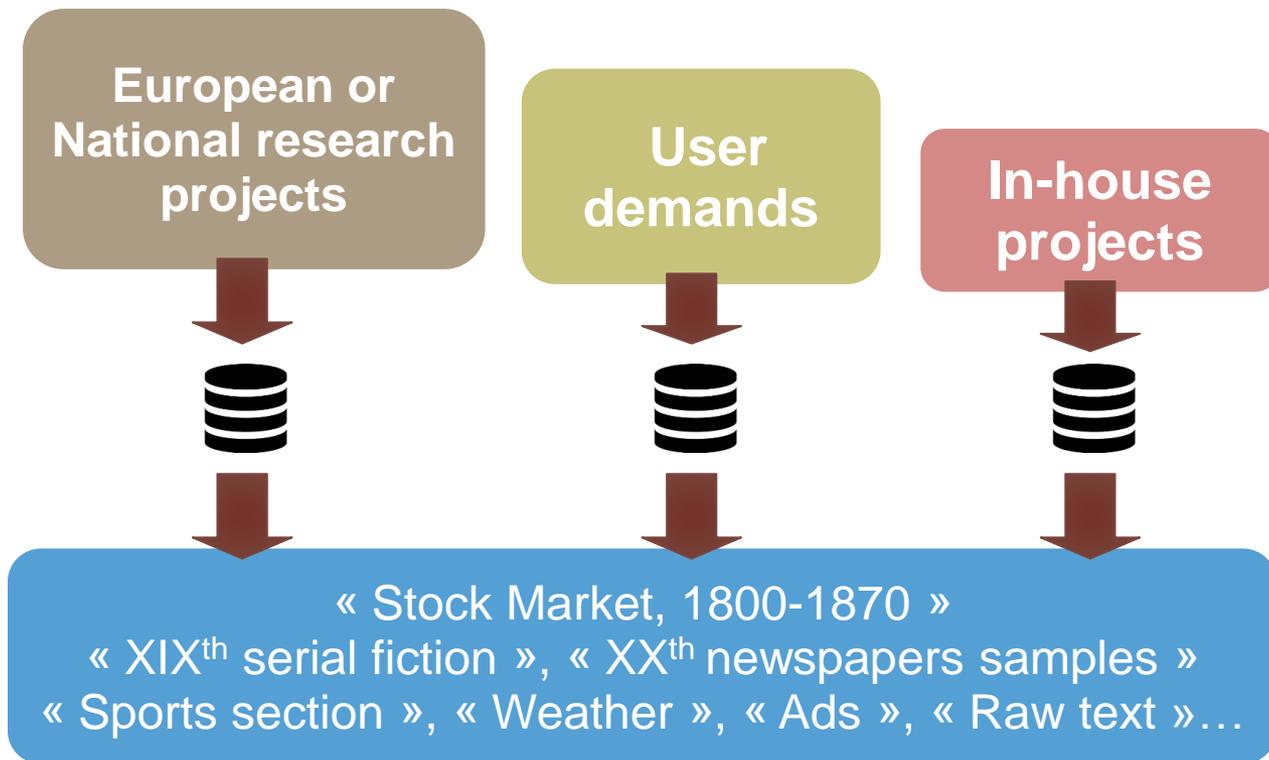


What is OLR?

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

Pre-processed Datasets

Leveraging on research projects, in-house projects... to satisfy generic needs



Collection of pre-processed datasets (text, metadata, image)

It is likely that some researchers will have the same needs...



Working with a pre-processed Dataset

As a researcher:

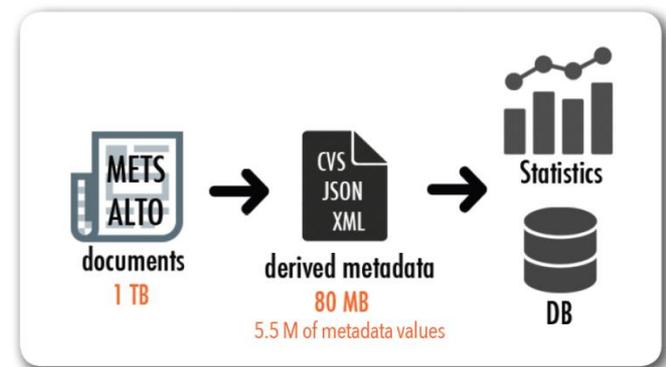
- You don't need to write some **code** to extract your dataset from the digital collection
- You don't need to **parse millions** of XML files
- Your dataset is **research friendly** (XML, JSON, not PDF...)
- The dataset **format** fulfil your needs (e.g. from raw text to heavy METS/ALTO)
- Your dataset is fully described with **metadata** (coverage, completeness, quality metrics...) and **context** (politics of digitisation, formats)
- You don't need to **wait** for DLs to process your request



Example of Production of a Quantitative pre-processed Dataset

OCR and OLR files are full of informational objects tagged into the XML that can be counted: number of words, articles, illustrations, tables, content types classification...

- 880k pages from Europeana Newspapers OLRed corpus
- 7 metadata extracted at issue level, 5 at page level
- 5.5M of metadata values



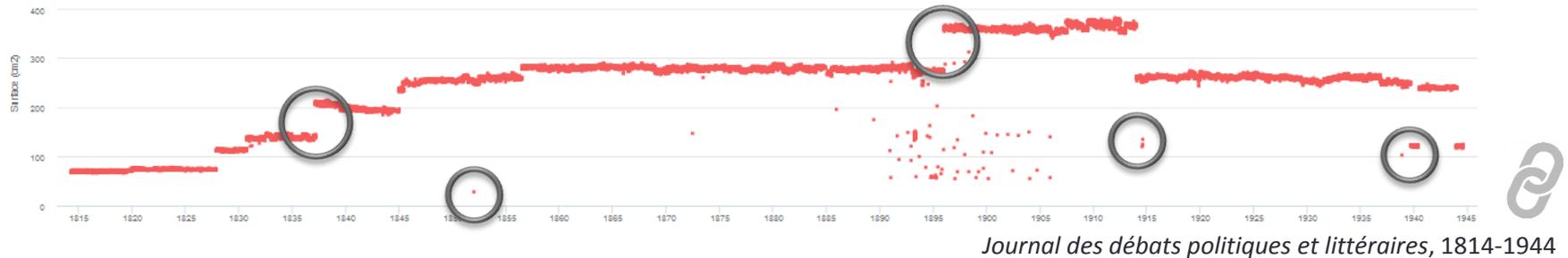
BnF Proof of Concept: 880k pages,
6 titles, 1814-1944

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

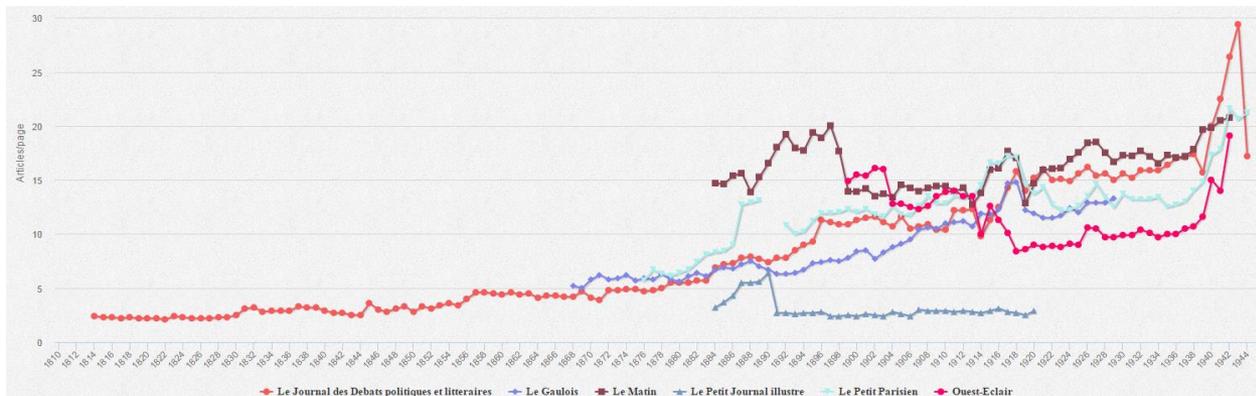
Quantitative Metadata Analysis

Now we can perform quantitative analysis and dataviz.

