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)

Here...

and here
A True Story about the Researchers’ Needs

- Obviously, he has to request the digital library catalog.
A True Story about the Researchers’ Needs

• And he needs a text retrieval functionality.
A True Story about the Needs of Researchers

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

+ 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**.

**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, like **articles** and **sections**.

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

---

*Image of a newspaper page with various annotations indicating layout recognition and structural elements such as headline, section, article, ad, illustration, and table.*
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

**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)
- Digital Curators & Mediators
- Digitization Program Managers
- DL Professionals
Digitization or IT People might be Interested by those Metadata

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**

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

© Highcharts

→ Better knowledge of the collection
Discovering Knowledge through Visualization

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

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

- **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 focus 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

The quantitative dataset is of a great help 😊

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

Types of content are tagged, like Stock Market

 ➔ 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)

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

Facts extracted thanks to dataviz can then enrich other digital objects like timelines.
Engaging new Audiences with Dataviz

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

⇒ 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 visualization can contribute to quality control.

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

⇒ "Inform end users about data quality!"
Querying the Dataset

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

- **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.

  ![Image search solution](http://localhost:8984/rest?run=findIllustratedPages.xq&fromDate=1900-01-01&toPage=1)

  ➜ "As a digital mediator, seeking for illustrations in our xM pages newspaper collection is a nightmare..."

  ➜ Let’s remember that newspapers are full of illustrations not described in our catalogs!
Advanced Search Mode for Newspapers

- 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”
Advanced Search Mode for Newspapers

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

  ![Trove Advanced Search](http://trove.nla.gov.au)

  **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

  ![Le Matin, Oct. 29th 1886](http://example.com)

  articles from 1886
  title OR subtitle con
  "statue & liberté"

  ![Cours des Metaux](http://example.com)

  Le Matin, Oct. 29th 1886
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.

?**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)

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!

- Fireman (NY police), *Le Matin*, 1932
- 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...
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!
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

Facetted search on bibliographical and semantical criteria

Timelines, Folders

Personal space
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

Faceted 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 length of the newspapers long enough to bring meaningful 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.
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 XIXth 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.
Thank you for your attention!

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

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

http://www.retronews.fr/

Thanks to all the Europeana Newspapers partners!