Fournir des corpus à destination de la recherche à des fins de text & data mining : le projet CORPUS

Emmanuelle Bermès

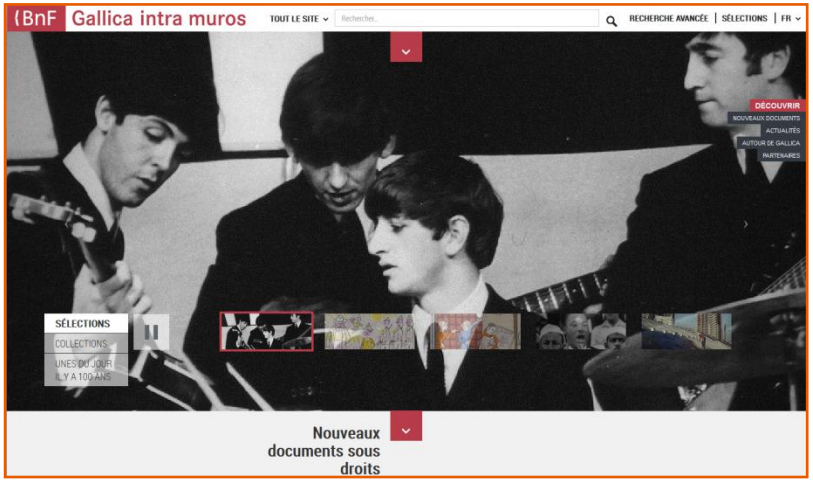
18/10/2017

Le projet DSR-CORPUS

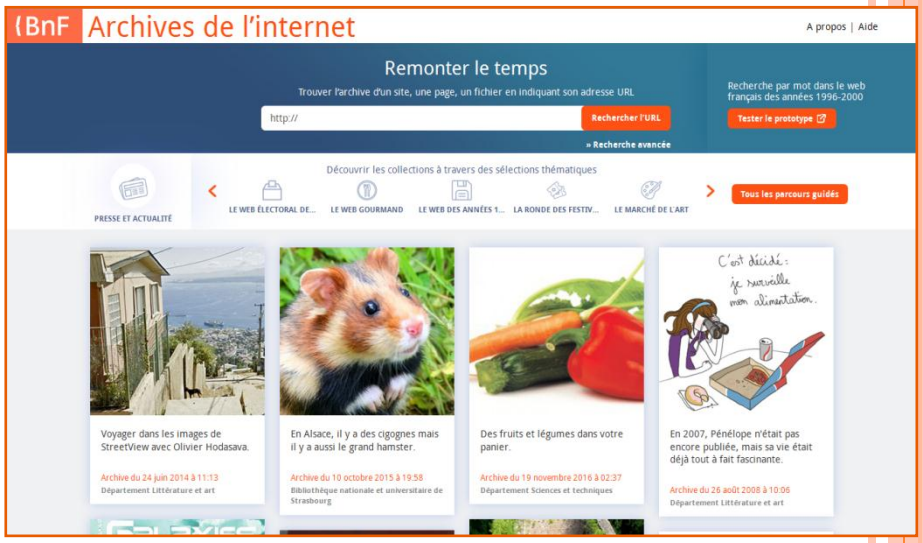
- Un projet inscrit au plan quadriennal de la recherche de la BnF 2016-2019
- Objectifs :
  - préfigurer un service de fourniture de corpus numériques à destination de la recherche
  - fournir à des chercheurs des données et des outils pour les analyser, dans le respect du droit d'auteur et de la vie privée.
- 3 années d'expérimentation (archives web, numérisation, métadonnées) + 1 année de bilan

