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

Jean-Philippe Moreux
Bibliothèque national 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...


EXCELSIOR

L'ÉTAT DU MARÉCHAL S'EST LÉGÈREMENT AMÉLIORÉ TOUTEFOIS LE PRONOSTIC DEMEURE AUSSI SÉRIEUX

UN IMPORTANT CONSEIL DES MINISTRES

M. BASCOM SLEMP A VISITÉ HIER LA SECTION DES ÉTATS-UNIS A L'EXPOSITION COLONIALE

LA CIRCULATION DES PIÉTONS SUR LA CHAUSSÉE

L'Ouest-Eclair

JOURNAL RÉPUBLICAIN DU MATIN

LES MYSTÈRES DE L'INSTINCT

Terrible collision d'autos

TROIS TUÉS. QUATRE BLESSÉS

Un halluciné est trouvé mort dans un champ

UN CALVAIRE BRETON va être transporté en Belgique

APRÈS LA RÉVOLTE DE LISBONNE

SOURIRES

LA POLITIQUE INTERNATIONALE

A LA HAYE La Cour de Justice condamnerait l'Anschluss

A GENÈVE Les travaux de la Société des Nations

VERS UNE CONCORDE CRÉATRICE

LA RECONDUCTION DU BUDGET

Le raid du Zeppelin vers l'Amérique du Sud

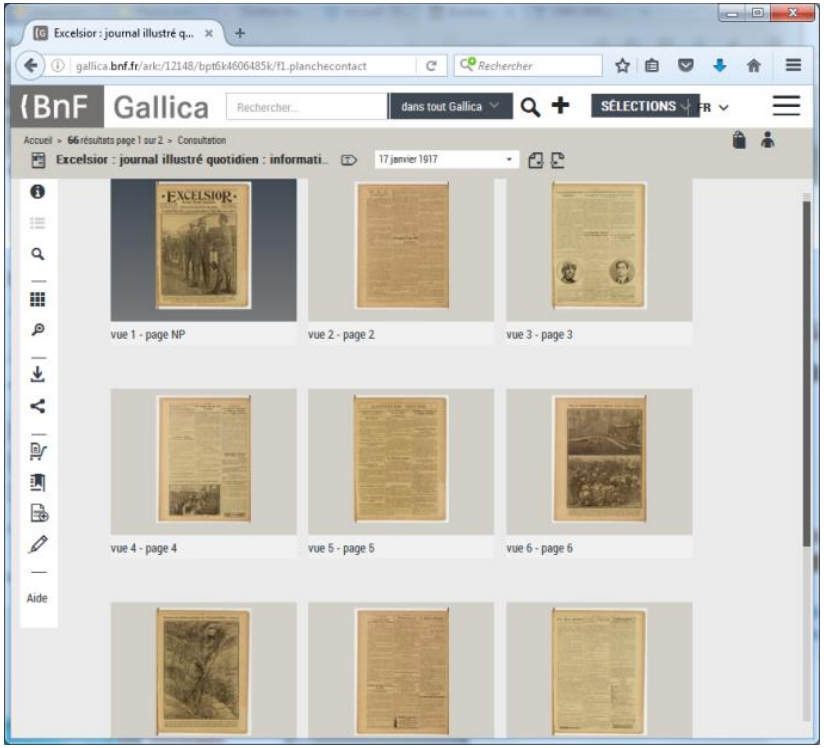
UN GANGSTER ASSASSINE PAR DES BANDITS A SA SOLDE

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 OCRed 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)

Here, and only here

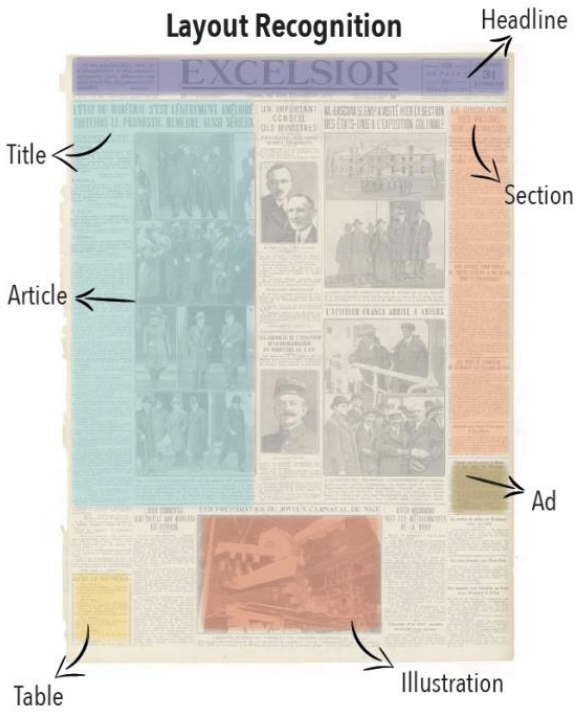
Article separation, Layout recognition

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



What is OLR?

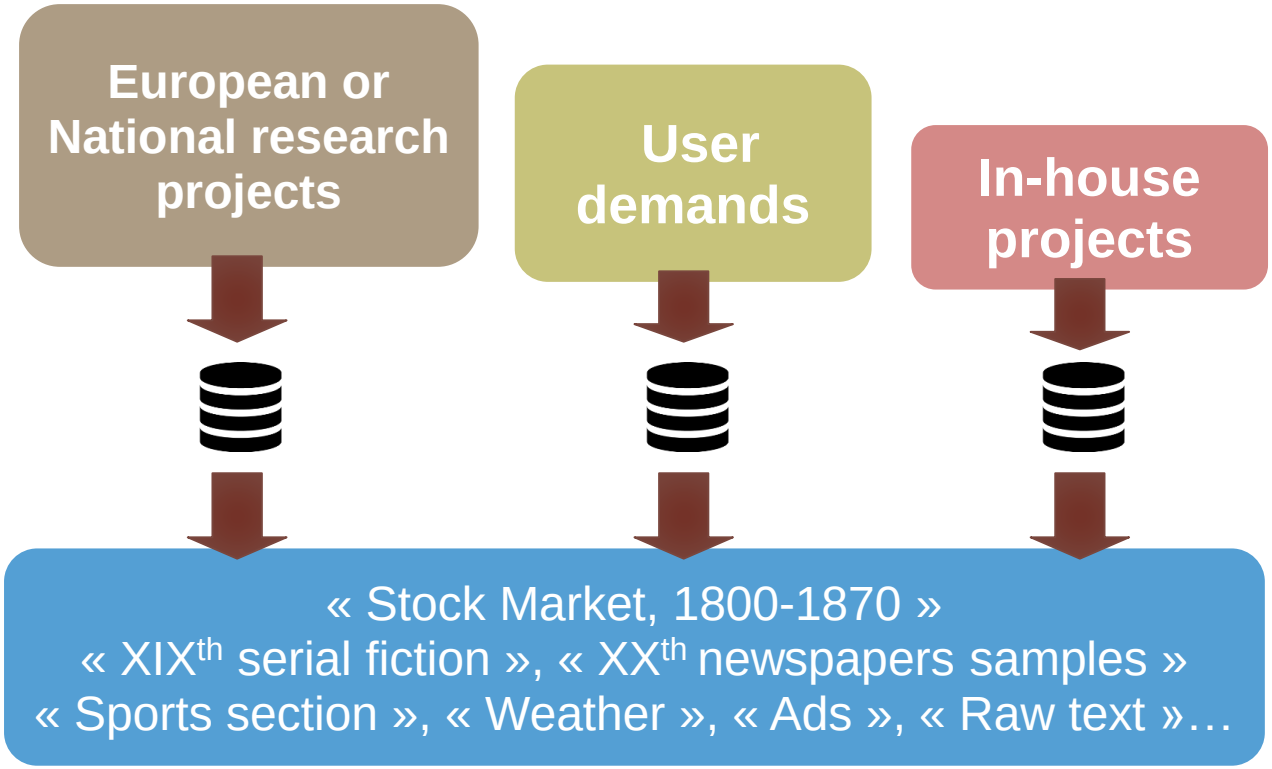
- Identification of structural elements, including separation of articles and sections
- Classification of types of content (ads, offers, obituaries…)

europeana newspapers

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

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
- Your don’t need to **parse millions** of XML files
- Your dataset is **research friendly** (XML, JSON, not PDF…)
- The dataset **format** fullfil 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…

