



FOURNIR DES CORPUS À DESTINATION DE LA RECHERCHE À DES FINS DE TEXT & DATA MINING : LE PROJET CORPUS

Emmanuelle Bermès

18/10/2017

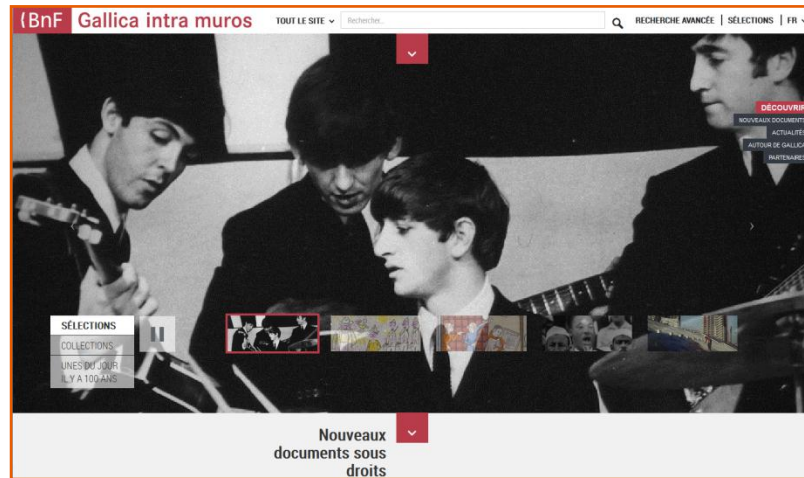
LE PROJET DSR-CORPUS

- 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

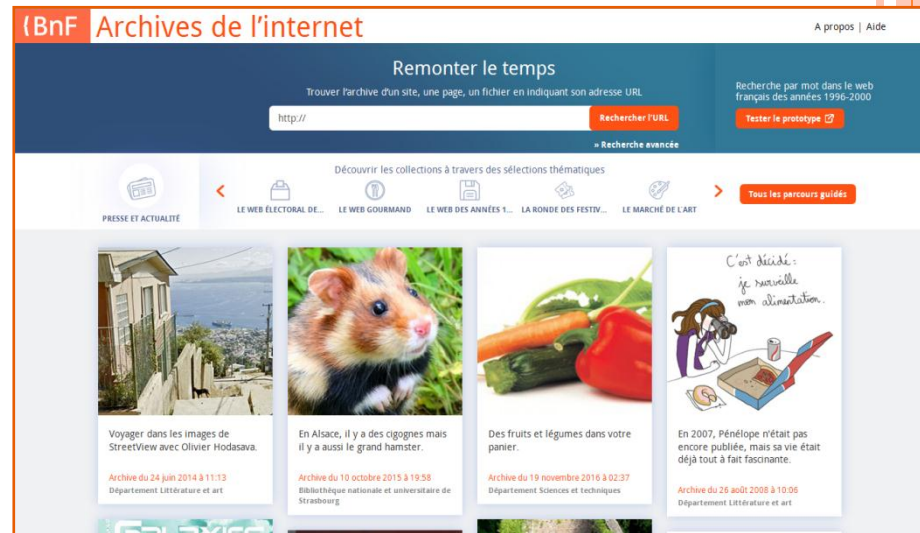
➔ Fiche dans la base ANIR : <http://c.bnf.fr/fom>



QUELS CORPUS ?

