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

Emmanuelle Bermès - BnF

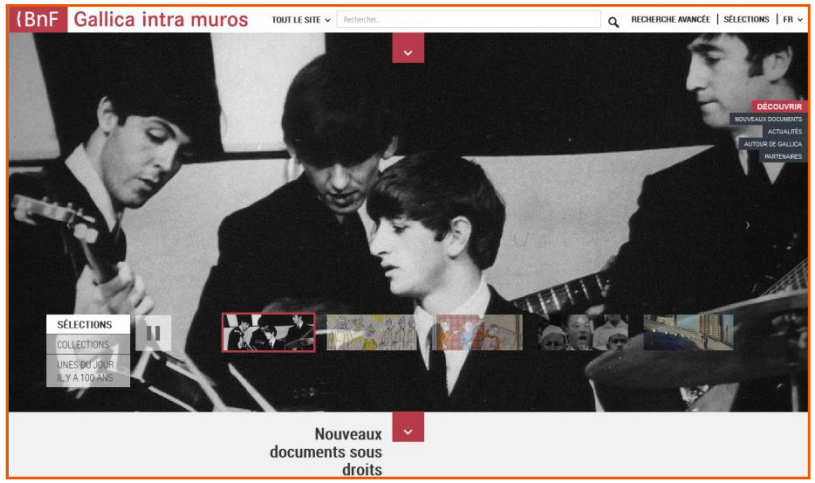
IFLA Satellite Meeting 2017: Digital Humanities

The CORPUS Project

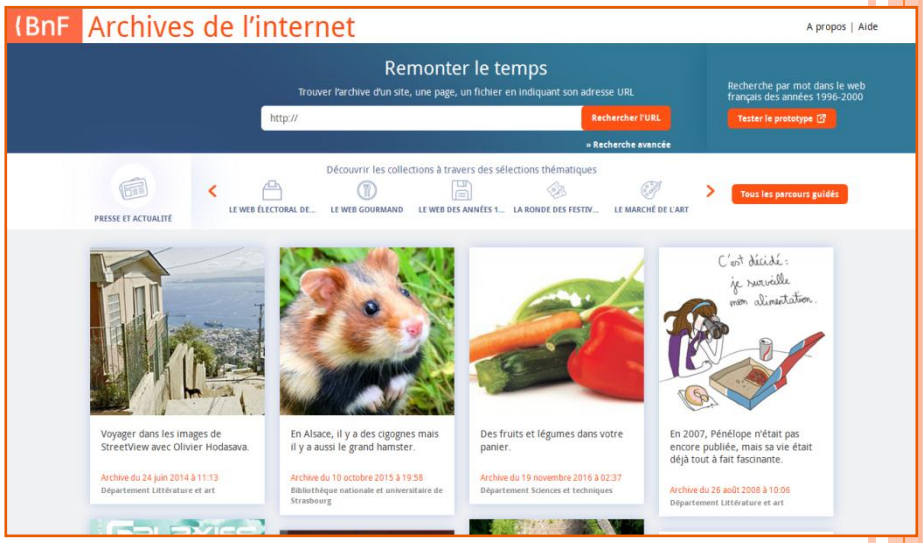
- Part of the BnF's 4-year internal research programme 2016-2019
- Objectives :
  - designing a future service for providing access to digital corpora for researchers
  - Providing researchers with data and tools to analyse them, in agreement with IPR and privacy legislation
- 3 years of experimentation (web archives, digitization, metadata) + 1 year for conclusions