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

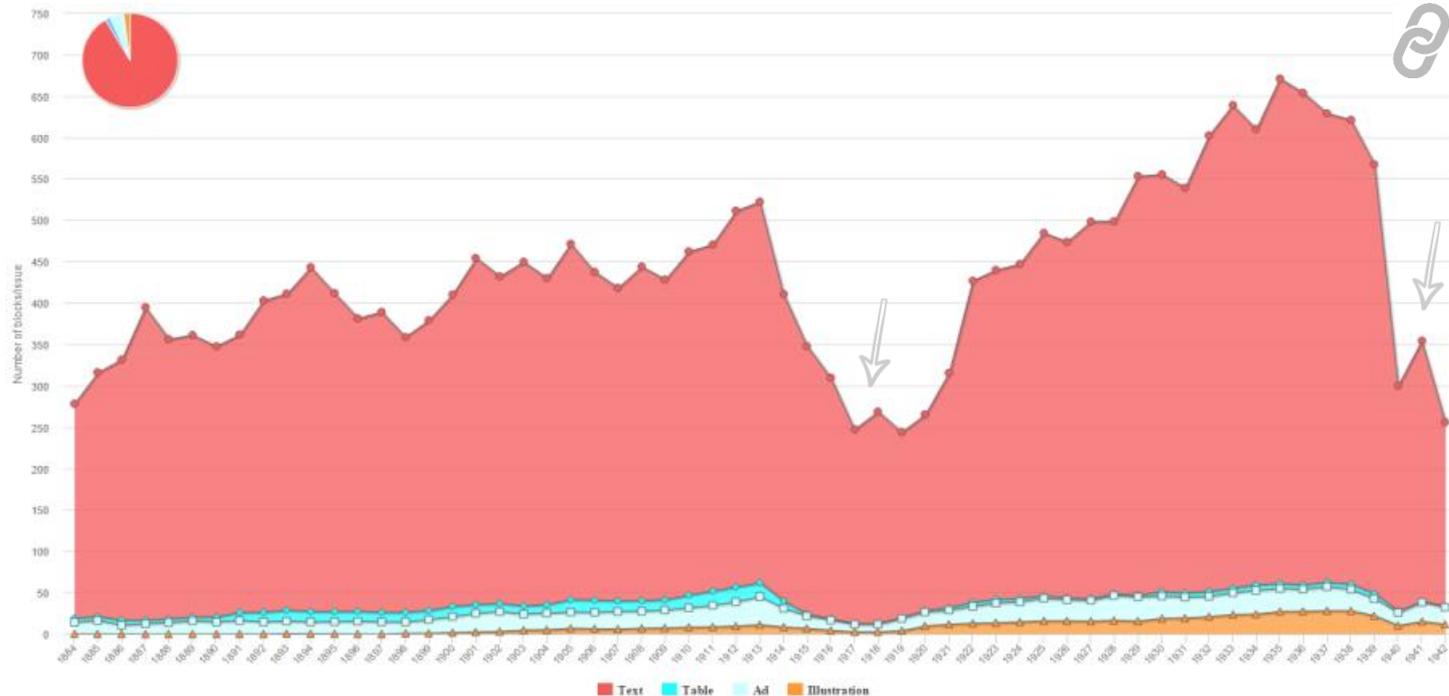


- **History of press/layout:** Visualization of the articles density per page reveals the shift from XVIIth “gazettes” to modern dailies.



Quantitative Metadata Analysis

- **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 ten years).

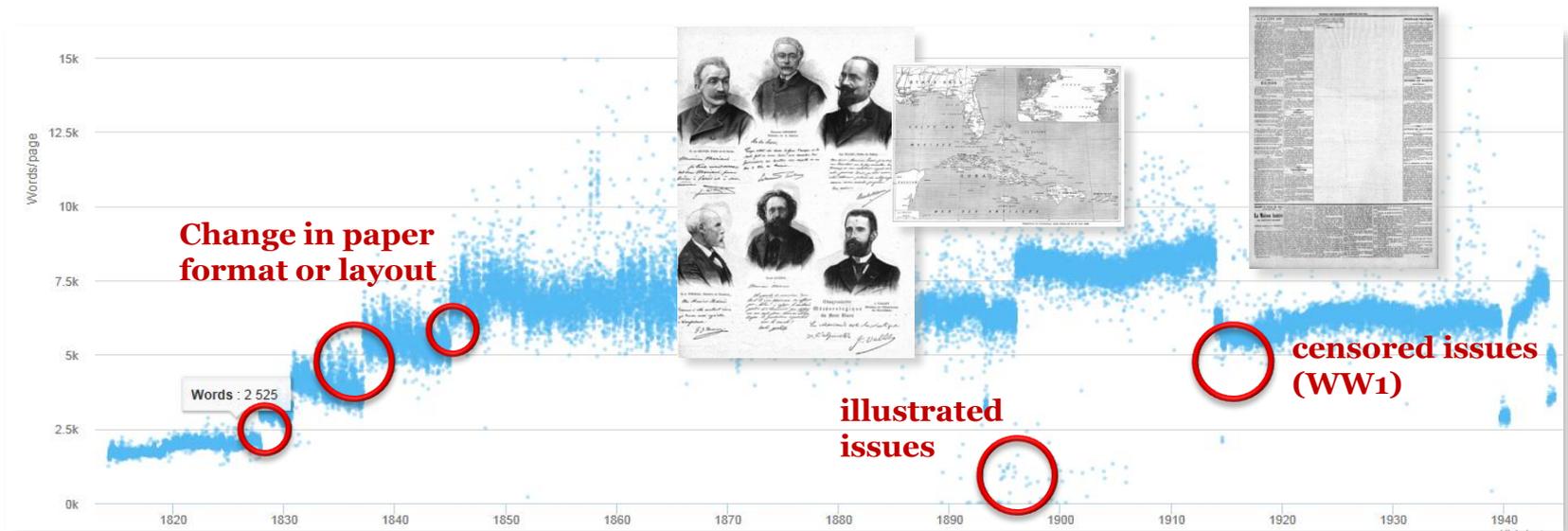


Le Matin, 1884-1942, types of content (articles, text blocks, tables, illustrations, ads)

Quantitative Metadata Analysis

Graph of **words density** reveals breaks due to changes in layout & paper format, outlier issues