➔ Fiche dans la base ANIR : http://c.bnf.fr/fom

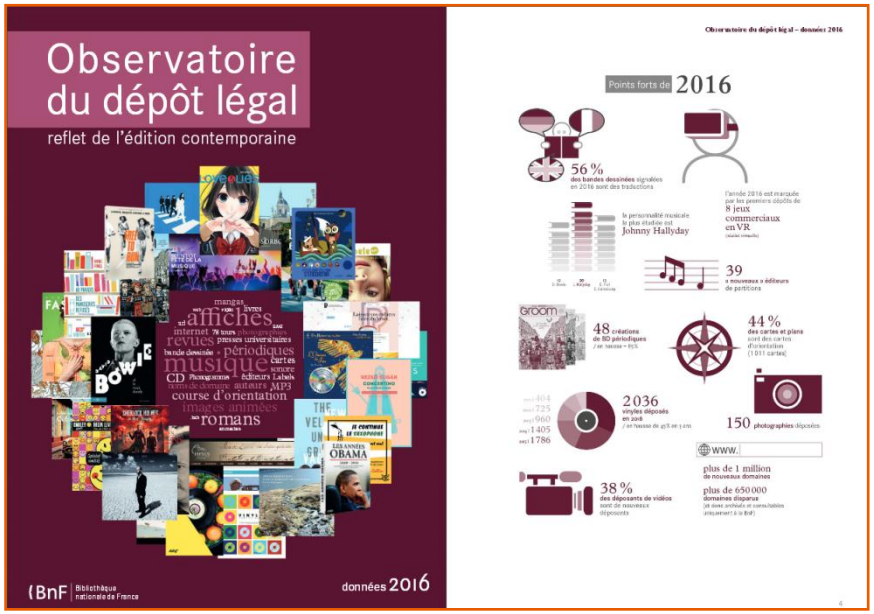
Quels corpus ?



Gallica + Gallica intra muros
4.5 millions de documents numériques
http://gallica.bnf.fr