➔ See: http://c.bnf.fr/fom

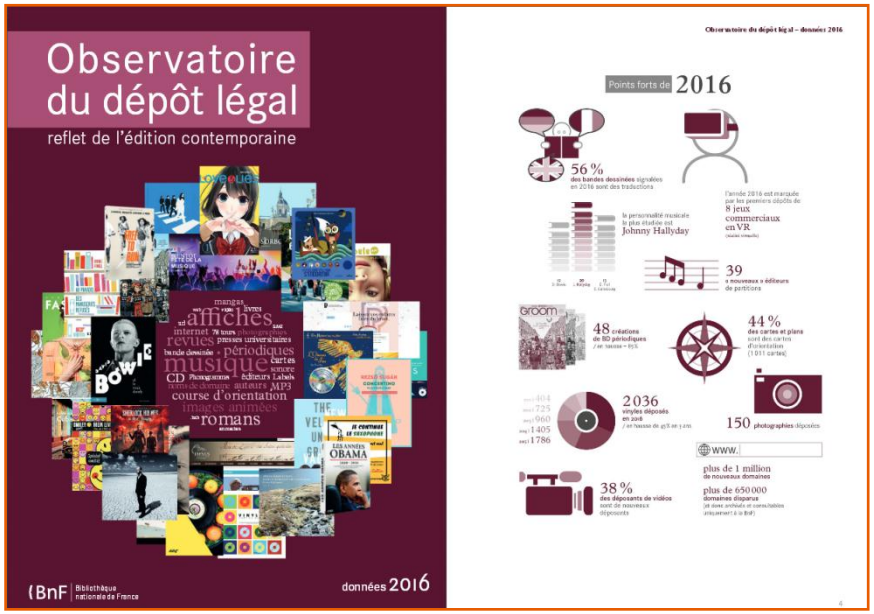
Digital Collections at the BnF



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

1.	**Shakespeare, William**	16.	Gildon, Charles
2.	**Horace**	17.	Young, Edward
3.	Pope, Alexander	18.	Congreve, William
4.	**Milton, John**	19.	Rider, William
5.	**Virgil**	20.	Cibber, Colley
6.	Ayscough, Samuel	21.	Griffith, Mrs. (Elizabeth)
7.	Bysshe, Edward	22.	**Fénelon, François de Salignac de...**
8.	**Ovid**	23.	Goldsmith, Oliver
9.	**Terence**	24.	Fenning, Daniel
10.	**Dryden, John**	25.	Addison, Joseph
11.	Becket, Andrew	26.	Walker, John
12.	Thomson, James	27.	**Voltaire**
13.	**Cicero, Marcus Tullius**	28.	Garrick, David
14.	**Jonson, Ben**	29.	Cibber, Theophilus
15.	Chambers, Ephraim	30.	Enfield, William

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 pa faded, (...)
(**Shakespeare**, *The Tempest*, Act 4, Scene 1) ca. 1611

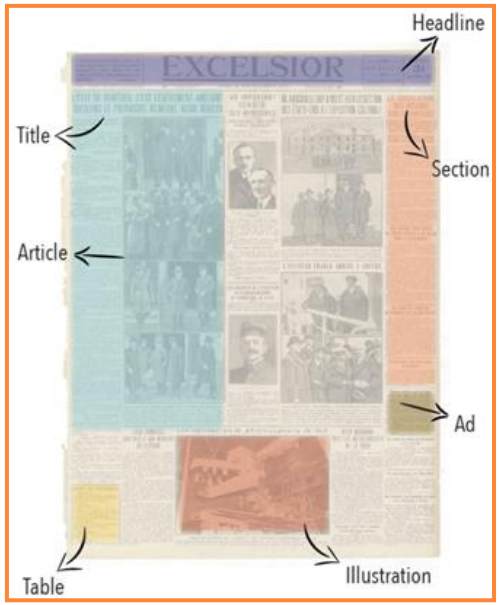
Of this obfervationl Shakefpear gives a beautiful example, in the passage lass quoted: The cloud-capt the gorgeous pa- laces, The solemn temples, the great globe it- self, Yea all which it inherit, (hall dilT
And like the bafeclfs fabric of a vision, lc.ive not a rack behind.
James Elphinston (1771)

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 diso/lve, And, like the baseless Fabrick of a Vision, Leave not a Wreck behind.
Pratt, Mr. (1776)

Non diffimilc Lda quet'idea è quella di Shakefpeare, in quei bei verli, Tlte cloid-capp'd Towers, the gorgeous Palaces, The folemn Temples, the great Globe itfel, YEa, all which it inherifitall diffilve, Aid. liie e btle Ij.les Fabric d'a vfioin, L'ave nol a ratc behind!
Giovanni Rucellai (1779)