Close reading (links to gallica.bnf.fr) / **distant reading**

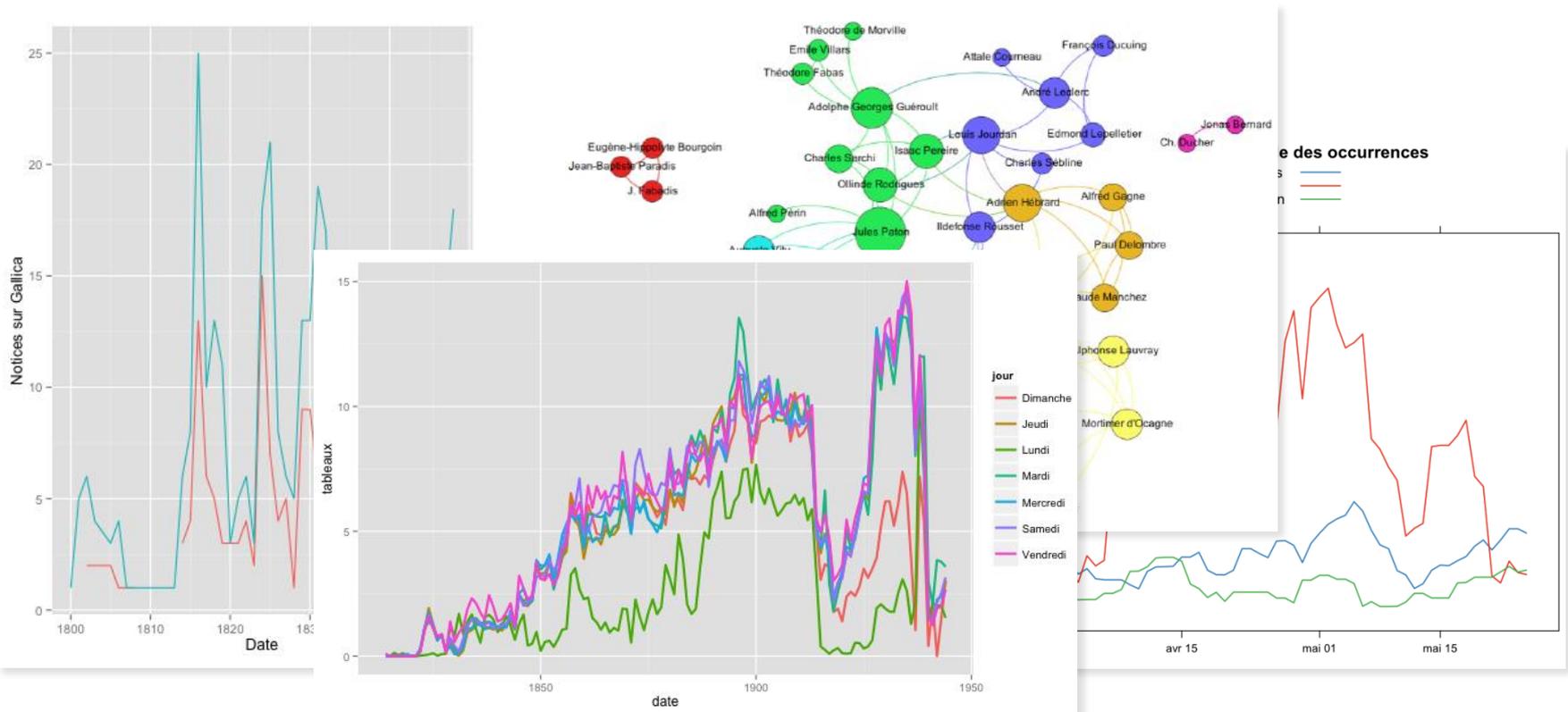


Journal des débats politiques et littéraires, 1814-1944, 45,334 issues displayed



Example of Hybrid* Digital Research

Sources: catalog metadata, pre-processed datasets (Europeana Newspapers + Quantitative EN-BnF), other data

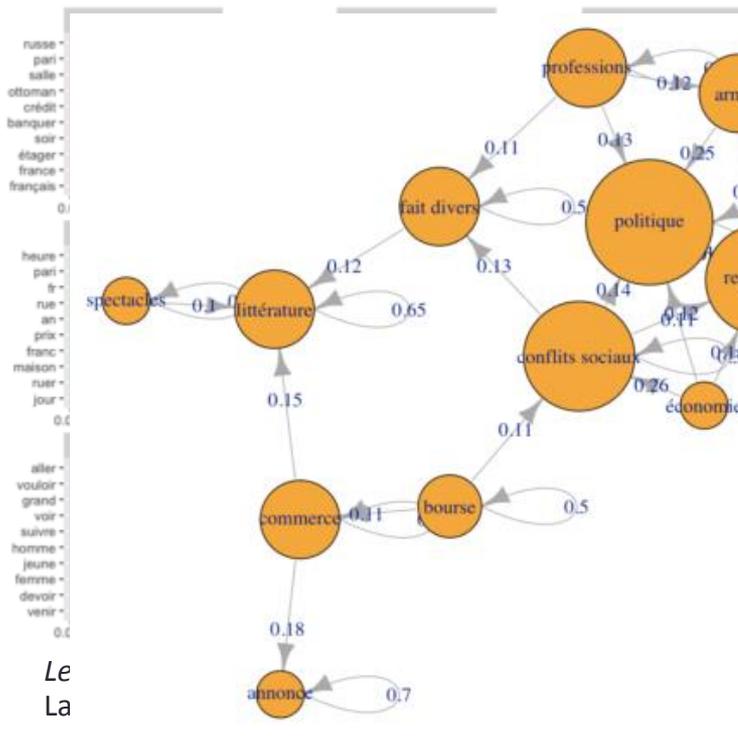


CELSA/GRIPIC, univ. Paris-La Sorbonne
 “Stock Market quotes creation and development in French newspapers”
 (1800-1870), P-C Langlais, PhD in Information Sciences, 2015

*Zaagsma, G., (2013). "On Digital History". *BMGN - Low Countries Historical Review*. 128(4), pp.3–29

What if you don't have Article Separation?

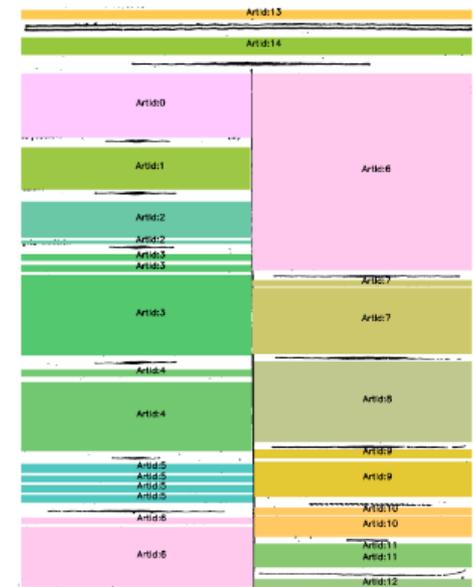
You can apply topic modeling, layout analysis, mixed techniques...



Markov chains for modeling the passage from one section (topic) to another



Layout analysis (pixel based)



T. Palfray, D.Hébert, P. Tranouez, S Nicolas, Thierry Paquet. "Segmentation logique d'images de journaux anciens". Conference Internationale Francophone sur l'Ecrit et le Document, Mars 2012, Bordeaux, pp.317, 2012

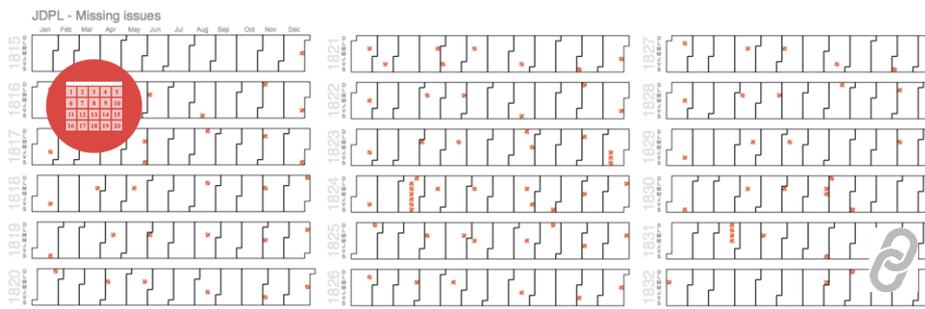
P-C Langlais, <https://numapresse.hypotheses.org>