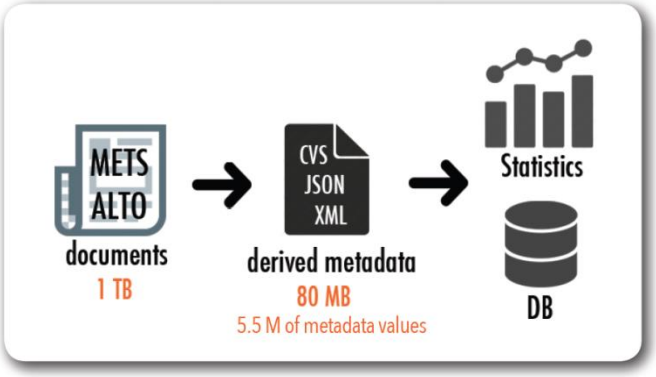
➔ 880k pages from Europeana Newspapers OLRed corpus

➔ 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

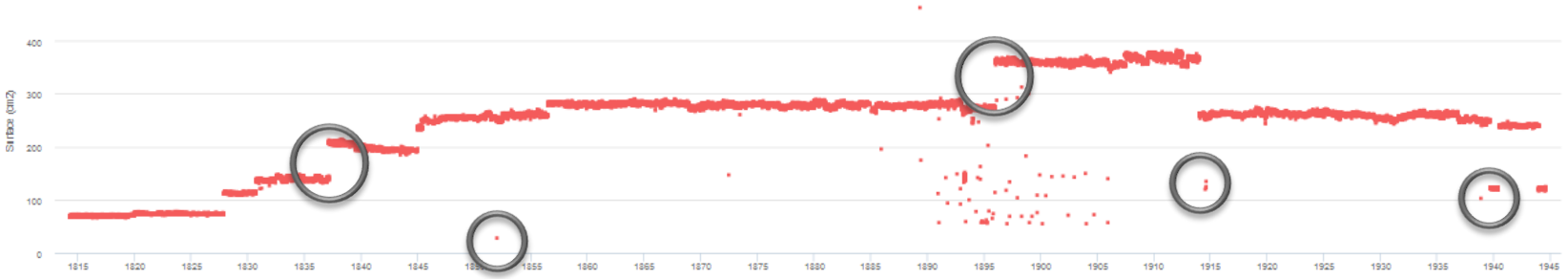
BnF Bibliothèque nationale de France



Quantitative Metadata Analysis

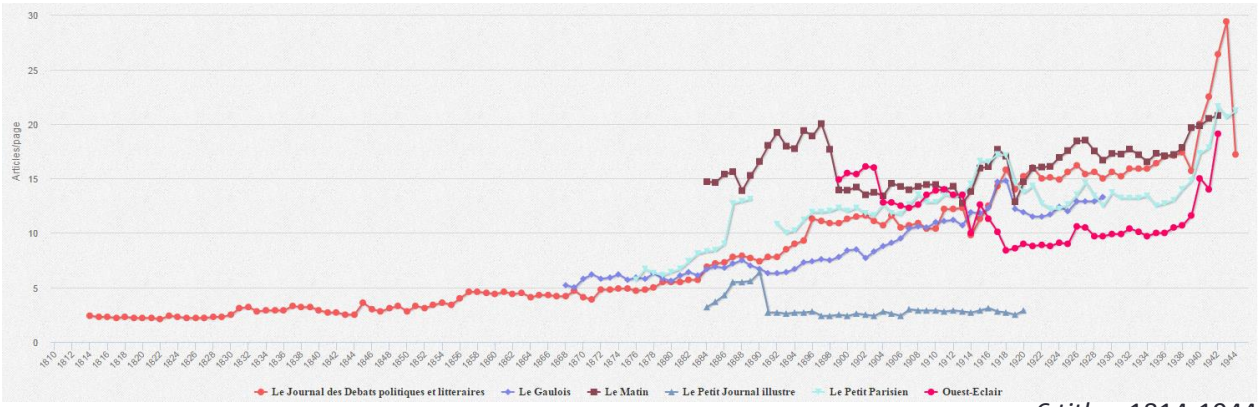
Now we can perform quantitative analysis and dataviz.

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



Journal des débats politiques et littéraires, 1814-1944

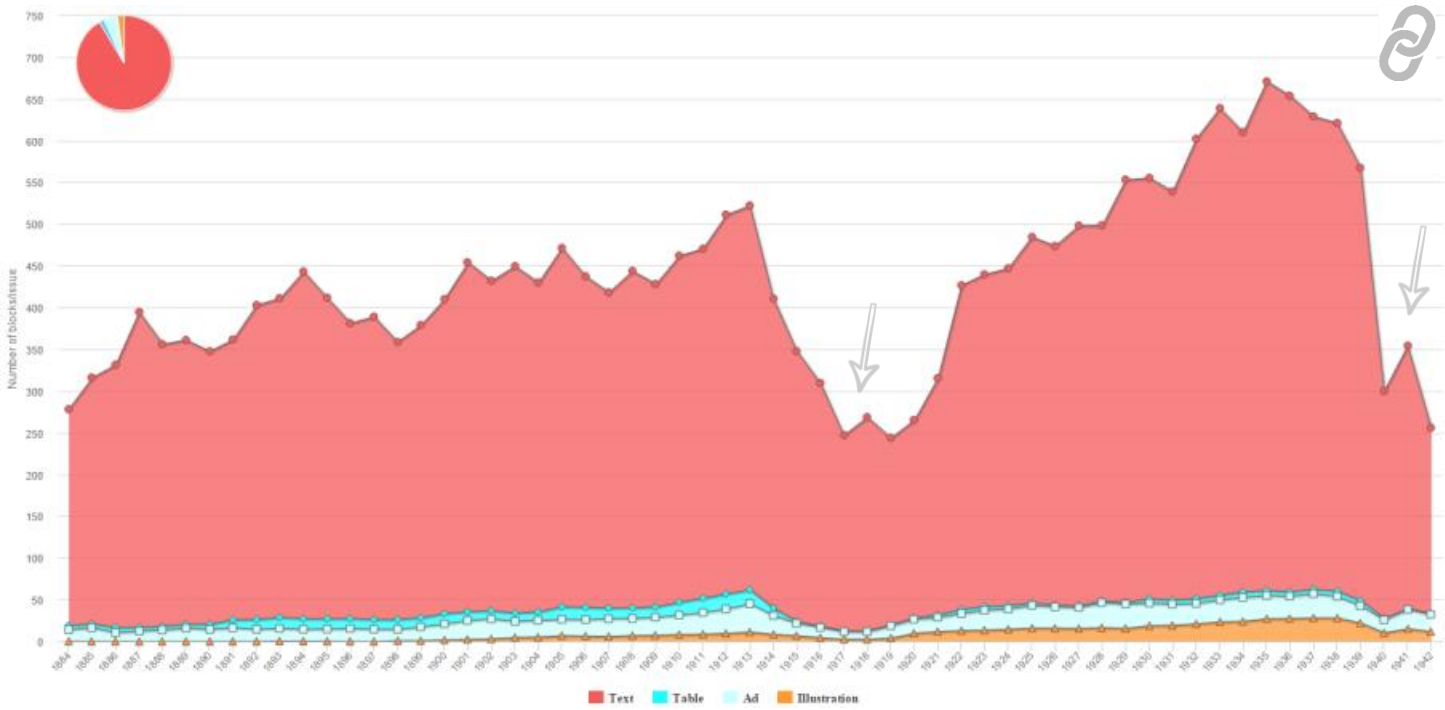
- **History of press/layout:** Visualization of the articles density per page reveals the shift from XVIIth "gazettes" to modern dailies.