Before CORPUS… (the corpus as a source)

- The Europeana Newspapers project



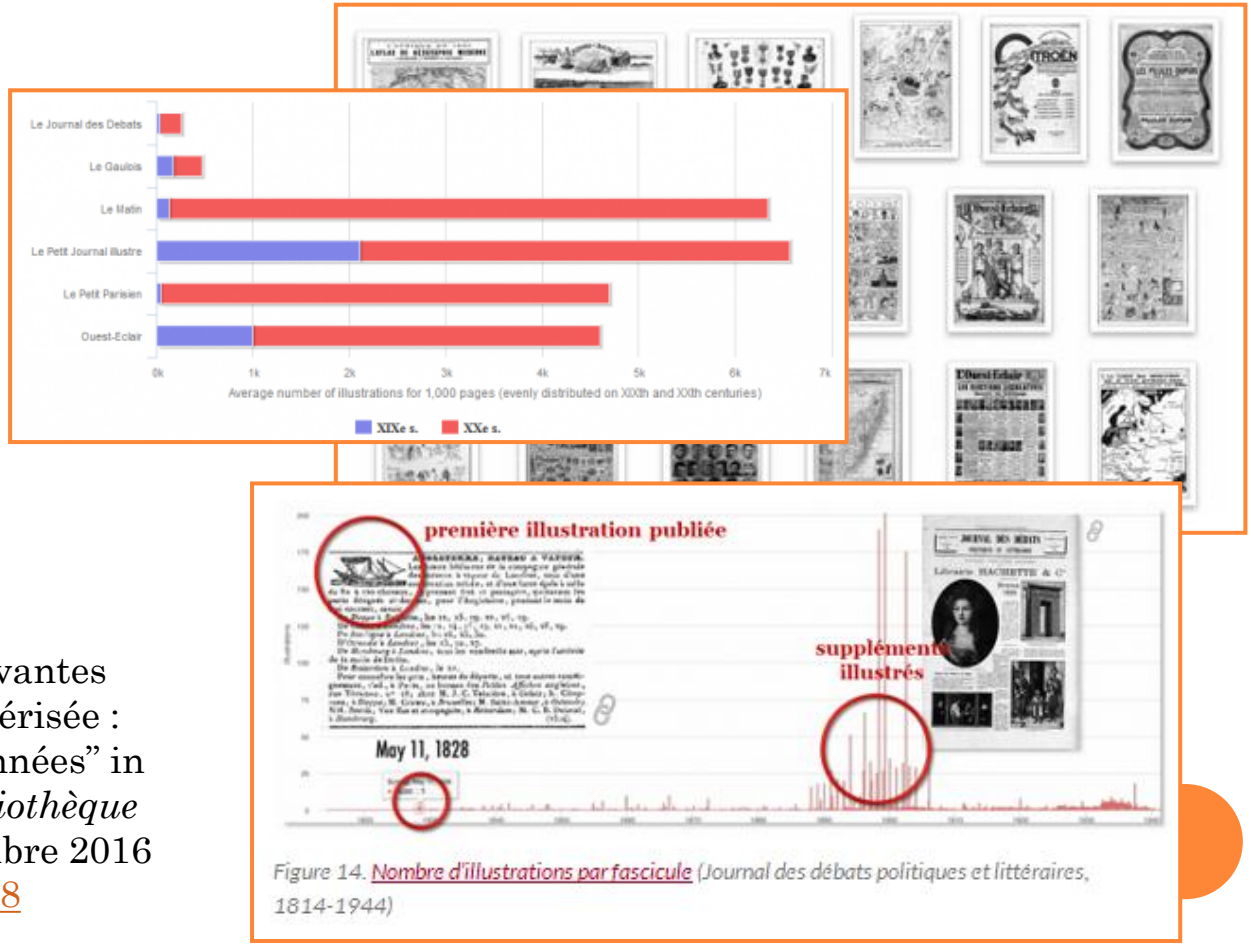


Figure 14. Nombre d'illustrations par fascicule (Journal des débats politiques et littéraires, 1814-1944)