Quality Assessment

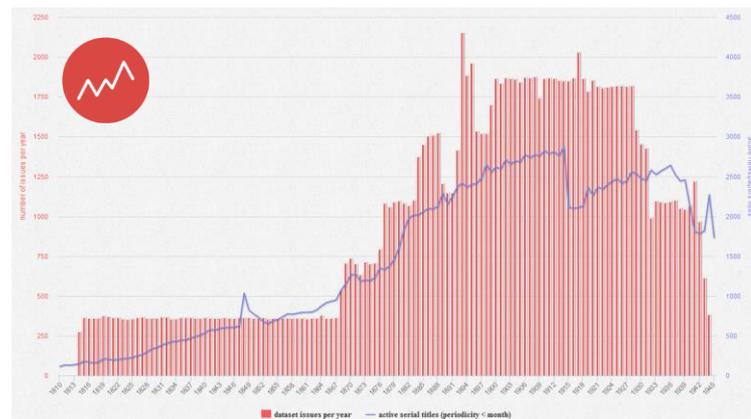
The quality of datasets can affect the validity of the analysis and interpretation. Irregular data in nature or discontinuous in time may introduce bias. → A qualitative assessment should be conducted.

Data vizualisation can contribute to quality control and end-users awareness

A calendar display of a newspapers title data shows rare missing digital issues, which suggests that the digital collection (for this title) is rather complete.



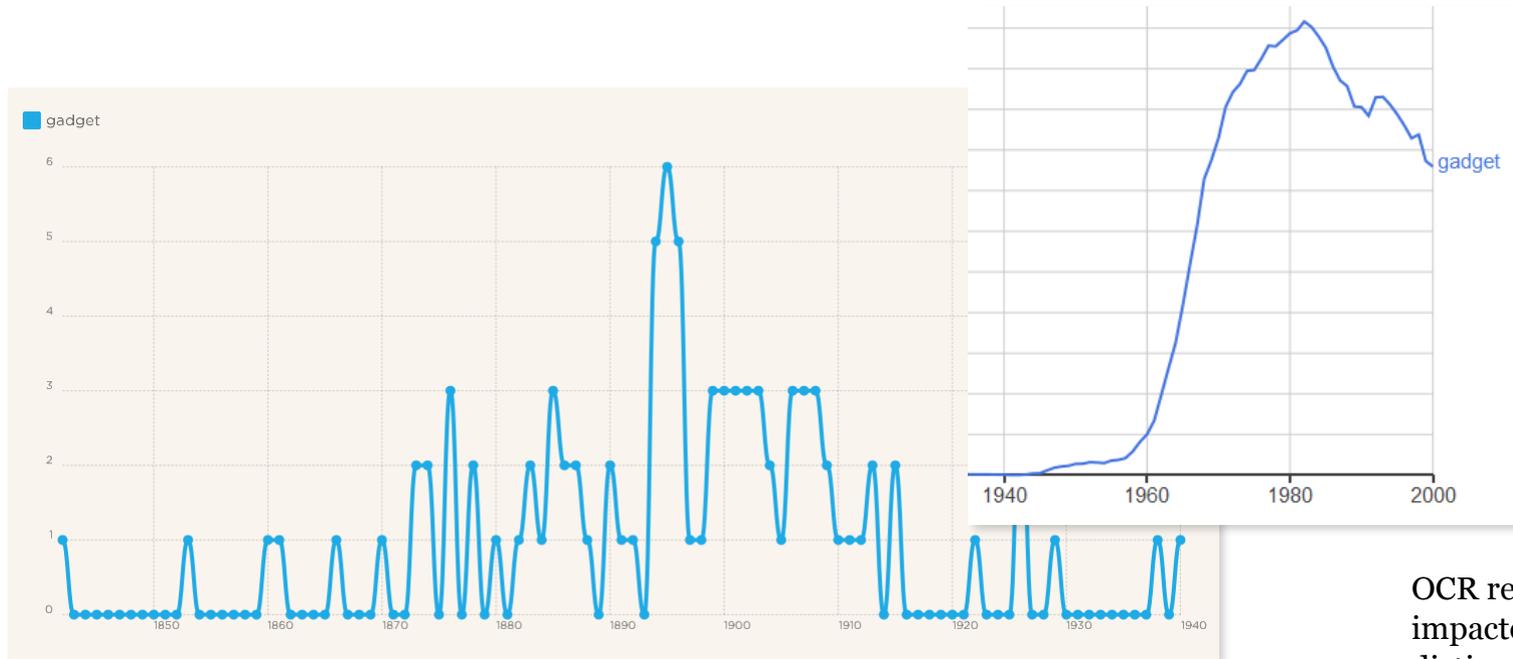
A statistical analysis can also help provide the necessary information on time coverage and data distribution. The period 1814-1867 is under-represented (by a single title). But the editorial production (number of active titles, blue curve) shows that the distribution of the dataset is correct.



Quality Assessment

The **quality of the OCR** is also of great importance for all **NLP** techniques

Pre-processed datasets should be delivered with **QA information**

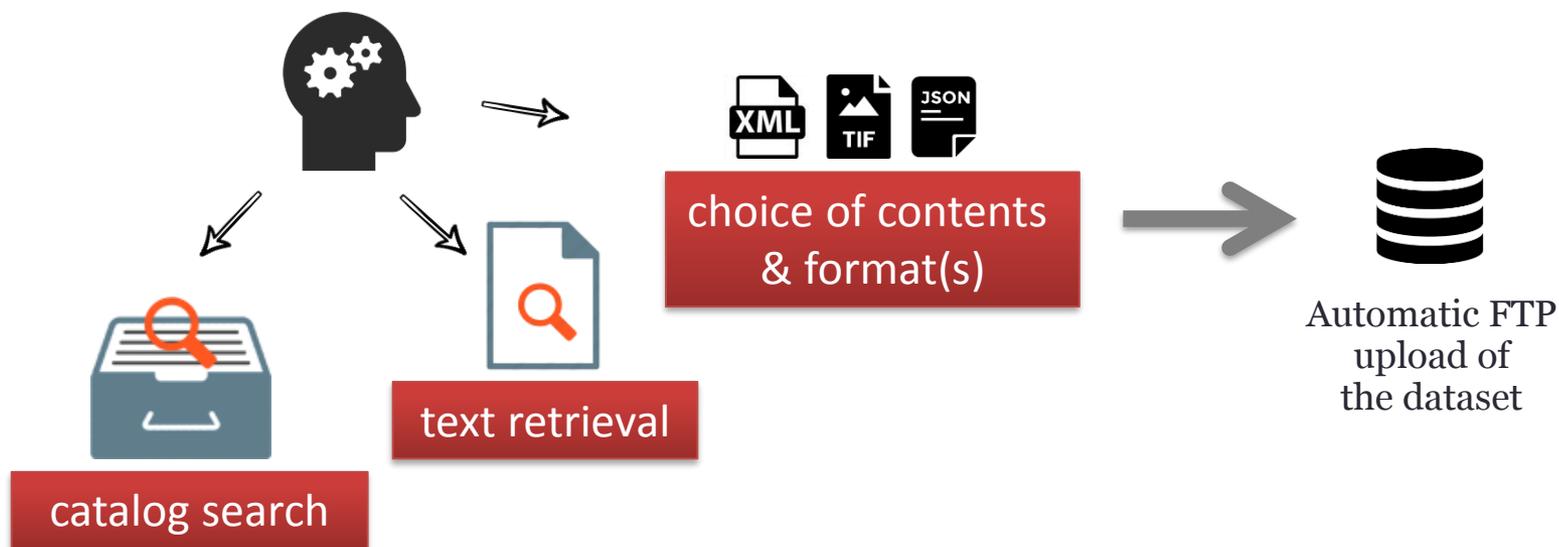


The “**gadget**” in this corpus (1840-1940, retronews.fr) are actually false positives of “**bugdet**”

OCR results are impacted by the dictionaries used, the document genre, the publication date...