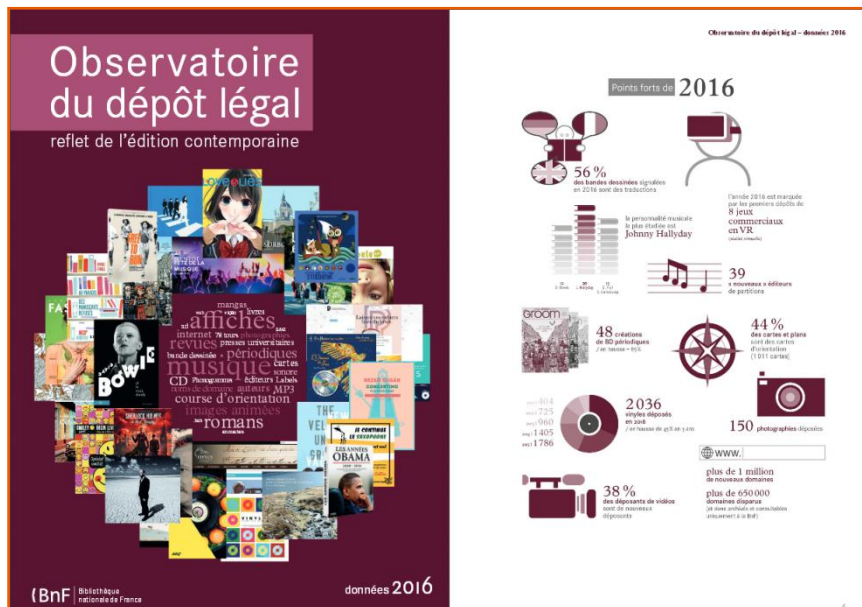


Gallica + Gallica intra muros
4.5 millions de documents numériques
<http://gallica.bnf.fr>



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

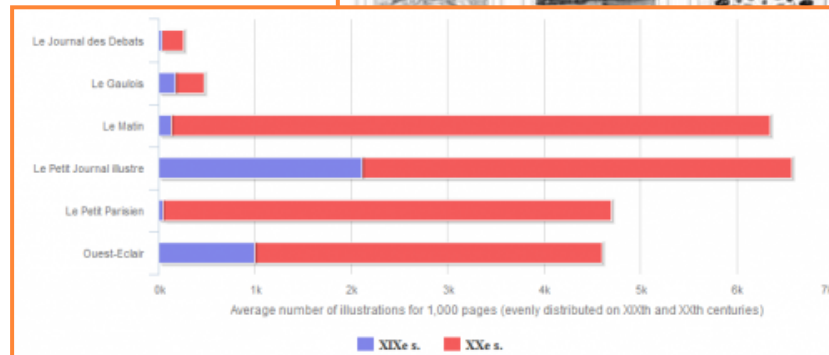
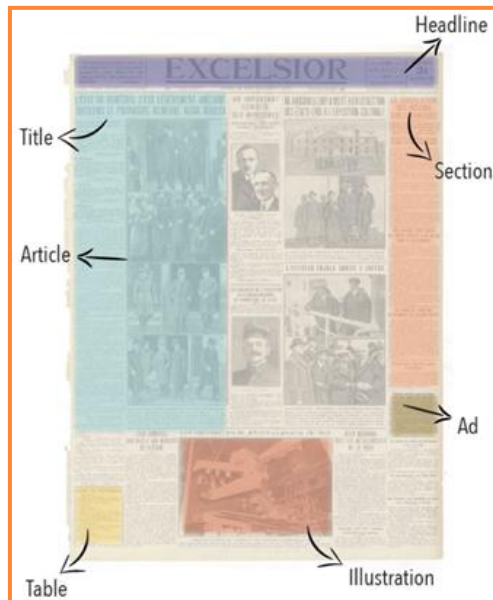


Figure 14. *Nombre d'illustrations par fascicule* (Journal des débats politiques et littéraires, 1814-1944)

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>

AVANT CORPUS... (LE CORPUS COMME BAC À SABLE)