JP Moreux, "Approches innovantes pour la presse ancienne numérisée : fouille et visualisation de données" in *Carnet de recherche à la bibliothèque nationale de France*, 3 décembre 2016 https://bnf.hypotheses.org/208

Before CORPUS… (the corpus as a sandbox)

- ETIS : the ASAP project

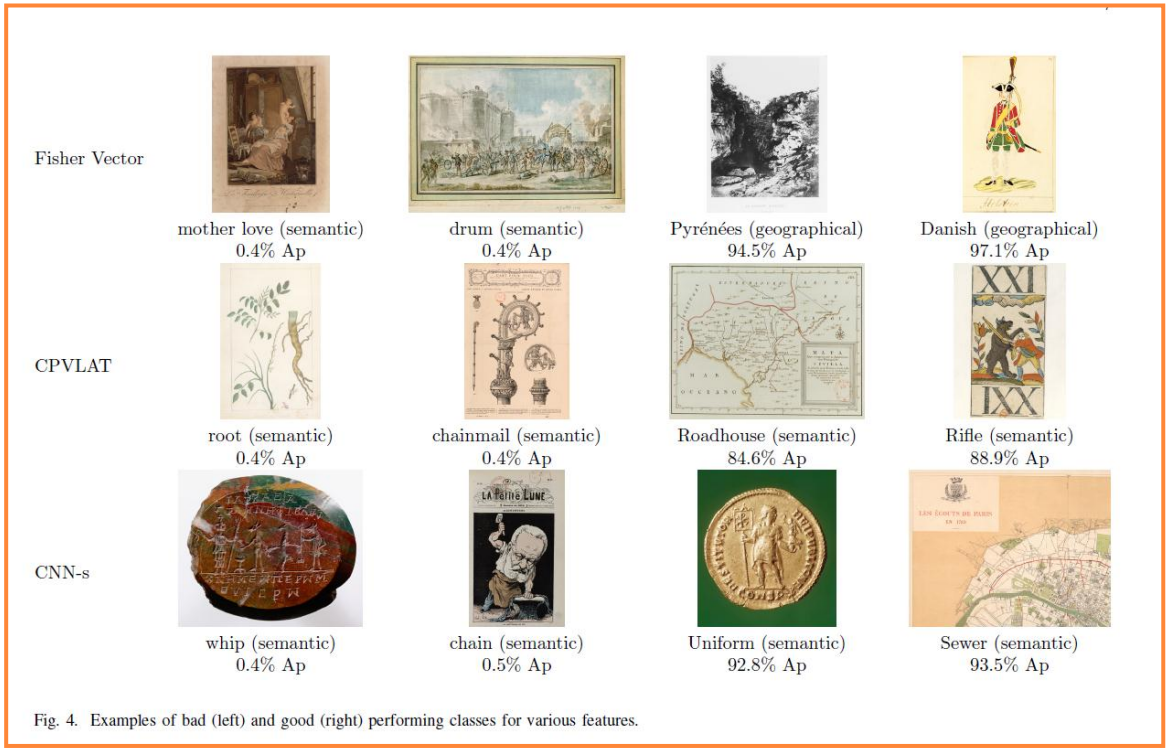


Fig. 4. Examples of bad (left) and good (right) performing classes for various features.

David Picard, Philippe-Henri Gosselin, Marie-Claude Gaspard. “Challenges in Content-Based Image Indexing of Cultural Heritage Collections.” *IEEE Signal Processing Magazine*, Institute of Electrical and Electronics Engineers, 2015, 32 (4), pp.95 – 102

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 »