On-demand Datasets

- Some users have **specific** needs, particularly on the selection step.
- **Automation** of the application and delivery process could make sense (for both the researchers and the DLs...)
- The on-going BnF “CORPUS” project is investigating these needs



APIs (Application Protocol Interface)

Machine-operable access to content for on-line dissemination:

- ✓ DLs must **disseminate** all their digital collections at various levels of **granularity** (document, page, article, paragraph, text fragment...)
- ✓ Thanks to **interoperable standards** (OAI-PMH, IIIF...)
- ✓ Making these digital object **machine-operable** (particulary for machine-harvesting: Europeana, CLARIN...)
- ✓ They must offer reference to these objects, and make these IDs **persistent**

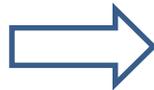


→ All these requirements are researchers friendly (autonomy, instant access, no administrative burden)

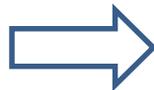
IIIF example

- ✓ **Canvas** are abstract containers for modeling a page content
- ✓ **Annotations** can reference identifiable fragments of text

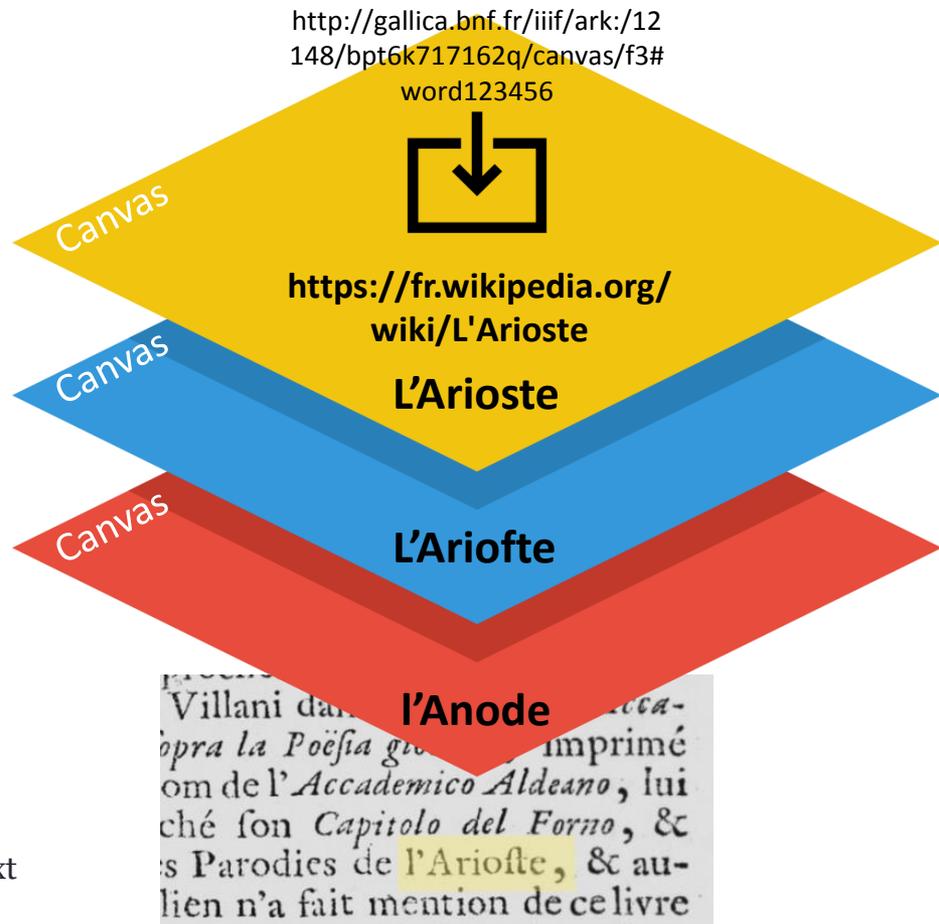
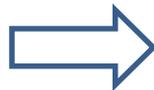
User annotations,
automatic annotations



Post-correction of
OCR (*crowdsourcing*)

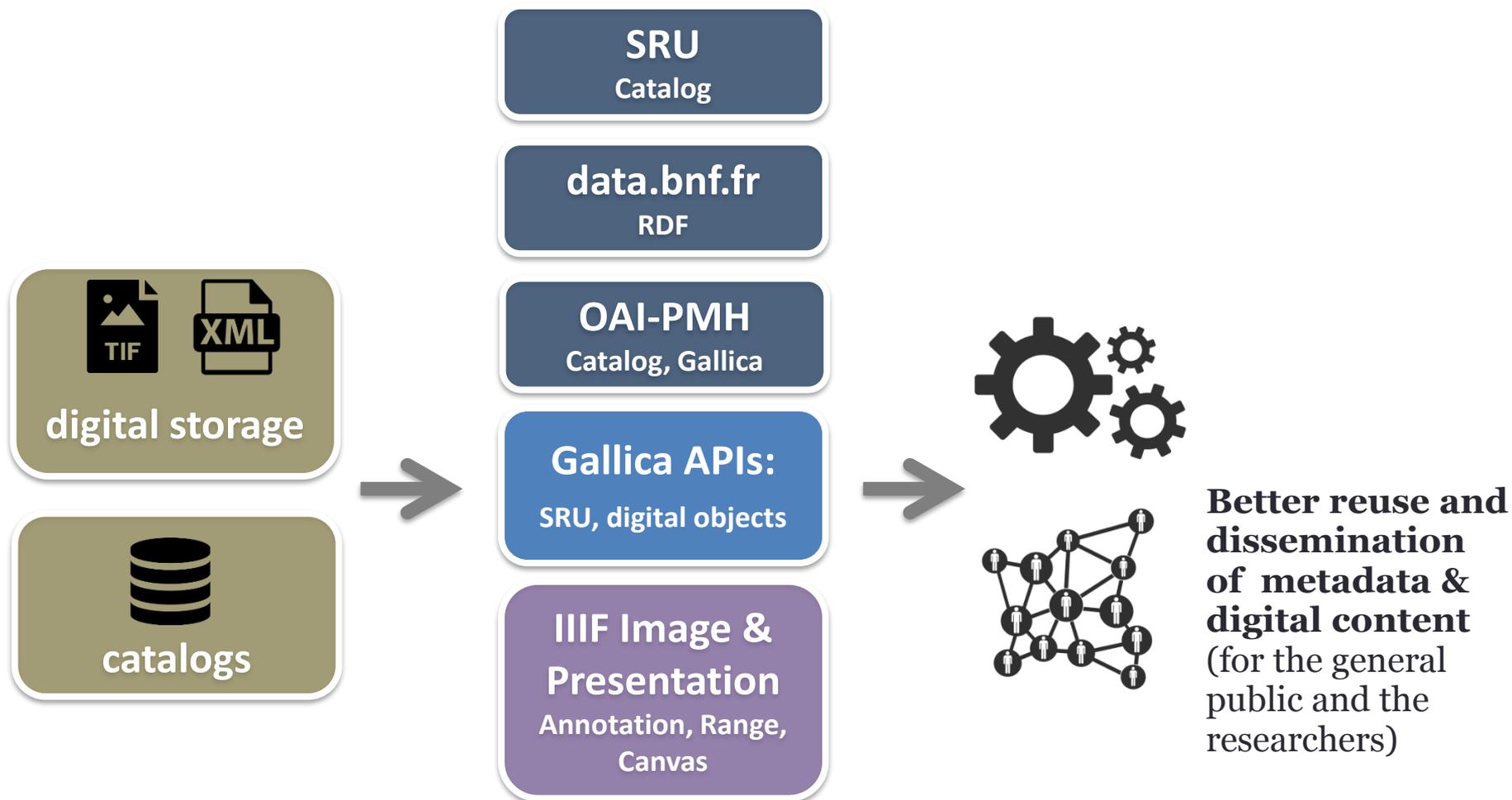


Raw OCR



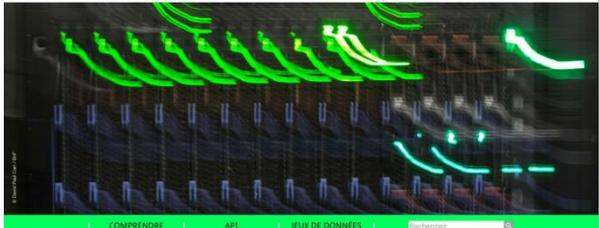
On-going work to link IIIF-Presentation API and text content (IIIF Newspapers Group, ALTO Board)