- ETIS : le projet ASAP

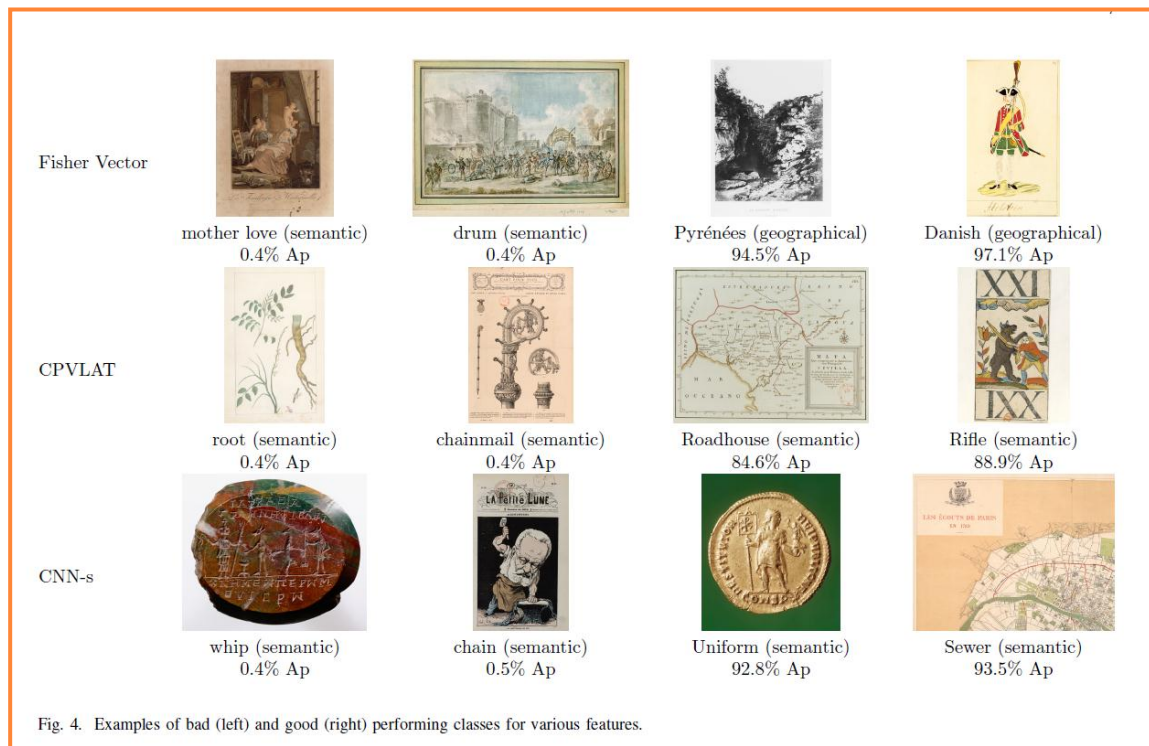


Fig. 4. Examples of bad (left) and good (right) performing classes for various features.