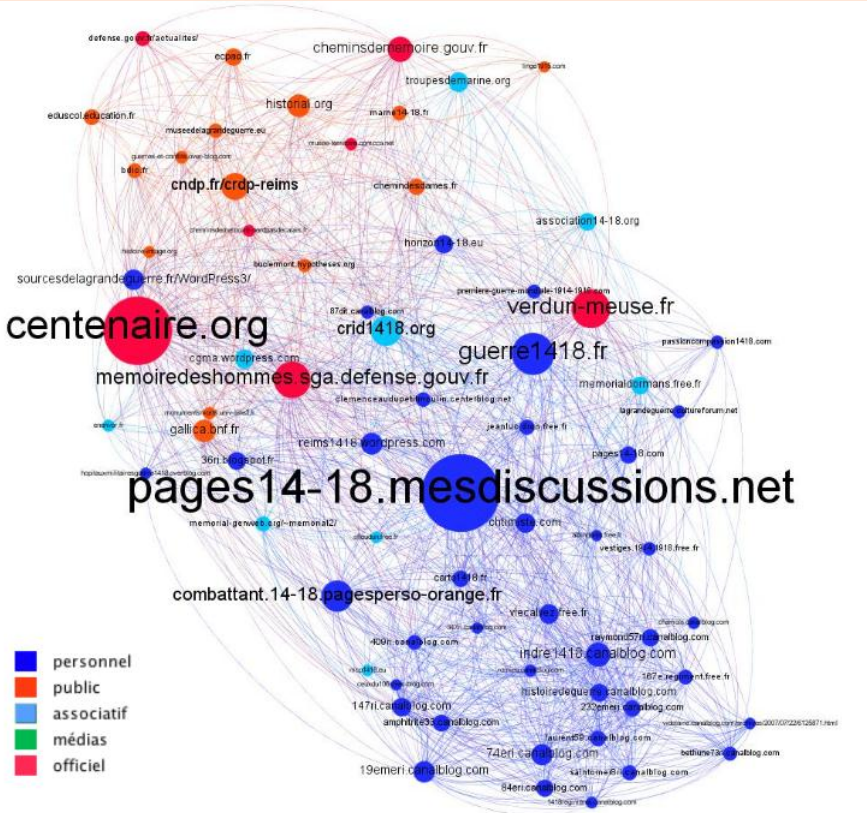
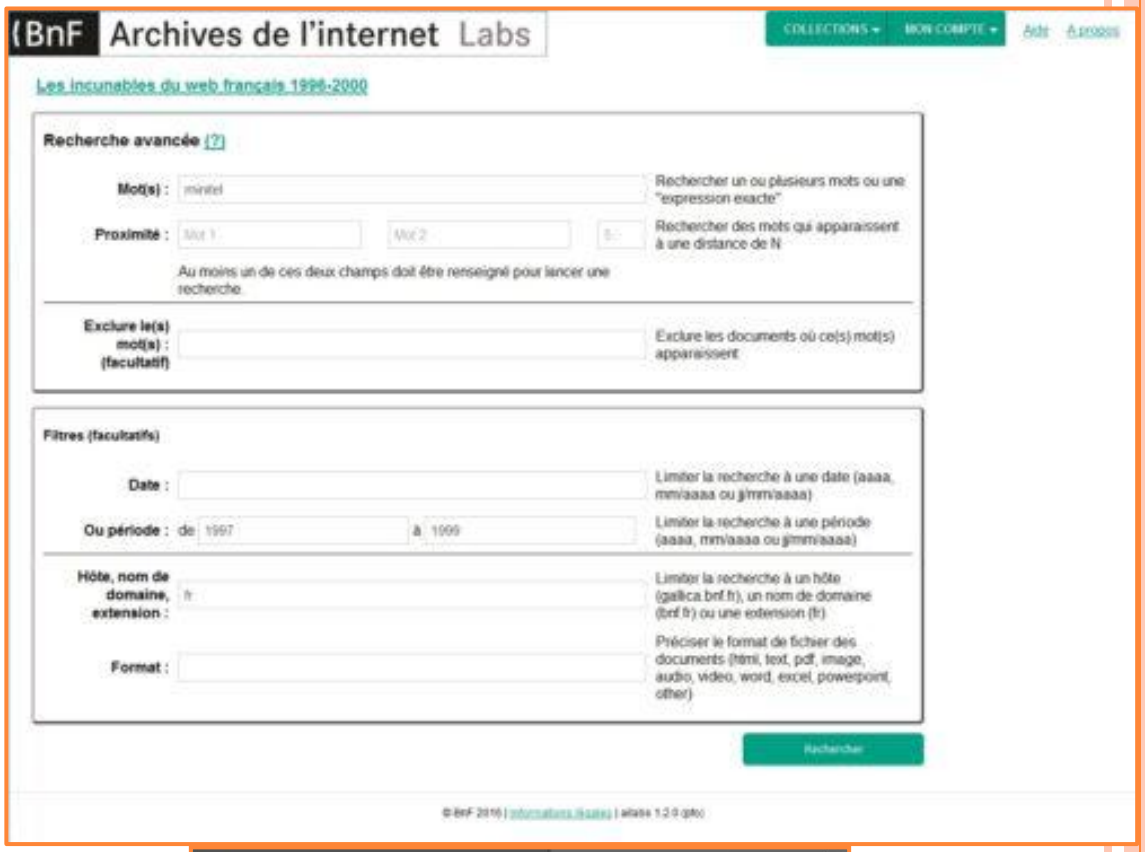