6 titles, 1814-1944

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).

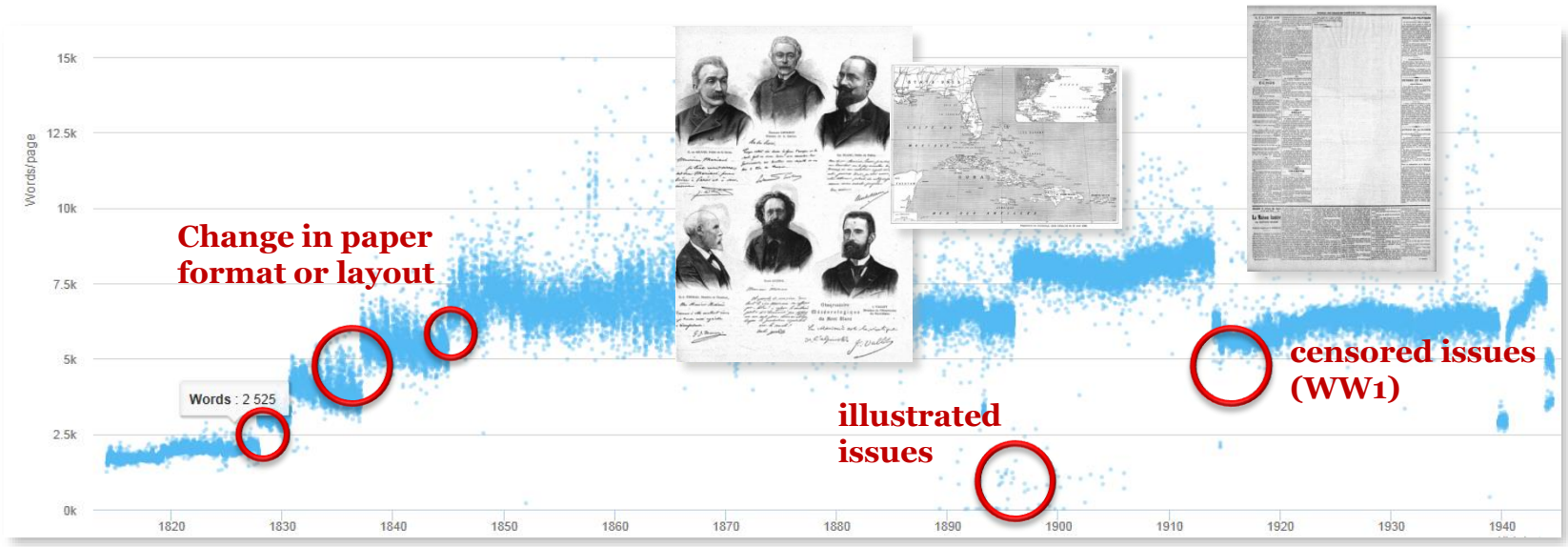


Le Matin, 1884-1942, types of content (articles, text blocks, tables, illustrations, ads)

Quantitative Metadata Analysis

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

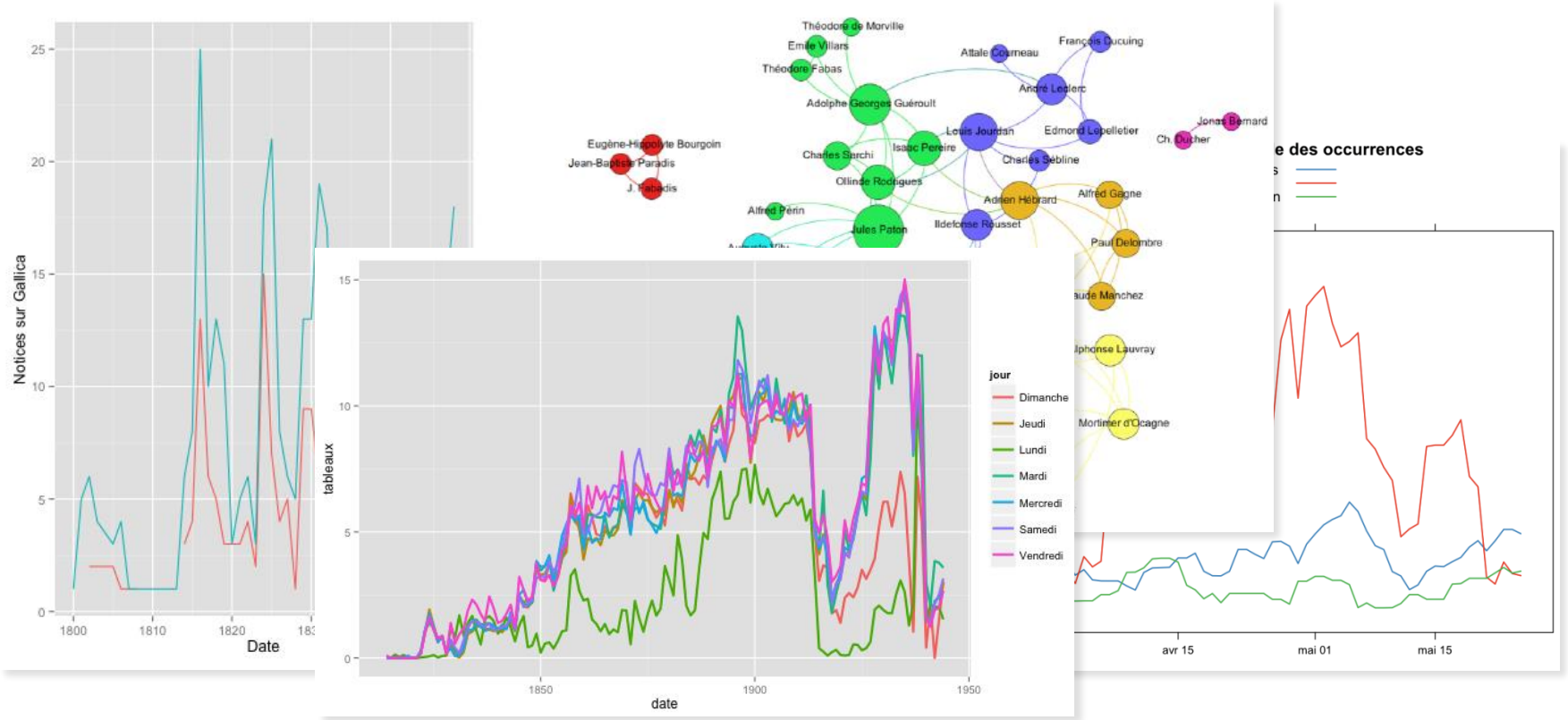
Close reading (links to gallica.bnf.fr) / **distant reading**