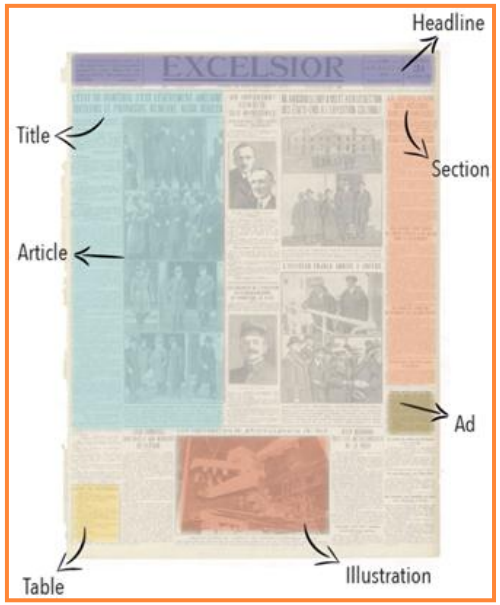
Archives de l'Internet
793 To

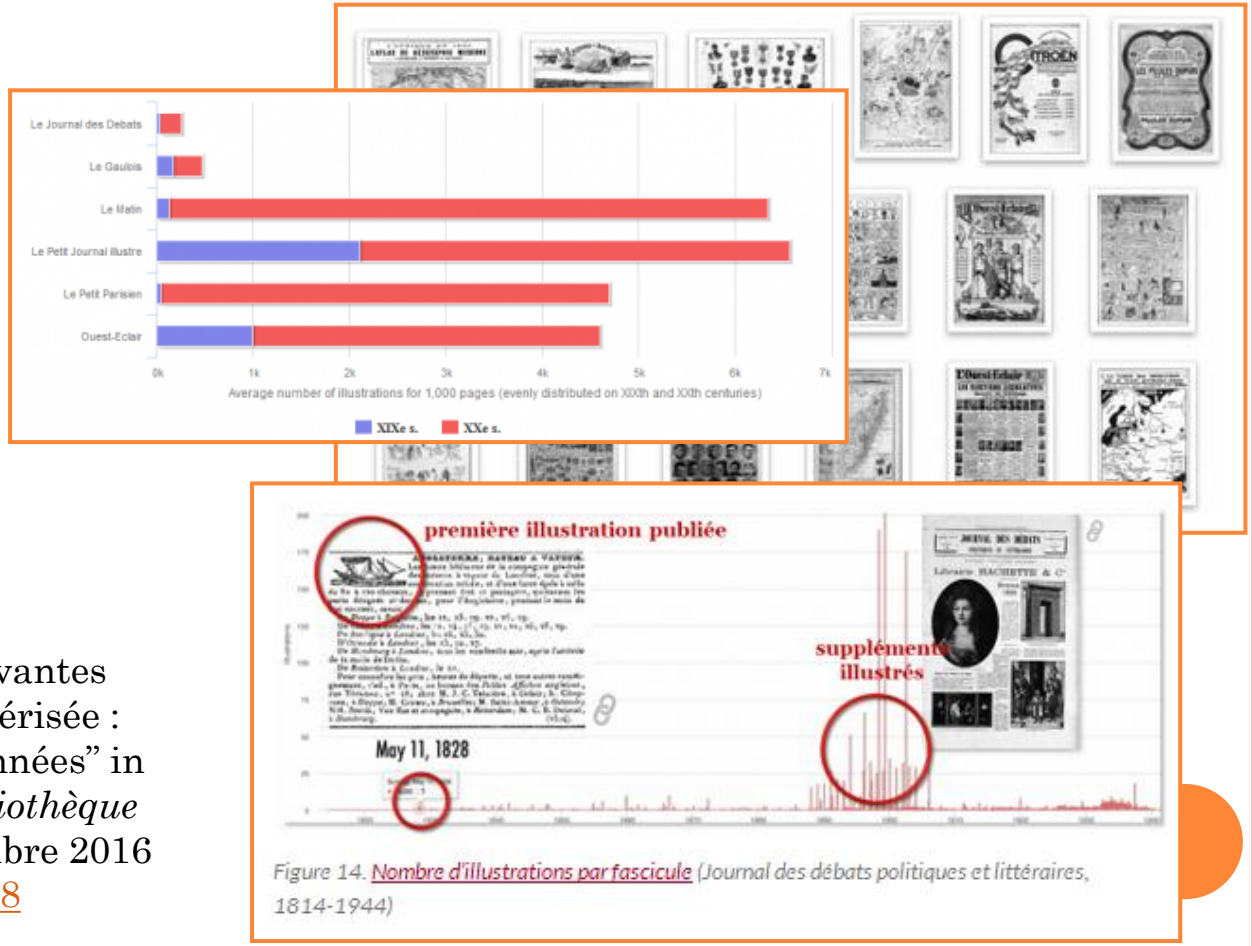


Métadonnées
+ de 20 millions de notices
http://data.bnf.fr

Avant CORPUS… (Le corpus comme source)

- The Europeana Newspapers project





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

Avant CORPUS… (Le corpus comme bac à sable)

- ETIS : le projet ASAP

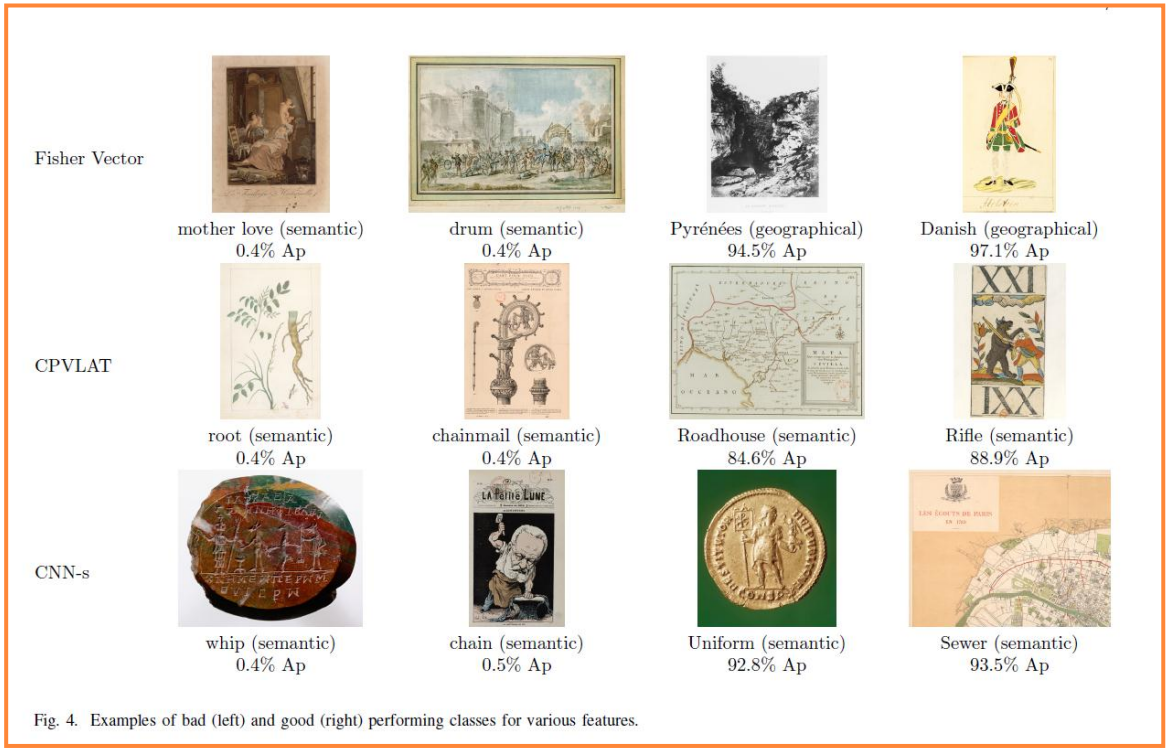


Fig. 4. Examples of bad (left) and good (right) performing classes for various features.

