Text, data and link-mining in digital libraries: looking for the heritage gold

26-27.03.18

Emmanuelle Bermès
National Library of France
Deputy Director for services and networks

LIBRARY SCIENCE TALKS
Delivering BnF digital collections as data to researchers in order to allow them, on-site, to perform computational analysis of digital corpora.

CORPUS, a BnF research project

Part of the BnF’s 4-year internal research programme 2016-2019

Objectives:
- Designing a future service for providing access to digital corpora for researchers
- Providing researchers with data and tools to analyse them, in agreement with IPR and privacy legislation

3 years of experimentation (web archives, digitization, metadata) + 1 year for conclusions

Performance and strategic report 2017-2021

For further information:
http://c.bnf.fr/fom
DIGITAL COLLECTIONS AT THE BNF

Web archives
900 To

Gallica + Gallica intra muros
4.5 million digitized items
http://gallica.bnf.fr

Metadata
More than 20 million records
http://data.bnf.fr
The Europeana Newspapers Project

Jean-Philippe Moreux,
« Approches innovantes pour la presse ancienne numérisée : fouille et visualisation de données »,
published on the Carnet de recherche à la bibliothèque nationale de France,
3 December 2016
https://bnf.hypotheses.org/208
BEFORE CORPUS...

The corpus as a sandbox

David Picard, Philippe-Henri Gosselin, Marie-Claude Gaspard
Corpus from: http://images.bnf.fr

ETIS: the ASAP project

(BnF)
The corpus as an interface

Valérie Baudouin, Zeynnep Pelhivan,
Project « Le devenir en ligne du patrimoine numérisé : l’exemple de la Grande Guerre »
https://hal.archives-ouvertes.fr/hal-01425600

Project « Le devenir en ligne du patrimoine numérisé : l’exemple de la Grande Guerre »
Lab « Les passés dans le présent »

(BnF)

Figure 19. Carte du Web Grande Guerre (collecte de novembre 2014, sites avec degrés ≥30)
Partnership with Web90 (CNRS/ISCC)

Weblog: http://webcorpora.hypotheses.org/

« Once upon a time in the web: 20 years of Web archiving in France »
Event
Video recordings: http://c.bnf.fr/fse
Initially conceived as an experimentation, this interface is now integrated in the tools’ production process:

- Full-text indexing function
- Statistics and metadata extraction
- 3 corpora:
  - collections carried out by the BnF after the attacks of January and November 2015
  - News 2010 - 2017

Web archives Labs, an interface designed to meet researchers’ needs.
Agreement with the OBVIL Lab

Presentation by Didier Alexandre, OBVIL Director, at JPAC 2016:
http://c.bnf.fr/fuK

OBVIL website:
http://obvil.lip6.fr/tgb/

135,000 ALTO files delivered for text-mining
related projects

Adrien Nouvellet, Valérie Beaudouin, Florence D'Alché-Buc, Christophe Prieur, François Roueff. « Analysis of Gallica Usage Traces: A Log Files Analysis of Gallica Website. » Télécom ParisTech; National Library of France, 2017. hal-01709264
a wide array of different situations...

Good knowledge of collections, lack of technical skills

Access interfaces, pre-processed data

Training and education

Expert on tools, wants raw data

Raw Data

RESEARCHERS

LIBRARY
The Giranium project

Partnership with the GRIPIC - CELSA

“Girardin digitized or the birth of the media industry”: analysis of 300 autograph letters from Girardin preserved in the Institute and other documents (notes, newspapers, magazines ...).

- Need for a workspace (DEP)
- Workshops with BnF experts about the tools developed for the project
Digitized documents

**YEAR**

2 events

2 workshops

« Describe, transcribe and delivery heterogeneous corpora: methods, formats, tools »

Minute published online in the Carnet de la recherche de la Bibliothèque nationale de France
https://bnf.hypotheses.org/2214

« Geolocalisation: heritage materials in space and time »

Minute published online in the Carnet de la recherche de la Bibliothèque nationale de France
https://bnf.hypotheses.org/2299
The Néonaute project: a search engine to find neologisms in the web legal deposit collections – submitted to the DGLFLF's "Languages and Digital 2017" call

Experimental project with several potential partners (INRIA, INA, EPFL...) on image mining and pattern recognition
Le portail BnF API et jeux de données décrit et documente l'ensemble des API qui permettent d'interroger et de récupérer les métadonnées des catalogues (notamment BnF Catalogue général, data.bnf.fr, Gallica) et les collections numérisées de la BnF. Pour faciliter l'accès aux données et leur utilisation, des jeux de données (images et textes, métadonnées, statistiques) sont directement téléchargeables via le portail.