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

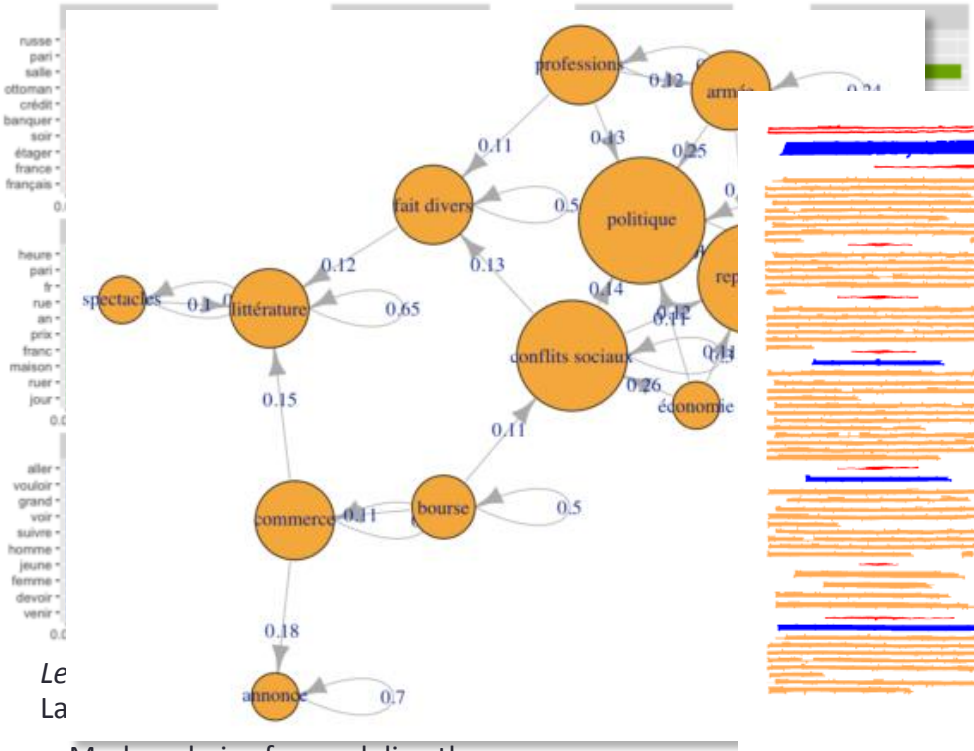


CELSA/GRIPIC, univ. Paris-La Sorbonne
"Stock Market quotes creation and development in French newspapers" (1800-1870), P-C Langlais, PhD in Information Sciences, 2015

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

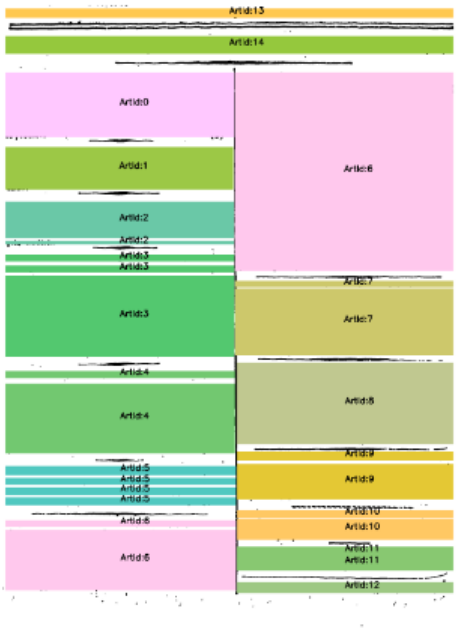
What if you don't have Article Separation?

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

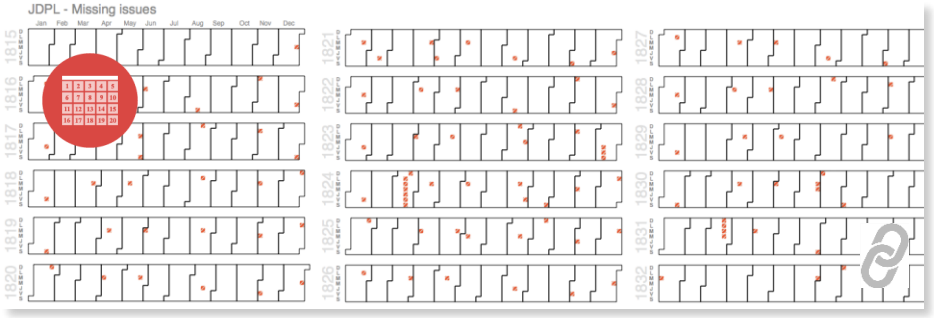
P-C Langlais, https://numapresse.hypotheses.org

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

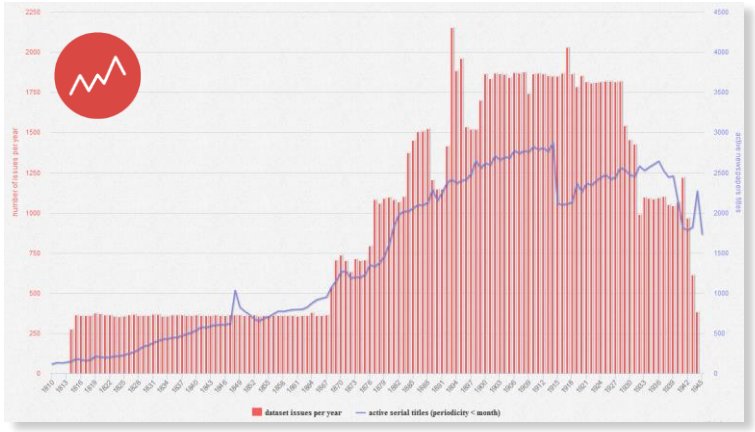
A calendar display of a newspapers title data shows rare missing digital issues, which suggests that the digital collection (for this title) is rather complete.



© Google Charts API

Le Journal des Débats

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.