APIs @ BnF



APIs @ BnF

(BnF) API et jeux de données Découvrez et utilisez les données de la BnF



Le portail API et jeux de données décrit les données diffusées par la BnF, sous forme d'API requêttables aussi bien que de corpus pré-construits. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Actualités

La BnF organise son deuxième hackathon les 25-26 novembre 2017

Nouveaux jeux de données

Découvrez et utilisez les API et jeux de données de la BnF

Laissez-vous guider dans les API et jeux de données

Découvrir les réutilisations remarquables

Les identifiants, un pont entre les API

Conditions d'utilisation

Conditions d'utilisation | Ecrire à la BnF



2nd Edition: November 2017

- ✓ Pre-processed datasets
- ✓ APIs
- ✓ Users guide

➔ Raise awareness among researchers about the availability of the DLs resources

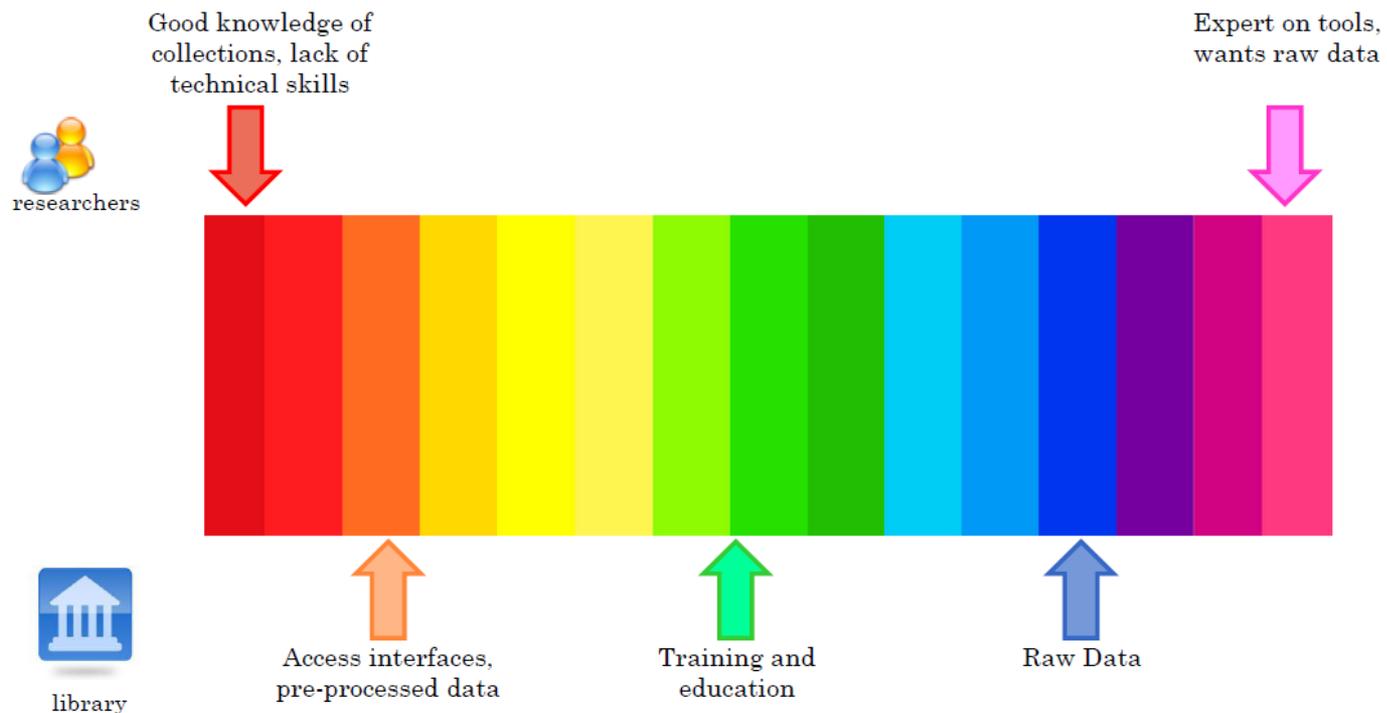
api.bnf.fr (November 2017)

Conclusion on Access to Digital Content

- ✓ Pre-processed datasets, on-demand datasets and APIs will **satisfy the majority of the user demands**
- ✓ It will **speed up the delivery of content** to researchers and **reduce** the administrative burden on both parts.
- ✓ It will contribute to **dissemination** and **reuse** of digital resources
- ✓ For other use cases, we still have **partnership** (for large academic projects) and **contracting** (when the dataset production is not costless on DLs side) solutions.

But Digital Scholarship is not only a matter of Access to Datasets...

Digital scholars are a **special** users group for DLs. They embody a **wide array of different situations**. The services a DL can deliver are **not only a matter of datasets and technical formats**.



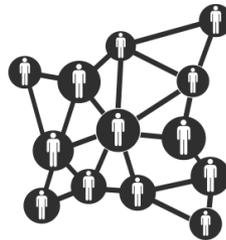
DHs and DLs: same practices

We can see **Digital Humanities** as a **community of practices** (textual editing, data modeling, creation of structured and enriched data...), aiming at **generating new knowledge**, offering services (**access** to content, tools, methods)

Digital Libraries have more or less **the same practices**; they also need to gain a **better knowledge of their digital assets** (which also implies to supplement catalog information with **distant reading**); they have a long history in providing **access** to content



Digital Scholars:
research purposes



Digital Curators & Mediators: insights
on the digital collections



Digitisation Managers and Experts: knowledge on the
digital collections

Example of common practices

The Retronews heritage press archive (BnF public/private partnership) makes heavy use of semantic treatments (named entity recognition, topic modeling, historical events extraction, article separation) to enhance information retrieval performances (for end-users)

The screenshot displays the Retronews search interface. At the top, there are navigation links: 'Contenu', 'Retronews Configuration', and 'Mes brouillons'. Below this, a search bar shows '484 résultats' and a dropdown menu for 'TRIÉ PAR PERTINENCE'. The left sidebar contains filters for 'FILTRER VOTRE RECHERCHE' (Par titre de presse, Par type de presse, Par périodicité, Par date de publication, Par lieu de publication) and 'SERVICES PREMIUM AFFINER VOTRE RECHERCHE' (Par thématique, Par évènement, Par sujet, Par personne). The 'Par personne' filter is highlighted with a red circle and lists: CAILLAUX (237), CALMETTE (101), POINCARÉ (88), LABORI (70), and JAURÈS (65). The main content area shows search results for 'LE GAULOIS' (N° 13430 P.1, 22 JUILLET 1914) and 'LE MATIN' (N° 10984 P.1, 25 MARS 1914). The article text for 'LE GAULOIS' is visible, mentioning 'M. Félix Belle' and 'M. Caillaux'. Below the article, there are social media sharing icons. The bottom of the page shows 'LE JOURNAL' (N° 7843 P.1, 18 MARS 1914) and 'L'ÉCHO DE PARIS' (N° 10935 P.1, 22 JUILLET 1914).