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

Avant CORPUS… (Le corpus comme interface)

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

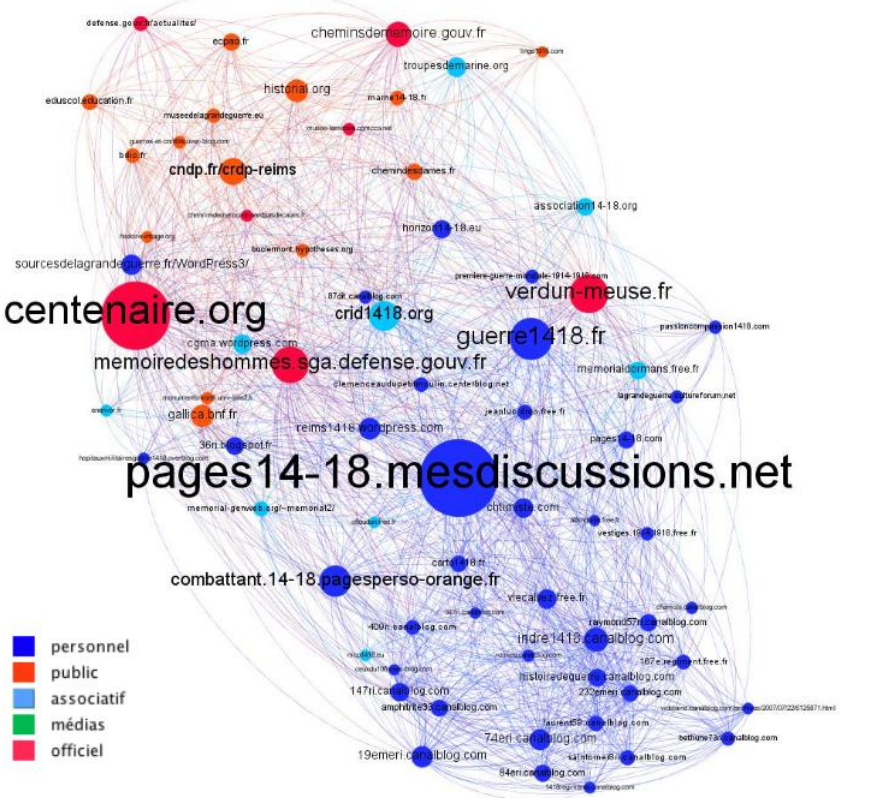
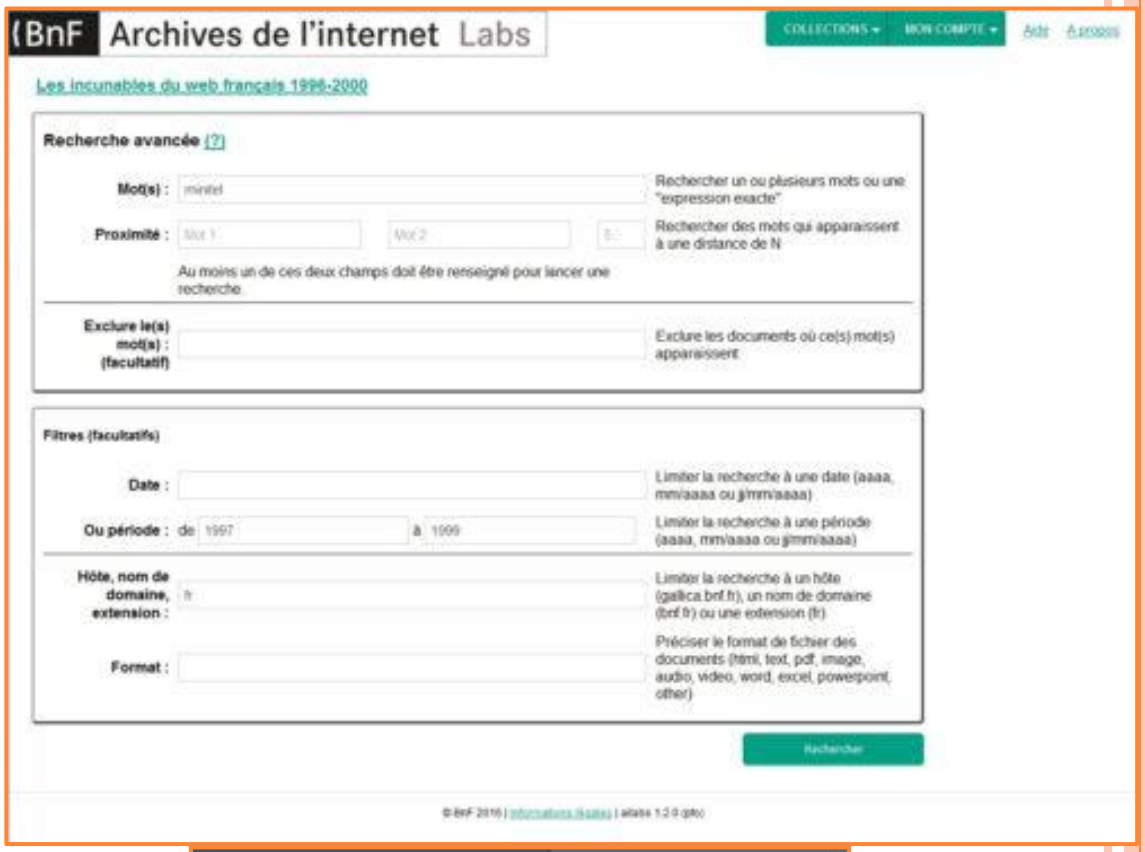