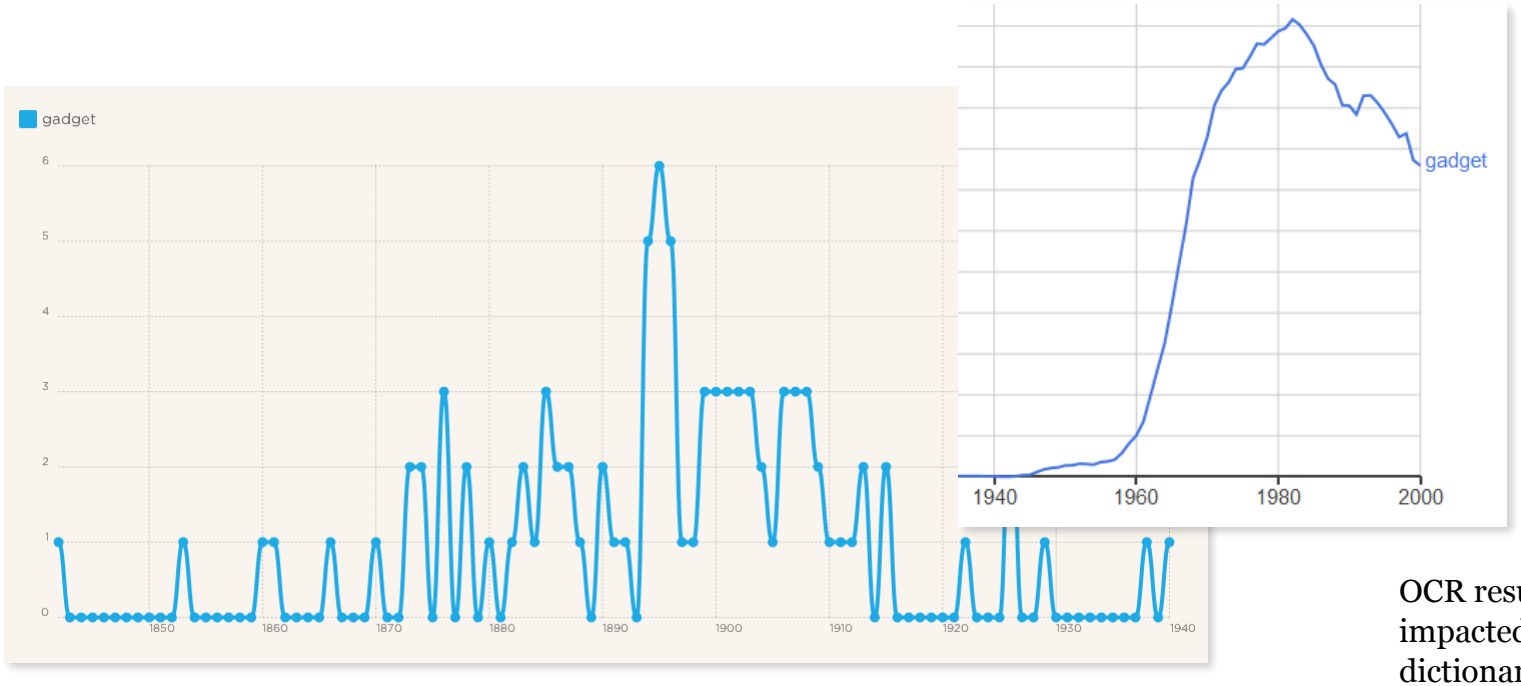


BnF Newspapers dataset

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**

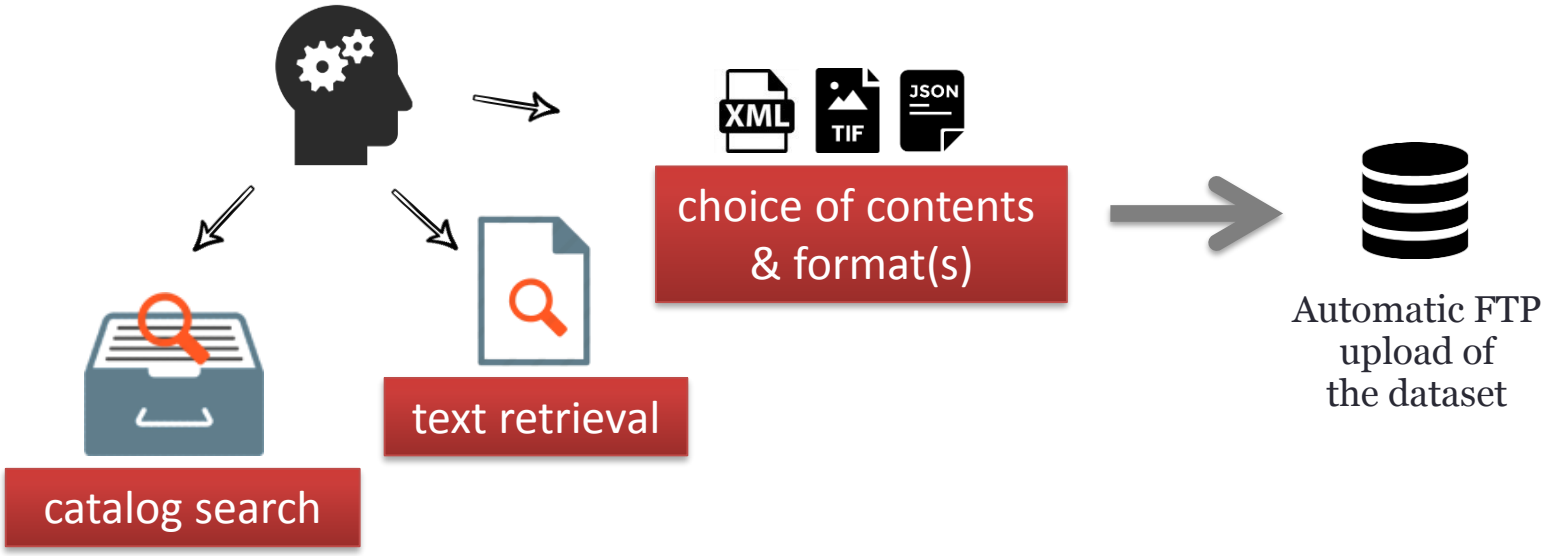


The “**gadget**” in this corpus (1840-1940, retronews.fr) are actually false positives of “**bugdet**”

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

On-demand Datasets

- Some users have **specific** needs, particulary 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)

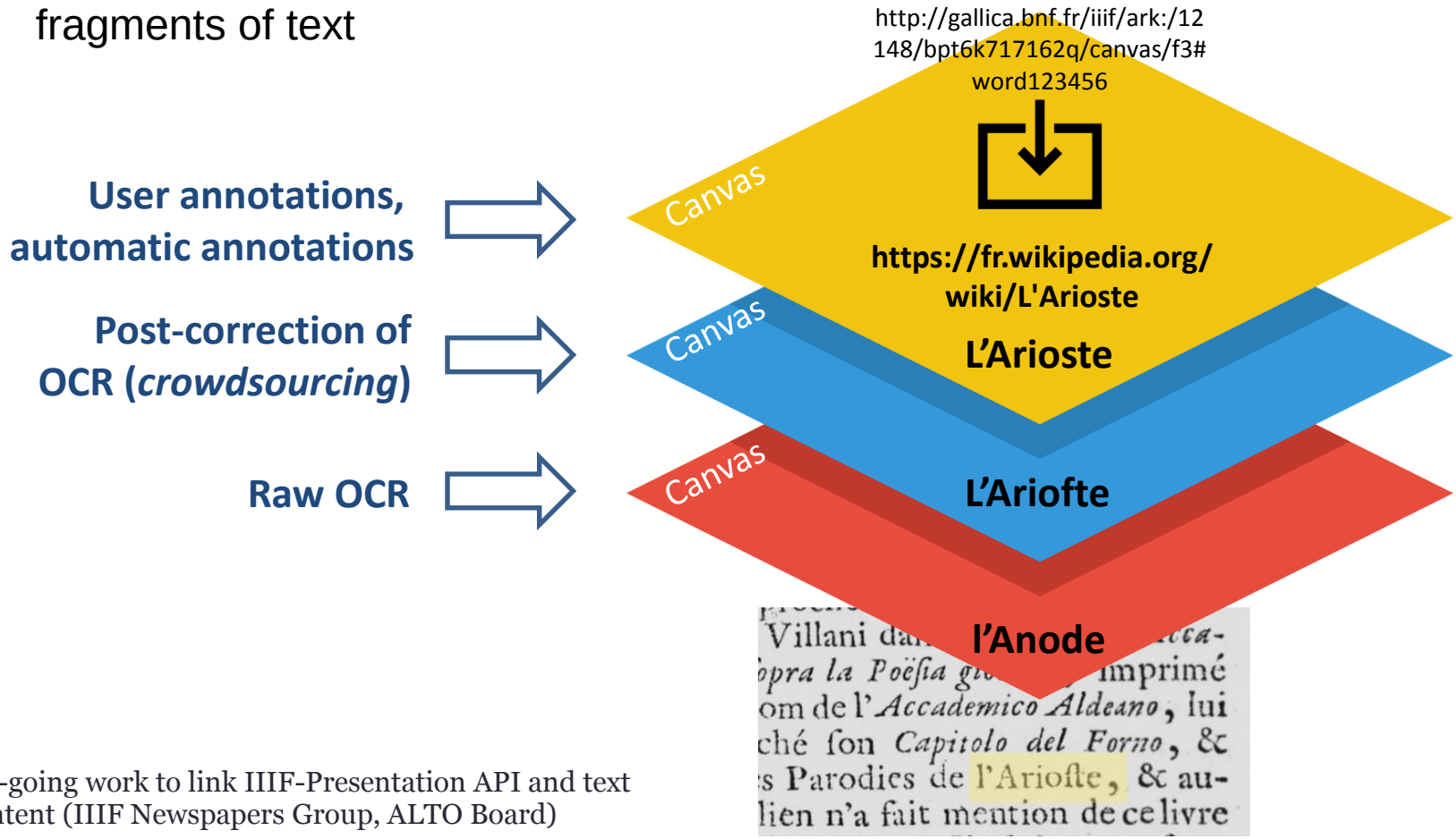
Machine-operable access to content for on-line dissemination:

- ✓ 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)

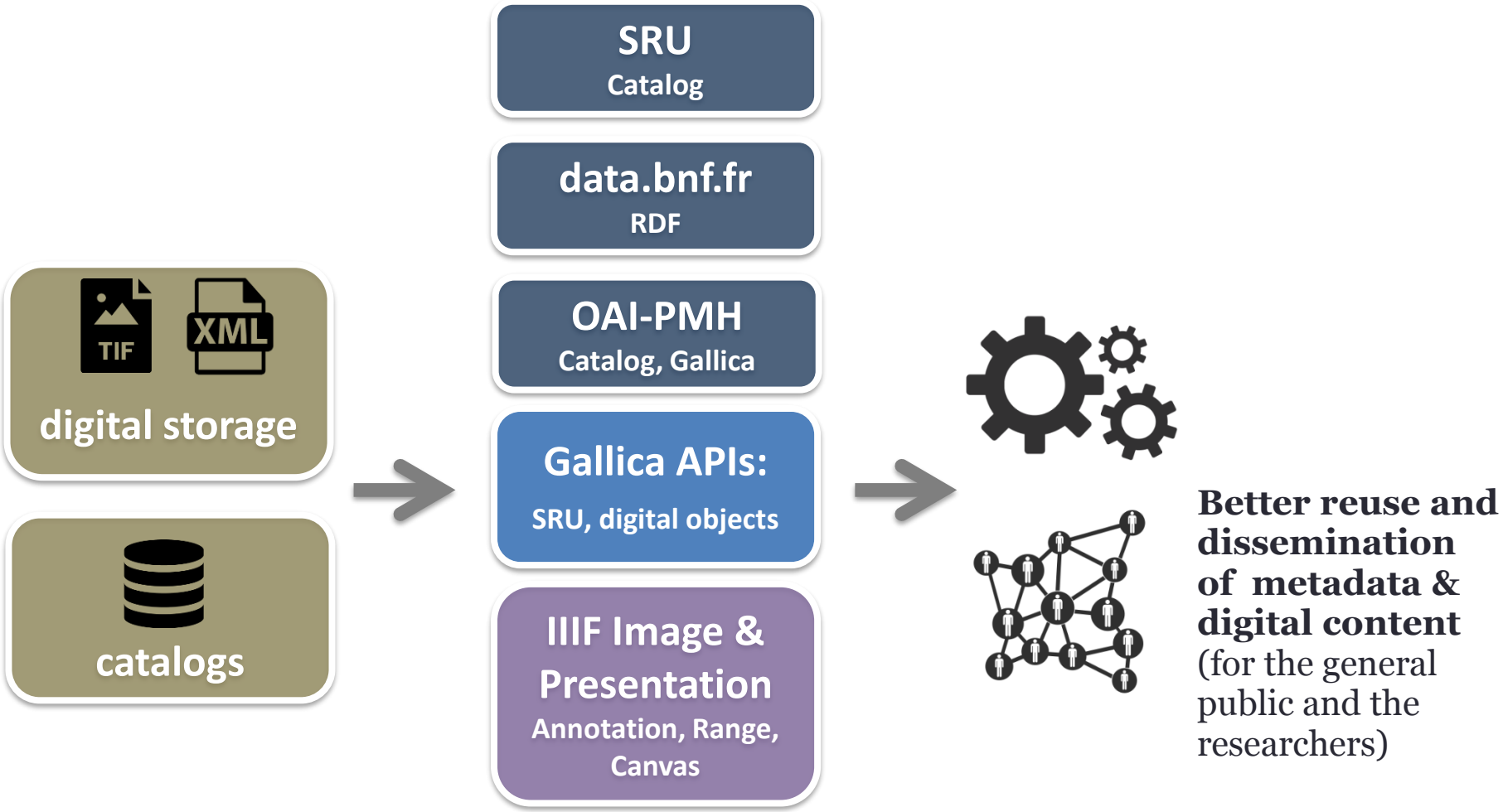
IIIF example

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



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

APIs @ BnF


TIF
XML
digital storage
catalogs
SRU
Catalog
data.bnf.fr
RDF
OAI-PMH
Catalog, Gallica
Gallica APIs:
SRU, digital objects
IIIF Image & Presentation
Annotation, Range, Canvas
Better reuse and dissemination of metadata & digital content
(for the general public and the researchers)

APIs @ BnF

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

api.bnf.fr (November 2017)

2nd Edition: November 2017

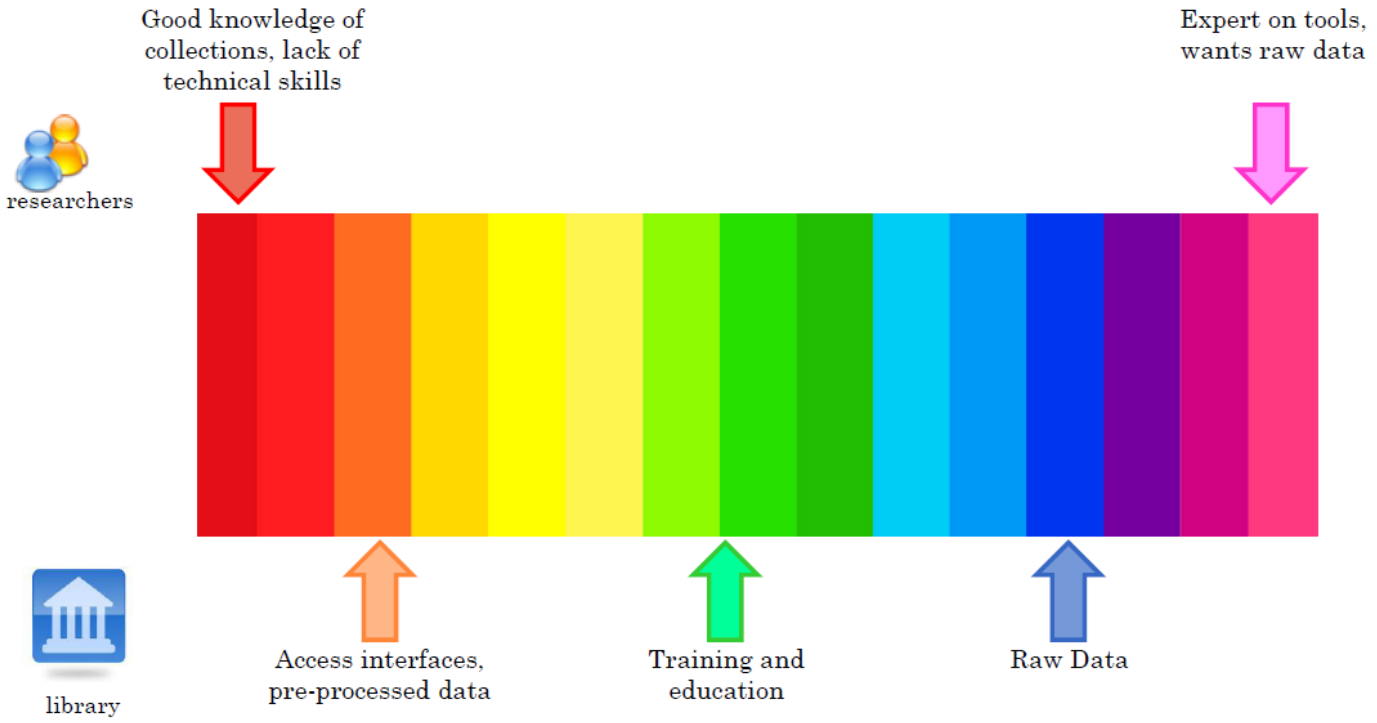
➔ Raise awareness among researchers about the availability of the DLs resources