Example of common practices

OCR Post-correction

- Improvement of OCR can occur in the DLs digitisation workflow and benefit to all users
- Or each DH project must assume responsibility for it...
- In an ideal world, only specific research projects should spend time and money on OCR improvement (rare scripts, ancient typography, high quality requirements, crowdsourcing project on specific corpus, etc.)

Post-OCR Text Correction

Home Challenge Dataset Evaluation

ICDAR2017 Competition on Post-OCR Text Correction

News:
Result submission deadline: June 29, 2017 (midnight UTC)

The accuracy of Optical Character Recognition (OCR) technologies considerably impacts the way digital documents are indexed, accessed and exploited. During the last decades, OCR engines have been constantly improving and are today able to return exploitable results on mainstream documents. But in practice, digital libraries have on shelves many transcriptions with a quality below expectation. In fact, ancient documents with challenging layouts and various levels of conservation such as historical newspapers still resist to modern OCRs. Moreover, formerly digitized resources processed with out-dated OCRs are rarely re-sent through the latest state-of-the-art digitization pipeline, as priority is often given to the ever-growing masses of new arriving documents. In this context, OCR post-correction approaches, either used on former digitized documents or on fresh challenging documents, could strongly benefit digital libraries.

Find and correct OCR errors



OCR-ed text

The law in that case was severe, for cowards and runaways were not only degraded from all honors, but it was also a disgrace.

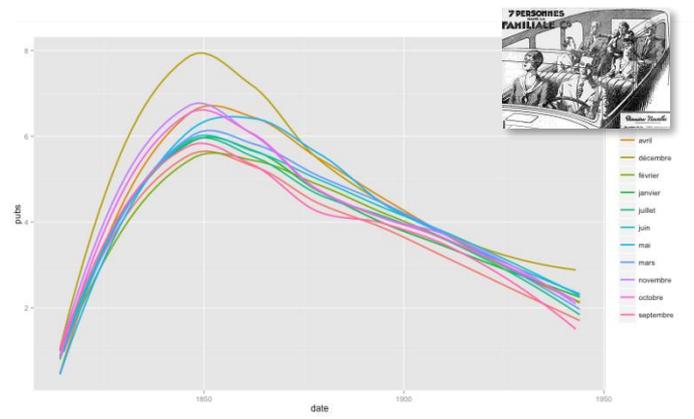
Aligned Gold Standard
The law in that case was severe, for cowards and runaways were not only degraded from all honors, but it was also a disgrace.

Challenge:
Find and correct OCR errors

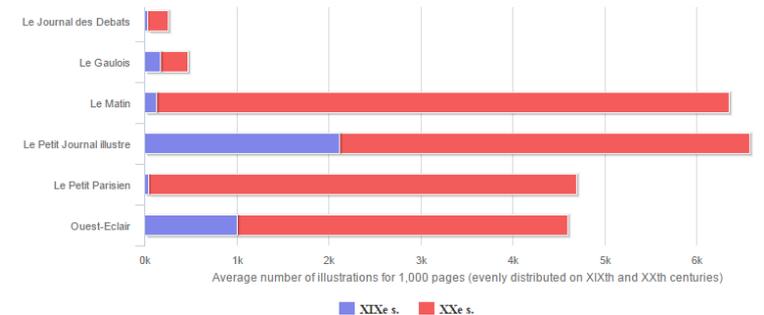
Example of common practices

Image retrieval in newspapers

- **Image bank project@BnF:** What titles contain illustrations? What is the total amount of images we can expect?
- **History of ads in the French dailies @GRIPIC/CELSA:** Where are the ads? Are they illustrated? What is the impact of Christmas on ads?



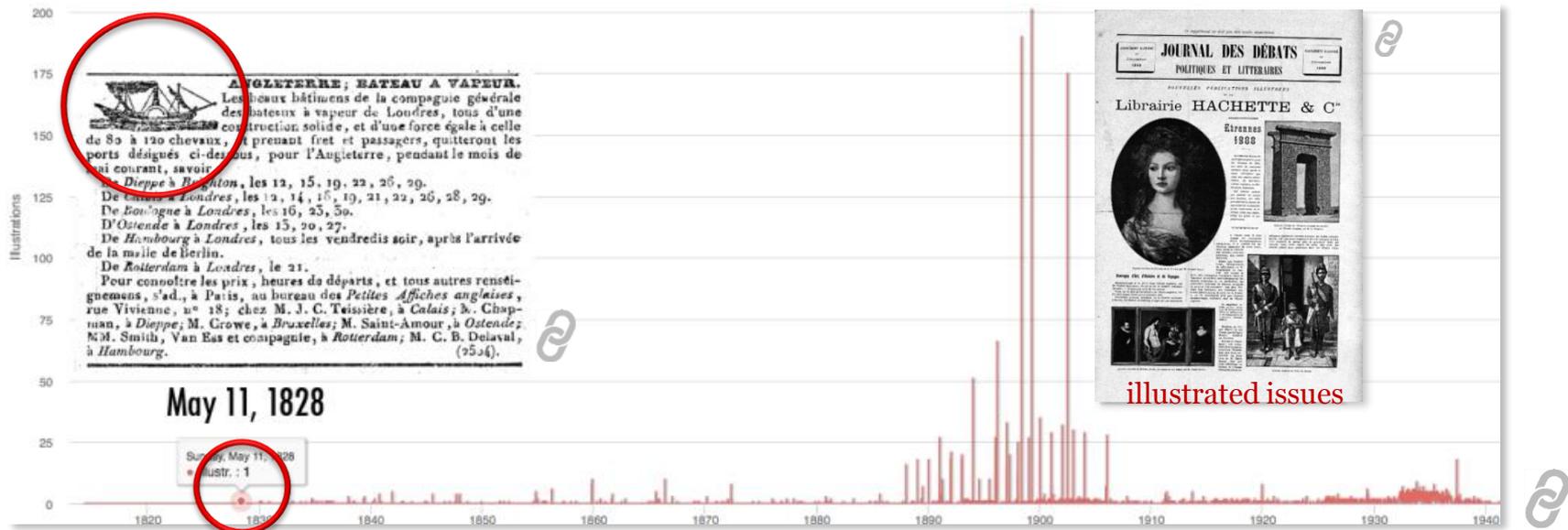
Using the same quantitative metadata set



Using the BnF Newspapers quantitative metadata set

Example of common practices

Image retrieval in newspapers: dataviz helps to spot the first published illustration (which turns out to be an advertisement). For the researcher, it's a fact/data. For the digital librarian, it's also an interesting fact that can be reused for digital mediation actions.



Journal des débats politiques et littéraires, 1814-1944, 45,922 fascicules (number of illustrations/issue)

Example of common practices

Image retrieval in newspapers and artificial intelligence techniques

- **Image retrieval project@BnF:** How can we automatically classify image genres and suppress noisy illustrations from newspapers?
- **History of ads in the French dailies@GRIPIC/CELSA:** How can we automatically extract illustrated ads from newspapers?