Figure 19. Carte du Web Grande Guerre (collecte de novembre 2014, sites avec degrés >30)

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

Bilan année 1

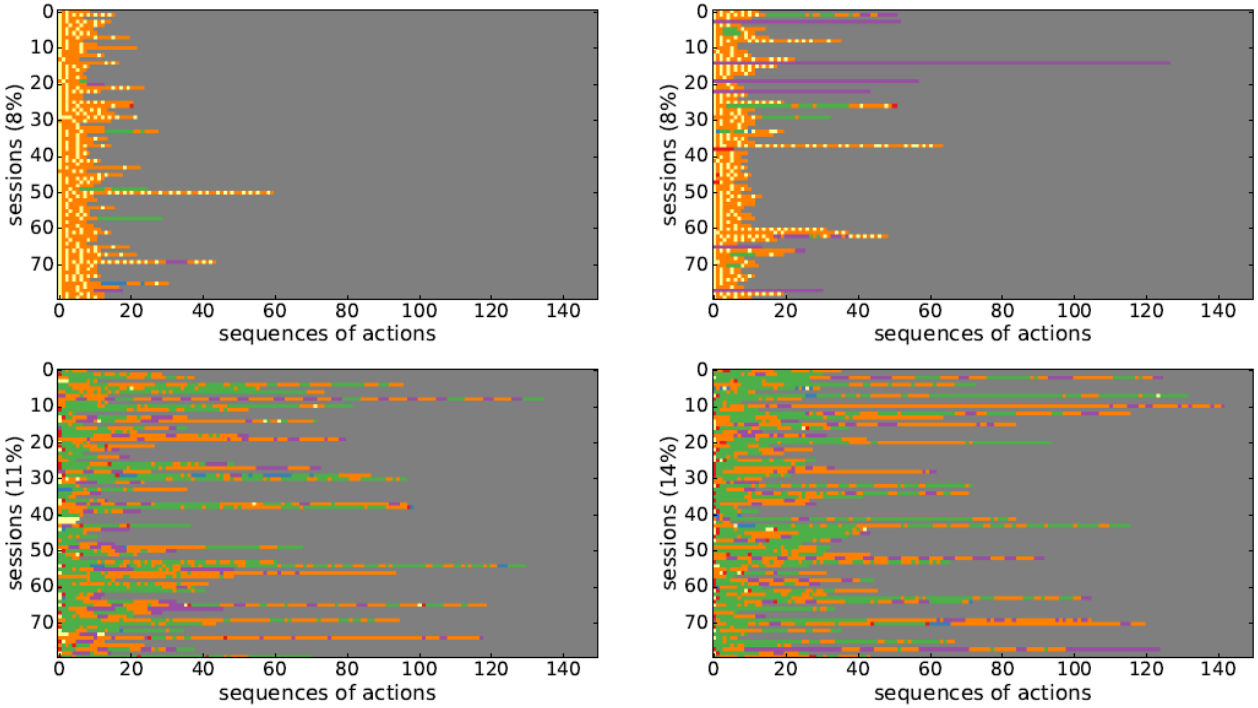
- Partenariat avec l'équipe Web90 (CNRS/ISCC)