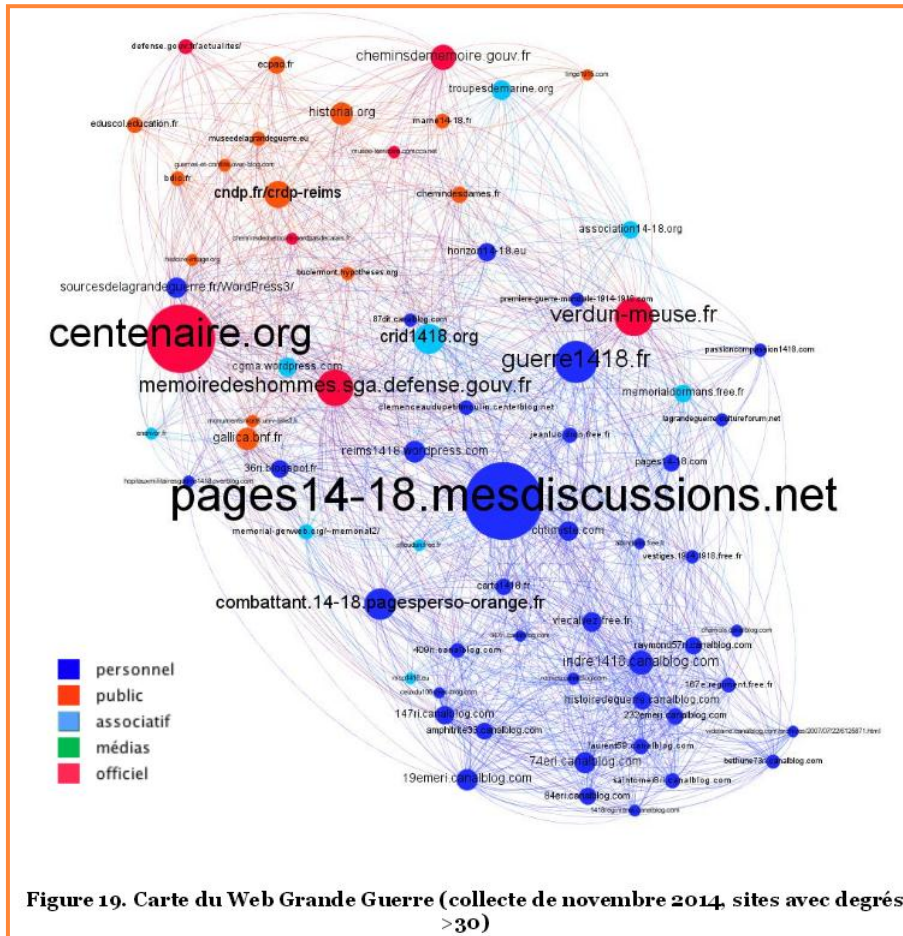
David Picard, Philippe-Henri Gosselin, Marie-Claude Gaspard. “Challenges in Content-Based Image Indexing of Cultural Heritage Collections.” *IEEE Signal Processing Magazine*, Institute of Electrical and Electronics Engineers, 2015, 32 (4), pp.95 – 102

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 »



Valérie Baudouin, Zeynep
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)



Les archives de l'internet comme sources : méthodes et représentations

Valérie BEAUDOUIN, enseignante-chercheuse (Télécom-ParisTech / Labex Les passés dans le présent), Sophie GEBEL, enseignante-chercheuse (Univ. Aix-Marseille), Francesca MUSIANI, enseignante-chercheuse (ISCC / Web90), Valérie SCHAFER, enseignante-chercheuse (ISCC / Web90), Marie-Luce VIAUD, cheffe de projet recherche et développement (Ina), Dana DIMINESCU, enseignante-chercheuse (Télécom-ParisTech)



Durée : 59 min

(BnF) Archives de l'internet Labs

COLLECTIONS MON COMPTE Aide À propos

Les incunables du web français 1990-2000

Recherche avancée (?)

Mot(s) : Rechercher un ou plusieurs mots ou une "expression exacte"

Proximité : Mot 1 : Mot 2 : Rechercher des mots qui apparaissent à une distance de N

Au moins un de ces deux champs doit être renseigné pour lancer une recherche

Exclure le(s) mot(s) : Exclure les documents où ce(s) mot(s) apparaissent

Filtres (facultatifs)

Date : Limiter la recherche à une date (aaaa, mm/aaaa ou jjmm/aaaa)

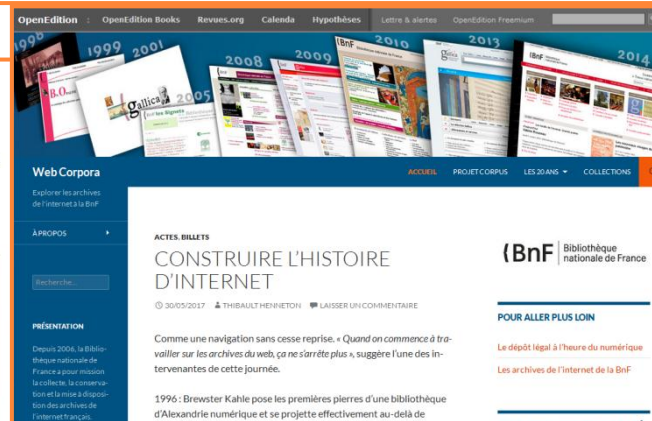
Ou période : de 1997 à 1999 Limiter la recherche à une période (aaaa, mm/aaaa ou jjmm/aaaa)

