TEXT, DATA AND LINK-MINING IN DIGITAL LIBRARIES: LOOKING FOR THE HERITAGE GOLD

Emmanuelle Bermès - BnF
IFLA Satellite Meeting 2017: Digital Humanities
THE CORPUS 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

See: http://c.bnf.fr/fom
Digital Collections at the BnF

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

Web archives
793 To

Metadata
More than 20 million records
http://data.bnf.fr
EPISTEMOLOGICAL QUESTIONS: THE EXAMPLE OF THE COMMONPLACES PROJECT

Clovis Gladstone, Glenn Roe, Robert Morrissey, and Mark Olsen, “Digging into ECCO: Identifying Commonplaces and other Forms of Text Reuse at Scale”, Digital Humanities 2016, Krakow, Poland, July 2016

And like the baseless fabric of this vision, The cloud-capped towers, the gorgeous palaces, The solemn temples, the great globe itself — Yea, all which it inherit—shall dissolve. And like this insubstantial pageant faded, (...) (Shakespeare, The Tempest, Act 4, Scene 1) ca. 1611

Of this observation Shakespeare gives a beautiful example, in the passage last quoted: The cloud-capt the gorgeous palaces, The solemn temples, the great globe itself, Yea all which it inherit, (hail diffil... And like the fadeLs fabric of a vision, lc.ive not a rack behind.

James Elphinstone (1771)

as in this well-known passage, where you may also mark the fine climax. The cloud-capt Towers, The gorgeous Palaces, the great Globe itself, Yea, all which it inherits, shall disolve, And, like the baseless Fabric of a Vision, Leave not a Wreck behind.

Pratt, Mr. (1776)

Non diffinimic Lda quet’idea è quella di Shakespeare, in quei bei verli, Tlce cloud-capp’d Towers, the gorgeous Palaces, The solemn Temples, the great Globe itsef, YEa, all which it inheriftall diffilve, Aid. liie e bte lj.les Fabric d’a viLoin. L’ave no1 a ratc behind!

Giovanni Rucellai (1779)
BEFORE CORPUS... (THE CORPUS AS A SOURCE)

- The Europeana Newspapers project

BEFORE CORPUS... (THE CORPUS AS A SANDBOX)

- ETIS : the ASAP project


Corpus from http://images.bnf.fr
BEFORE CORPUS… (THE CORPUS AS AN INTERFACE)

- Project « le devenir en ligne du patrimoine numérisé : l’exemple de la Grande Guerre » - Labex « les passés dans le présent »

Valérie Baudouin, Zeynnep Pelhivan : Cartographie de la Grande Guerre sur le Web : Rapport final de la phase 2 du projet "Le devenir en ligne du patrimoine numérisé : l'exemple de la Grande Guerre"

https://hal.archives-ouvertes.fr/hal-01425600
THE CORPUS PROJECT, YEAR 1

- Study of web archives, in partnership with Web90 (CNRS/ISCC)

Video recordings: [http://c.bnf.fr/fse](http://c.bnf.fr/fse)
THE CORPUS PROJECT, YEAR 1: RELATED PROJECTS

- Bibli-Lab : analysis of logs from Gallica

Etude réalisée par Adrien Nouvellet, Télécom Paristech
A WIDE ARRAY OF DIFFERENT SITUATIONS...

- Good knowledge of collections, lack of technical skills
- Expert on tools, wants raw data

Access interfaces, pre-processed data
Training and education
Raw Data

library

researchers
BUILDING THE FUTURE

Online dissemination

Legal aspects

Organisational aspects

Skills and know-how

Human resources

IT infrastructure

Physical space

Digital scholarship Lab
THANK YOU!

emmanuelle.bermes@bnf.fr
@figoblog