Figure 19. Carte du Web Grande Guerre (collecte de novembre 2014, sites avec degrés >30)

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

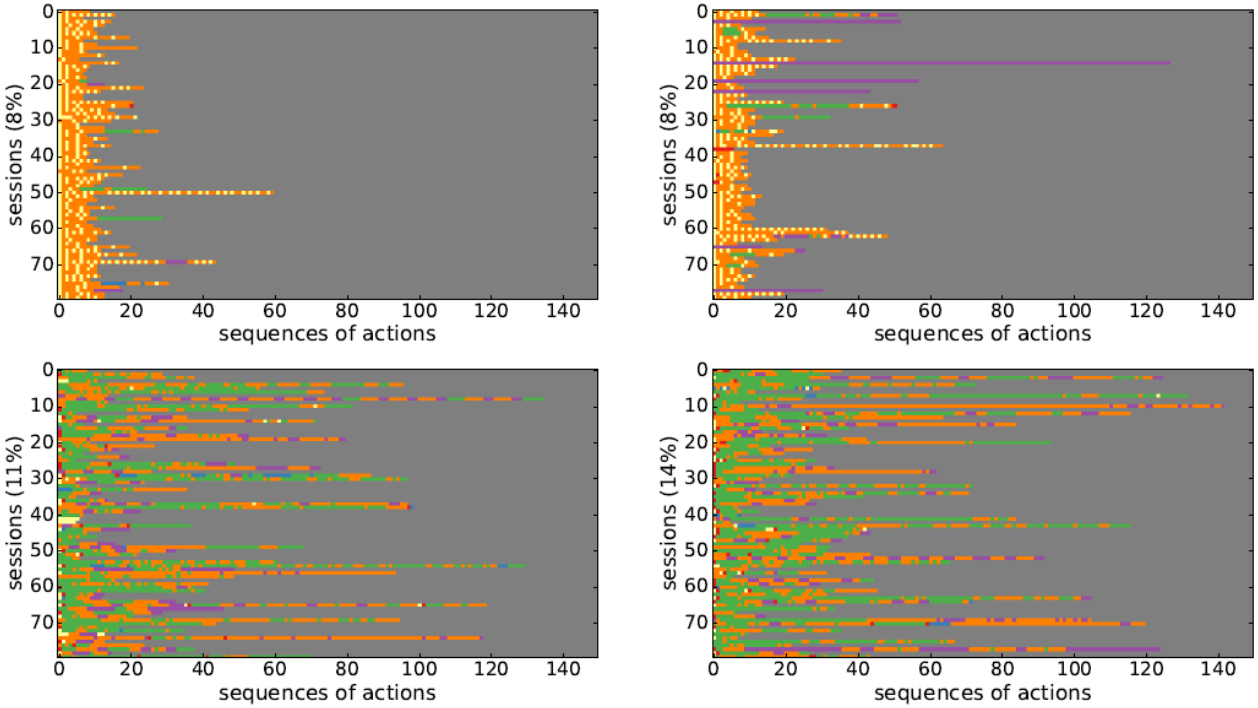
- Study of web archives, in partnership with Web90 (CNRS/ISCC)