Hôte, nom de domaine, extension : Limiter la recherche à un hôte (galica.bnf.fr), un nom de domaine (bnf.fr) ou une extension (.fr)

Format : Préciser le format de fichier des documents (html, text, pdf, image, audio, video, word, excel, powerpoint, other)

Rechercher

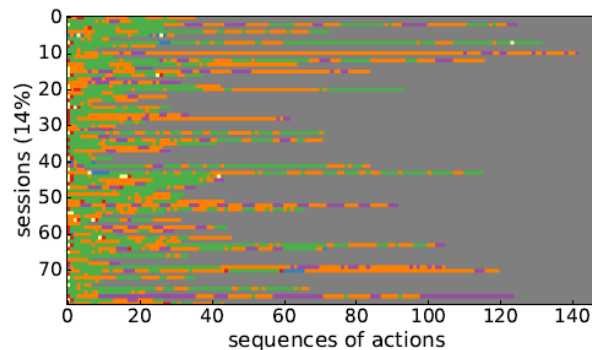
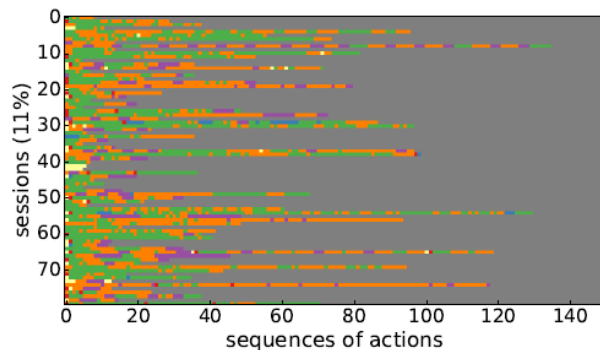
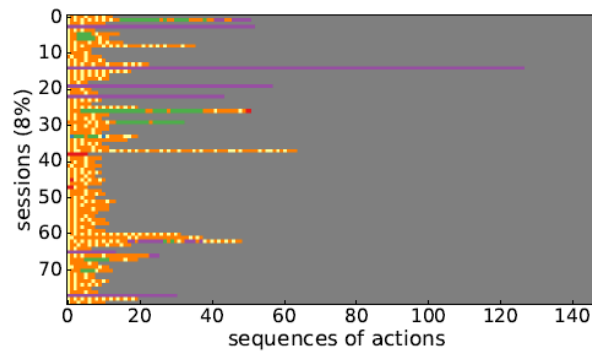
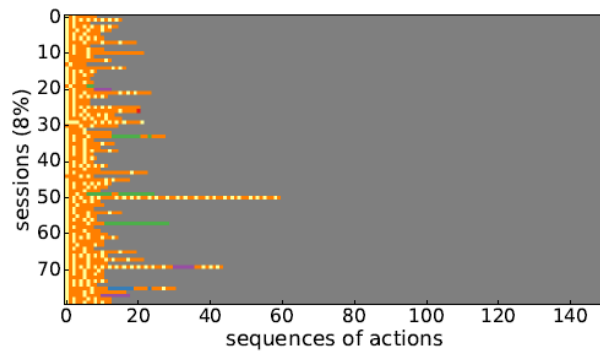
© BnF 2016 | informations@bnf.fr | version 1.2.0 (git)