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

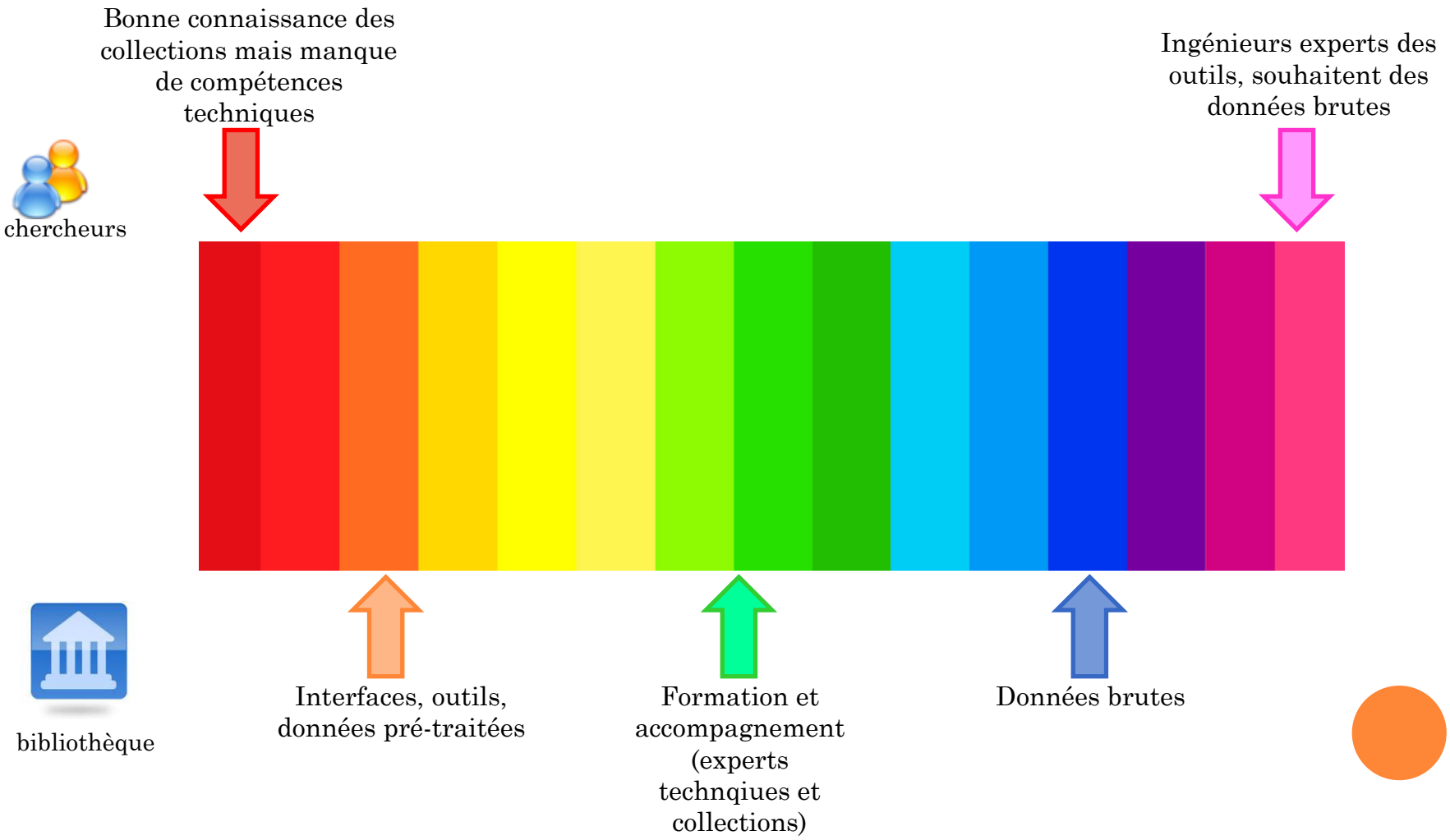
Bilan année 1 – Projets connexes

- Bibli-Lab : analyse des logs de Gallica

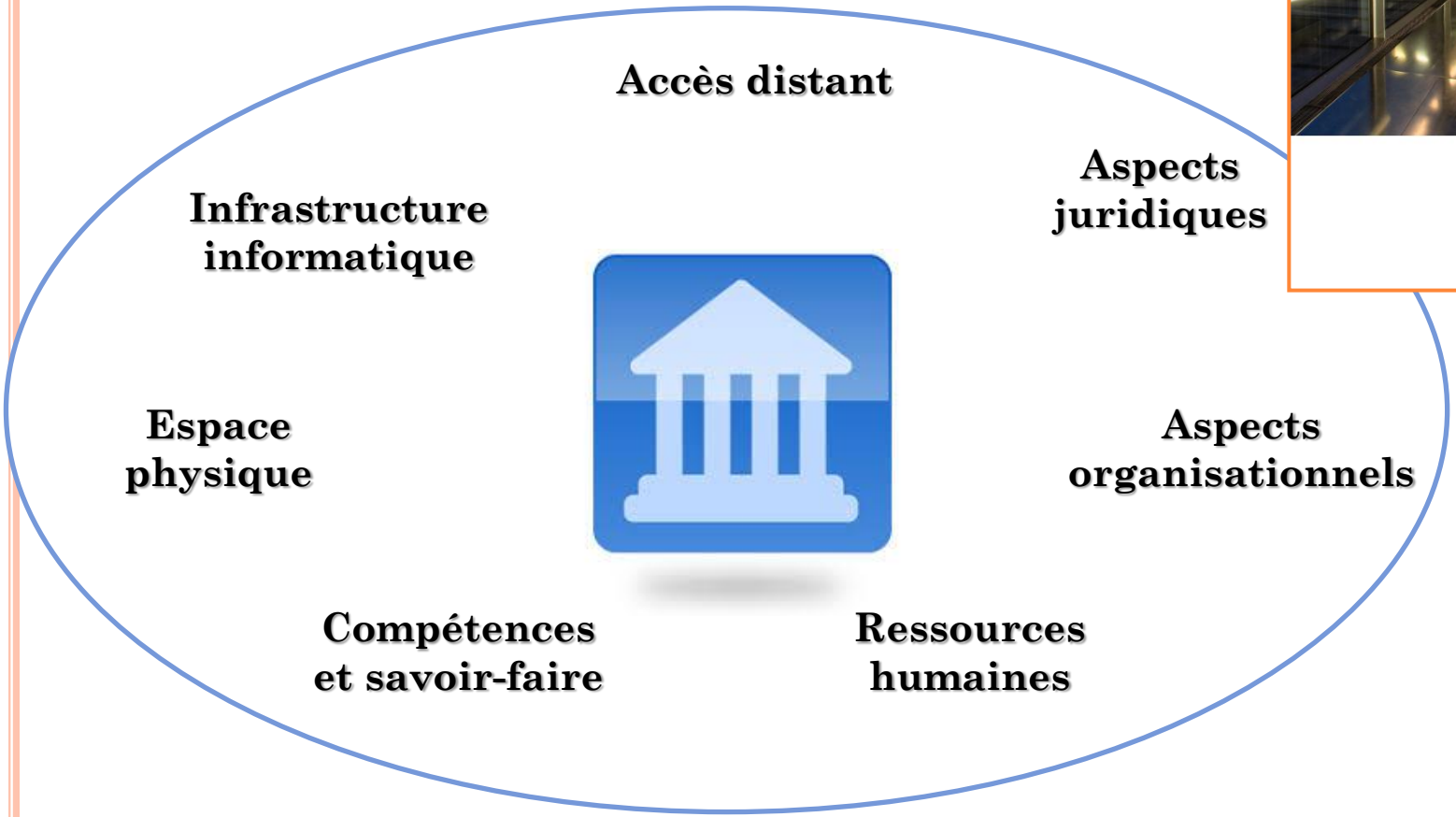


Etude réalisée par Adrien Nouvellet, Télécom Paristech

Une grande diversité de situations…



Construire l'avenir ?



Vers un laboratoire d'étude et d'analyse de corpus numériques à la BnF

Merci !

emmanuelle.bermes@bnf.fr

@figoblog