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.**



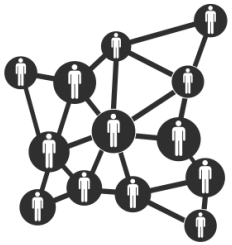
Emmanuelle Bermès, "Text, Data & link-mining in DLs: looking for the Heritage Gold", IFLA DH Satellite Meeting

DHs and DLs: same pratices

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

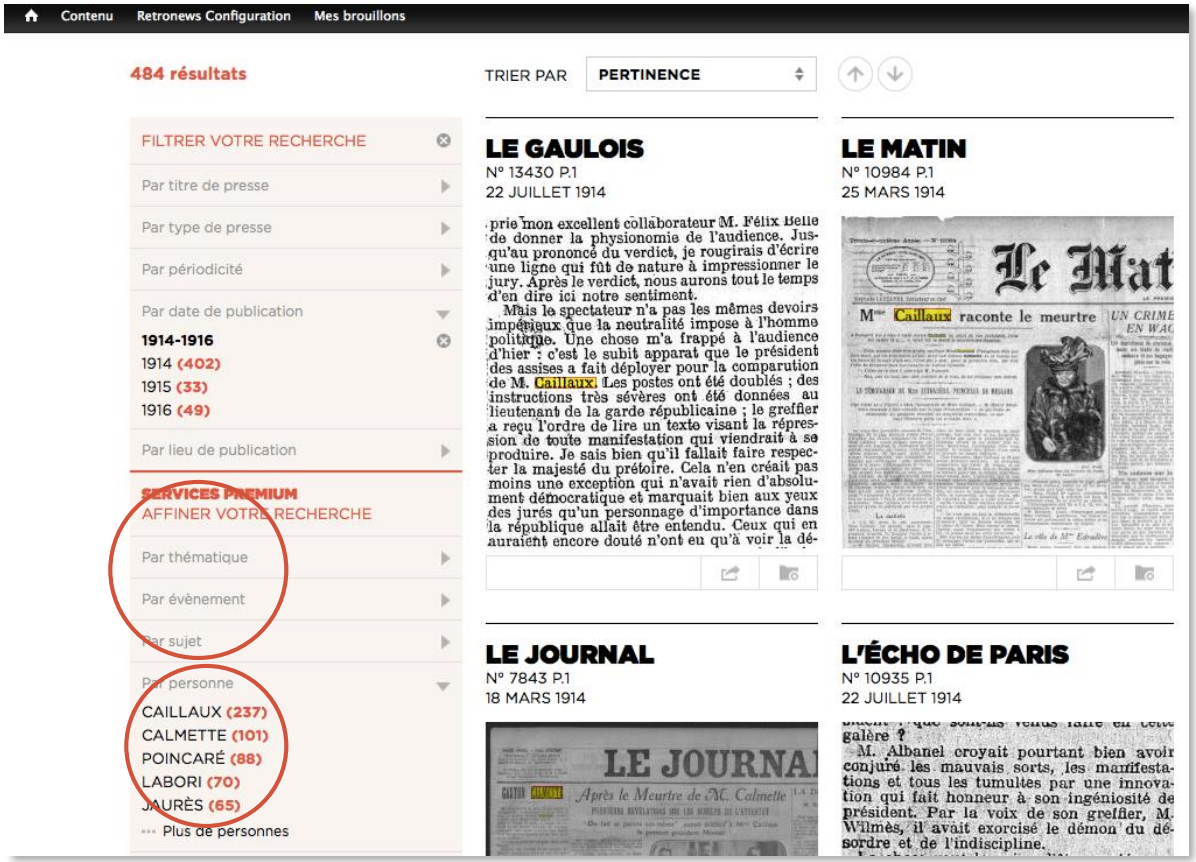


Digital Curators & Mediators: insights on the digital collections

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)

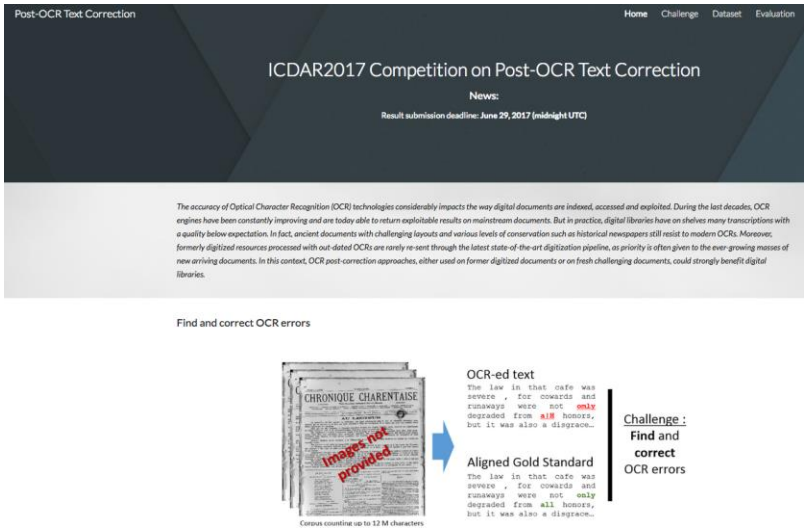


www.retronews.fr

Example of common practices

OCR Post-correction

- Improvement of OCR can occurred in the DLs digitisation workflow and benefit to all users
- Or each DH project must assume responsibility for it…
- In a ideal word, 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.)

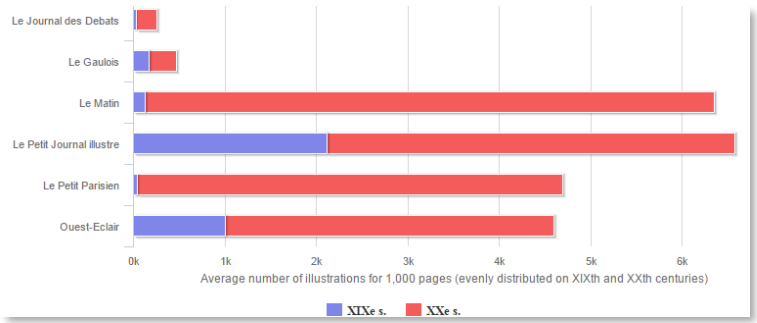


ICDAR Conference, Kyoto (November 2017)

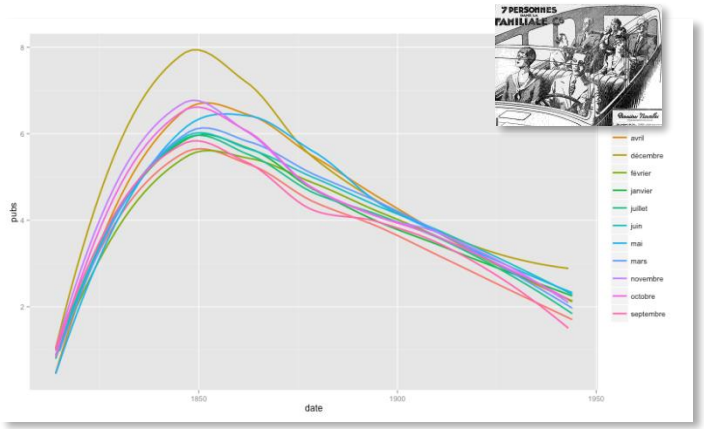
Example of common practices

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?