Drawings (2024)



Photos (2449)



Advertisings (364)



Scores (616)



Comics (212)



Handwritings (64)

Mariami à la Coea
bats calm vint Gringue
à tot, la tante! GORD.
ROCHARD

Engravings (1133)



Maps (282)



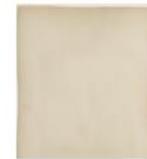
Ornaments (35)



Covers (86)



Blanks (178)



Texts (378)

« ...malheureux, rassurés que des vag
normes s'abattirent sur les navires,
après autrichien Androssy et divers
tes navires, subirent de graves avari
se quais furent balayés, un épais nu
ouvrit Messine.
On entendit des cris épouvantables,)
a silence tragique se fit.
Au lever du soleil, le désastre apparut d
sité son horreur. La ville entière s'é
les qu'un amas de décombres d'où sui
sient seulement les murs de l'Hôtel-de-V

BnF Image Retrieval PoC

http://altomotor.github.io/Image_Retrieval

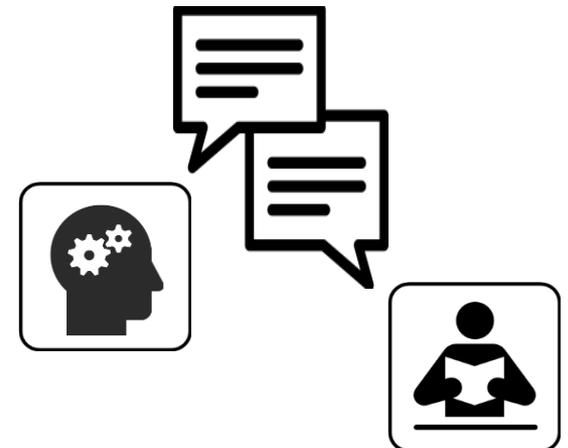
BnF Bibliothèque
nationale de France

Using a machine learning model (Google Inception-v3, deep convolutional neural network) trained on a heritage image dataset

Conclusion

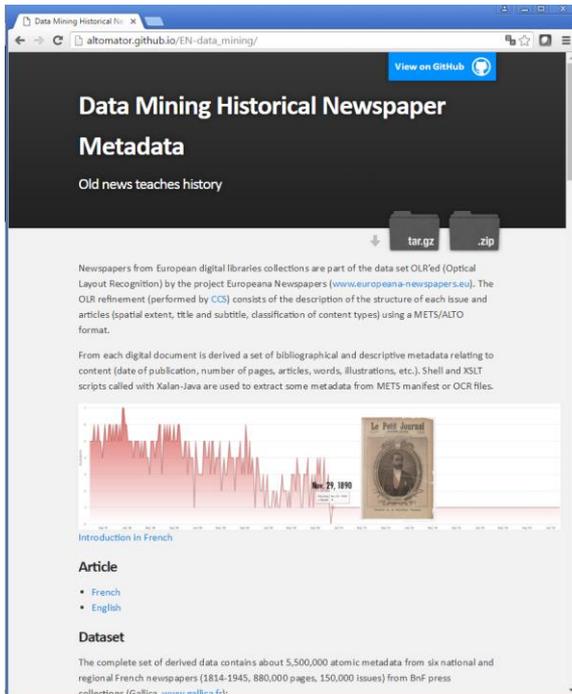
The opportunity to set up an DH lab for researchers and library professionals is becoming increasingly urgent.

- IT mining **infrastructure**
 - for **researchers**
 - for **in-house use cases**
- Physical space with **human resources**
- **Education and training** (for DL employees and DH scholars & students)
- Sharing **skills** and **know-how** (both ways)
- **Advice** on legal and organisational aspects



Thank you for your attention!

Datasets and scripts are publicly available. Just play with it!



The screenshot shows the GitHub repository page for 'Data Mining Historical Newspaper Metadata'. The page title is 'Data Mining Historical Newspaper Metadata' and the subtitle is 'Old news teaches history'. There are download buttons for 'tar.gz' and '.zip'. The main content describes the dataset, which consists of derived data from six national and regional French newspapers (1814-1945, 880,000 pages, 150,000 issues) from BnF press collections (Gallica). It mentions the use of Optical Layout Recognition (OLR) and machine learning techniques for metadata extraction.

Data Mining Historical Newspaper Metadata
Old news teaches history

tar.gz .zip

Newspapers from European digital libraries collections are part of the data set OLRed (Optical Layout Recognition) by the project Europeana Newspapers (www.europeana-newspapers.eu). The OLR refinement (performed by CDS) 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.

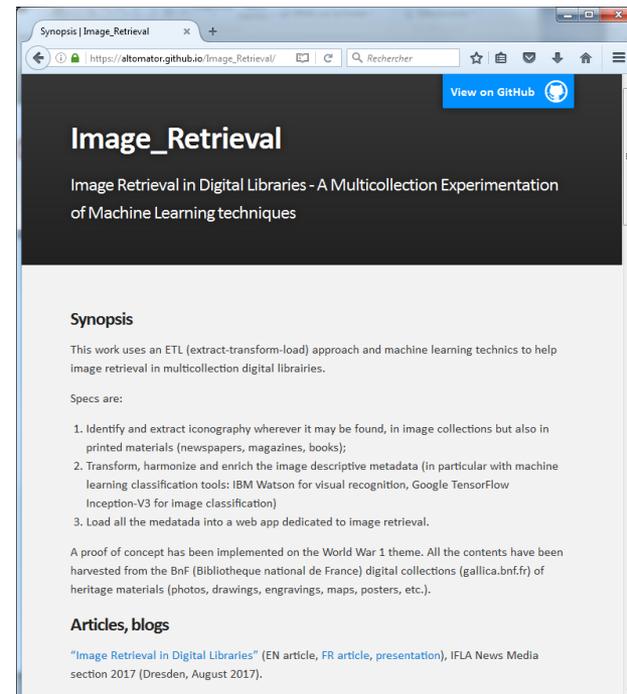
Article

- 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



The screenshot shows the GitHub repository page for 'Image Retrieval'. The page title is 'Image Retrieval' and the subtitle is 'Image Retrieval in Digital Libraries - A Multicollection Experimentation of Machine Learning techniques'. The main content describes the work, which uses an ETL (extract-transform-load) approach and machine learning techniques to help image retrieval in multicollection digital libraries. It lists three specifications: 1. Identify and extract iconography wherever it may be found, in image collections but also in printed materials (newspapers, magazines, books); 2. Transform, harmonize and enrich the image descriptive metadata (in particular with machine learning classification tools: IBM Watson for visual recognition, Google TensorFlow Inception-V3 for image classification); 3. Load all the metadata into a web app dedicated to image retrieval. A proof of concept has been implemented on the World War 1 theme. All the contents have been harvested from the BnF (Bibliothèque nationale de France) digital collections (gallica.bnf.fr) of heritage materials (photos, drawings, engravings, maps, posters, etc.).

Image Retrieval
Image Retrieval in Digital Libraries - A Multicollection Experimentation of Machine Learning techniques

Synopsis

This work uses an ETL (extract-transform-load) approach and machine learning techniques to help image retrieval in multicollection digital libraries.

Specs are:

1. Identify and extract iconography wherever it may be found, in image collections but also in printed materials (newspapers, magazines, books);
2. Transform, harmonize and enrich the image descriptive metadata (in particular with machine learning classification tools: IBM Watson for visual recognition, Google TensorFlow Inception-V3 for image classification)
3. Load all the metadata into a web app dedicated to image retrieval.

A proof of concept has been implemented on the World War 1 theme. All the contents have been harvested from the BnF (Bibliothèque nationale de France) digital collections (gallica.bnf.fr) of heritage materials (photos, drawings, engravings, maps, posters, etc.).

Articles, blogs

"Image Retrieval in Digital Libraries" (EN article, FR article, presentation), IFLA News Media section 2017 (Dresden, August 2017).

http://altomator.github.io/Image_Retrieval