Video recordings: http://c.bnf.fr/fse
Weblog: http://webcorpora.hypotheses.org/

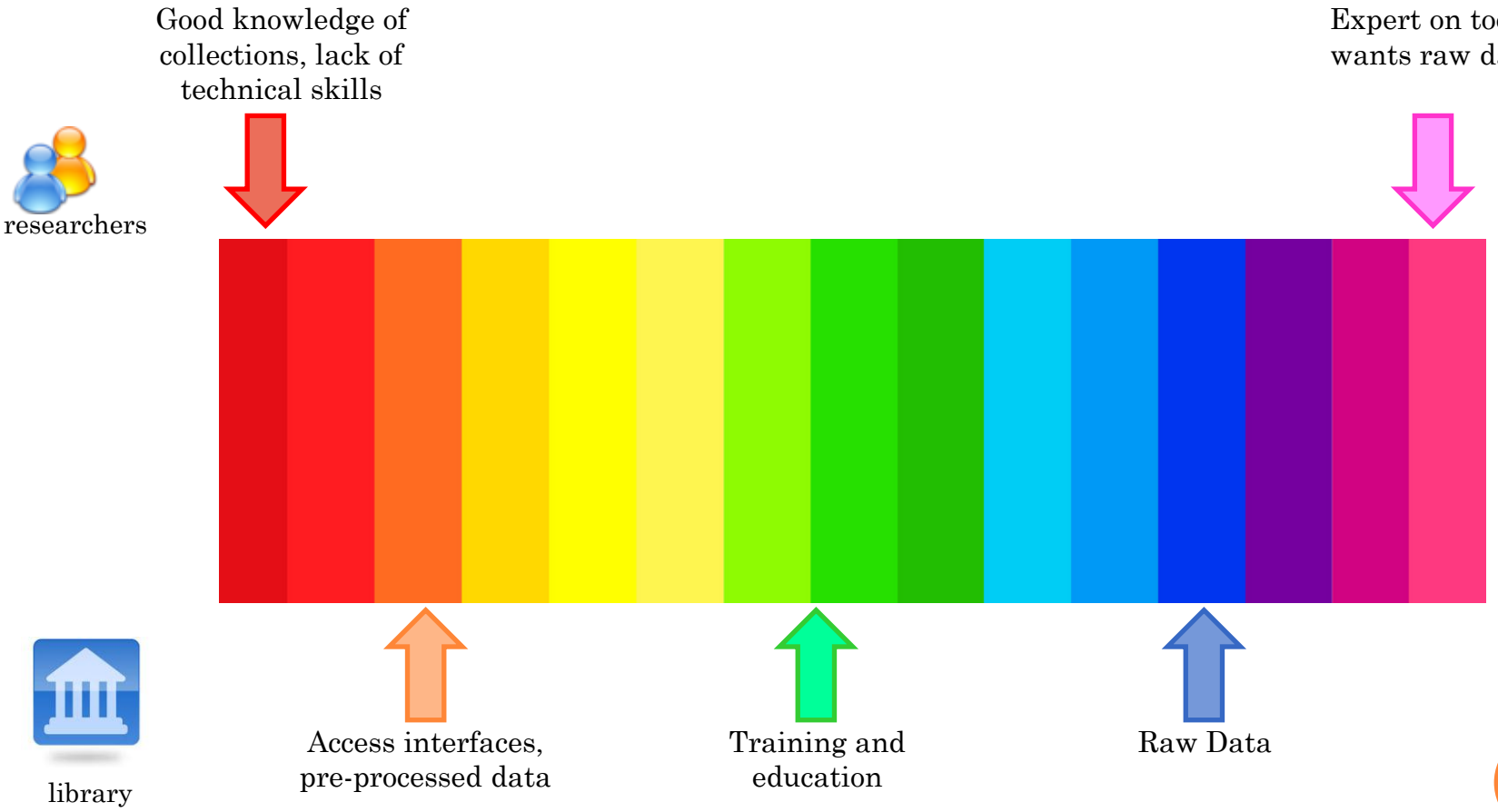
The Corpus project, year 1: related projects