Plusieurs formats, plusieurs technologies permettent de répondre à la diversité des usages des données de la bibliothèque : alimentation de catalogues, création de nouveaux services innovants, fuite de données, datavisualisation, etc.

Développeurs et développeuses, chercheurs et chercheuses, auteur et actrices du monde de la culture et de la chaîne du livre, digitaux humanists ou encore amateurs et amantes de culture, les données et métadonnées diffusées par la BnF n’attendent plus que vous !

Actualités

**Journée d'études : Quelles contributions à la circulation des données ? retrospective 2017**
(28 mars 2018)

**Gallica Studio : innovation et créativité autour des collections de Gallica**

**A propos du BnF API et jeux de données**

http://api.bnf.fr/
The Corpus project and its potential users.

A PROSPECTIVE STUDY OF USERS’ NEEDS AND EXPECTATIONS.

YEAR

2

study

By Eleonora Moiraghi
Research assistant
working on the Corpus project
January 2018
A QUALITATIVE RESEARCH

INTERVIEWS

3 survey populations
16 scholars, 11 members of staff and 3 subject-matter experts to determine potential users’ needs and expectations by adopting a qualitative approach

2 SYMPOSIA

networking events in order to encourage the Library and the research community to interact and share ideas, as well as to take note of informal observations

UX WORKSHOP

using the persona method
6 scholars and 15 members of staff to confirm the needs emerging during the interviews
FOCAL AREAS

Text and data mining
Digital corpora use or delivery and difficulties encountered
Similar services
Definition of the role of the BnF in the academic ecosystem
A physical space in the Library
Space organisation and friendliness
IT infrastructure, tools and functionalities
Support
TDM contributes to increase knowledge. TDM is not an exclusive research approach but it will have an impact and add value across all disciplines.
« The problem with a cross-disciplinary approach is the exploitation of one discipline by the other. The main problem for the computer engineer is testing the effectiveness of his algorithms. To us, the focus is not there. »

« Often one has to spend a lot of time cleaning data. OCR is prone to a huge amount of errors... ». 
Il faut vraiment investir dans le personnel.

New York Public Library Labs?
MIT Libraries?
Institut national de l’audiovisuel?
Scholars and subject-matter experts hesitate
Or have yet to determine a clear answer.
«The library will play a major role in creating a community that doesn't exist in the academic system.»
“Today, the only real reason to move is to go and meet the human expert. We spend a lot of time gathering information but an expert would guide us through this much more efficiently.”

“...it’s become more and more evident that opening data doesn’t simply mean leaving data on the Web... [...] To understand data, it’s necessary to know its history... [...] Having a lot of documentation online isn’t enough, you need a person.”

Every scholar agreed on the desirability of a virtual laboratory from the beginning of the project. Nonetheless, the pros of a physical space emerged later:

- Working spaces;
- Training;
- Advice;
- Sharing.
« Everyone knows that the most important times during a seminar are coffee breaks. In order to encourage interactions between human beings, a pleasant meal or a coffee are good catalysts. »

A setting that caters for all the needs of future users would include:

- Spaces reserved for the BnF staff
- A relaxing coffee area
- Rooms provided with screens for teamwork and training
- Individual rooms
- Offices for professors and researchers-in-residence
- An events room to accommodate about sixty guests.

A friendly, multi-functional, flexible and ever-changing space.
« I only work on my laptop. All of my work is stored in a single device. It’s unbearable for me to install the software that I already installed elsewhere. »
The persona method helped to identify and define some potential users’ profiles.

The personas profiles were representative and paradigmatic, characterised by caricatural elements that contributed to developing empathy.
Towards a Digital Humanities Lab at the BnF
Designing a new kind of reading room
POC in partnership with TERALAB: choice of a use case to test the implementation of a secure infrastructure and make DSI teams increase their skills
THANK YOU!

EMMANUELLE BERMÈS
@bnf.fr