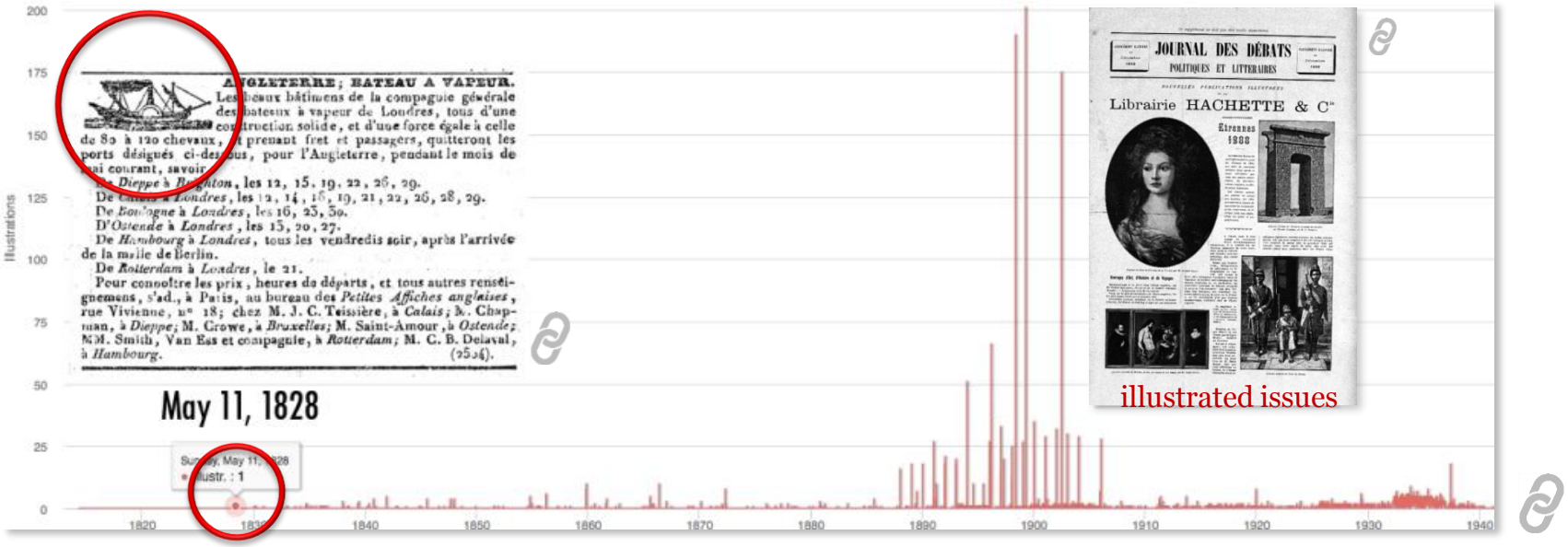
Using the BnF Newspapers quantitative metadata set



Using the same 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?

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

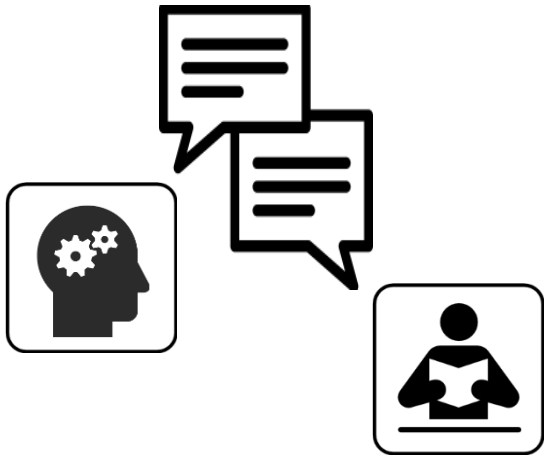
BnF Image Retrieval PoC
http://altomator.github.io/Image_Retrieval

BnF Bibliothèque nationale de France

Conclusion

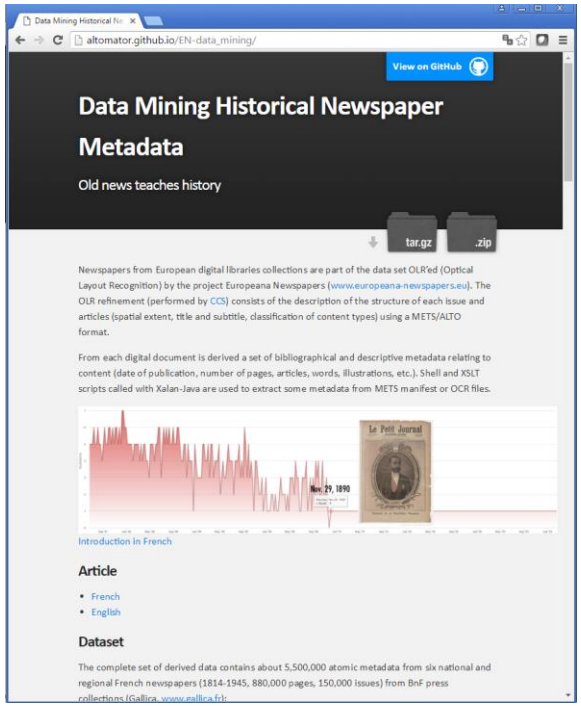
The opportunity to set up an DH lab for researchers and library professionals is becoming increasingly urgent.

- 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

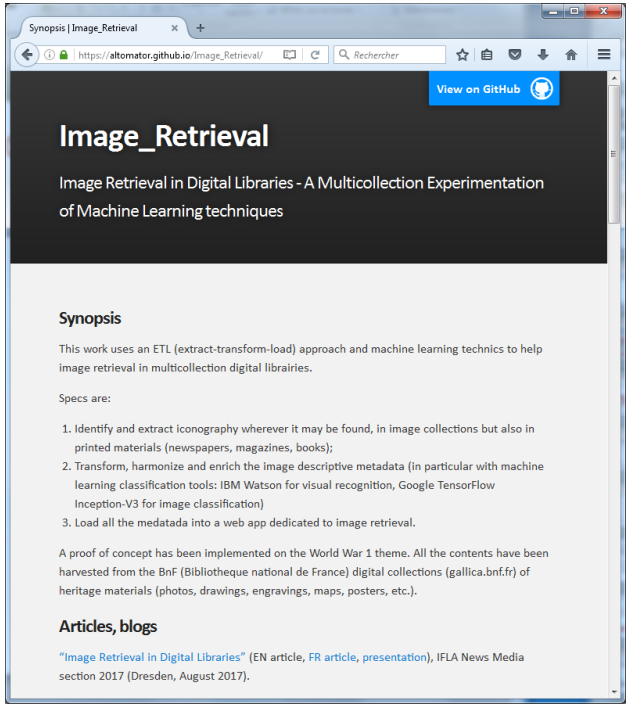


Thank you for your attention!

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



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



http://altomator.github.io/Image_Retrieval