- Bibli-Lab : analysis of logs from Gallica

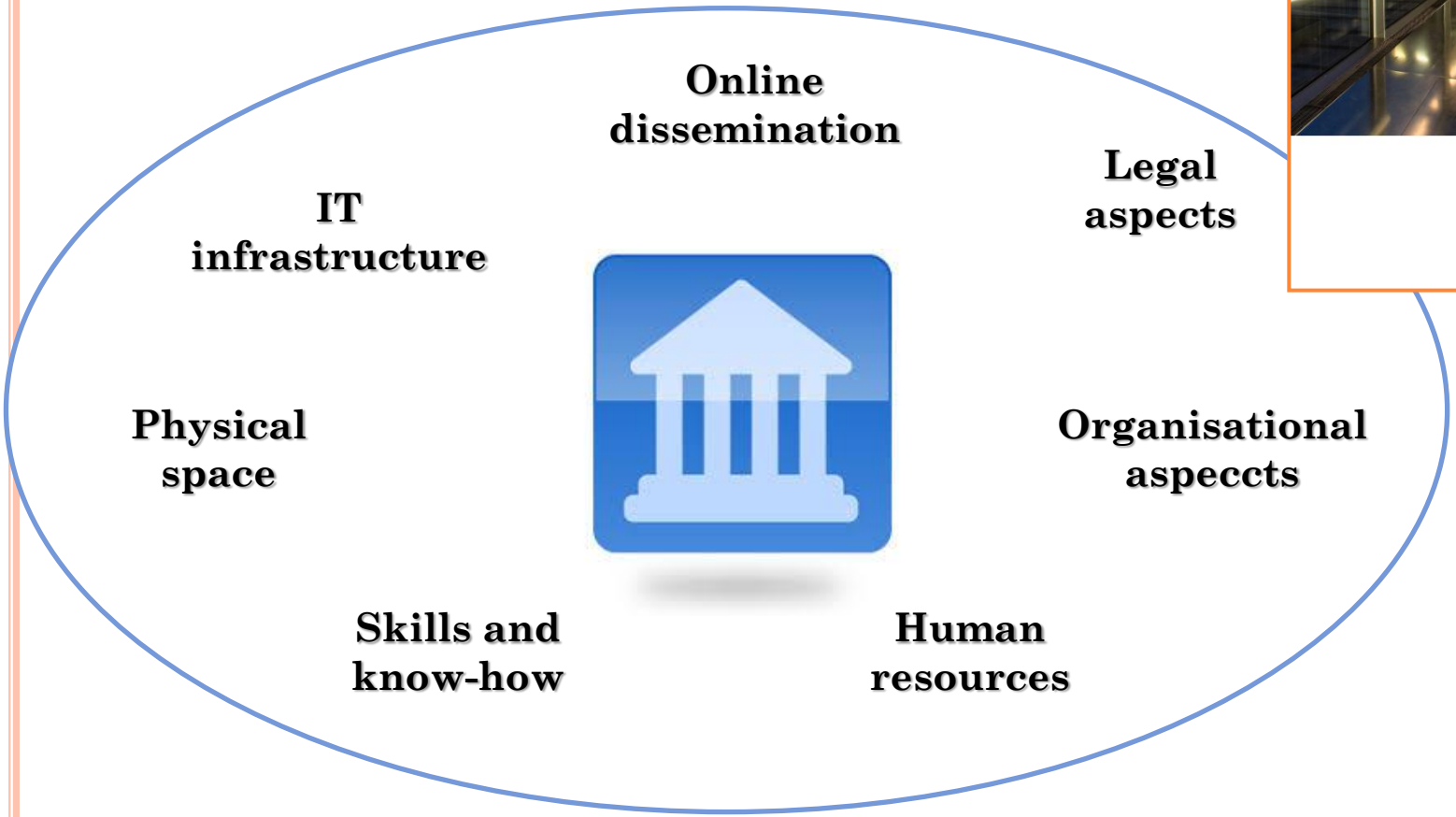


Etude réalisée par Adrien Nouvellet, Télécom Paristech

A WIDE ARRAY OF DIFFERENT SITUATIONS…



Building the future



Digital scholarship Lab

Thank you !

emmanuelle.bermes@bnf.fr

@figoblog