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

BILAN ANNÉE 1 – PROJETS CONNEXES

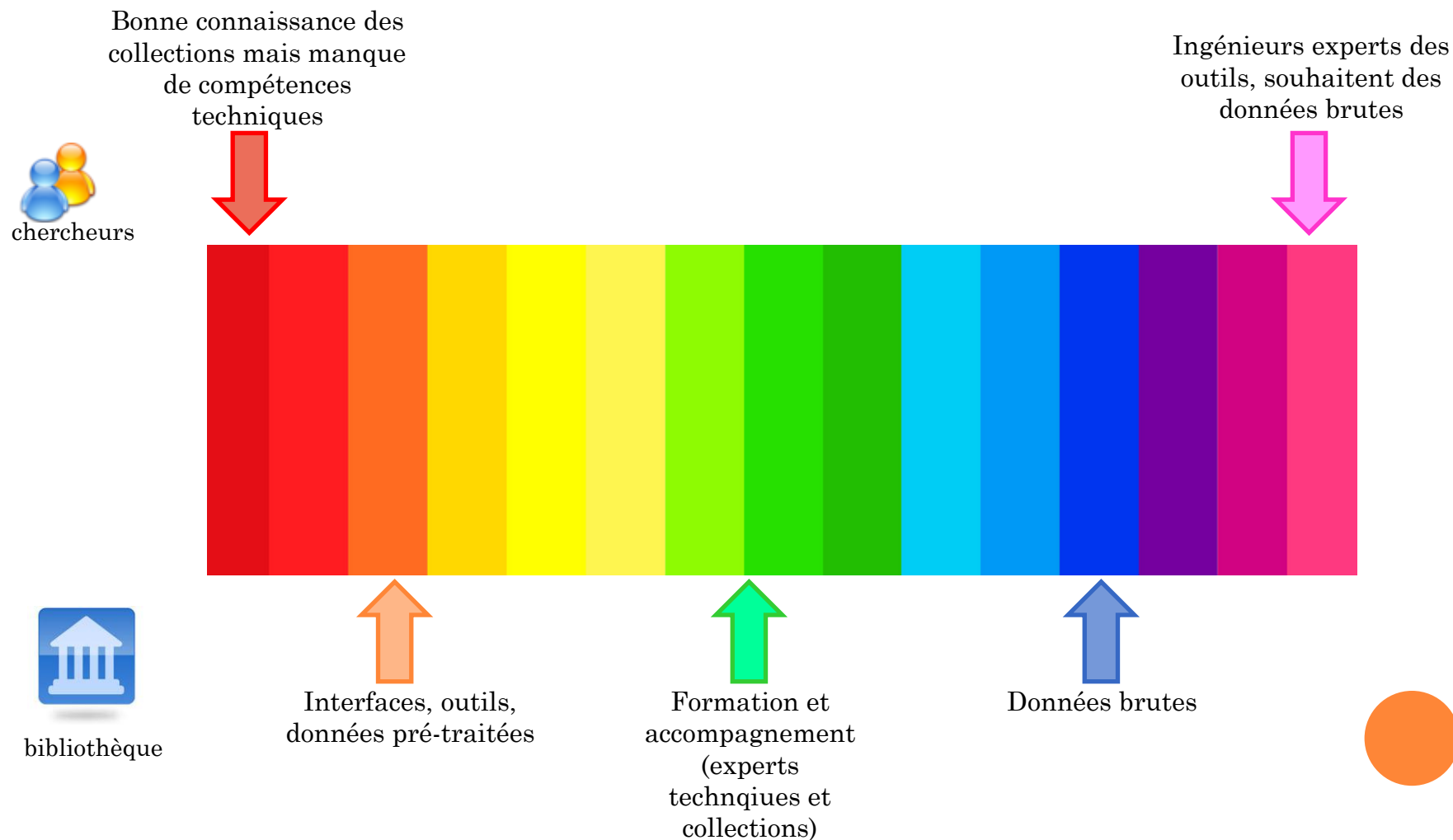
○ Bibli-Lab : analyse des logs de Gallica



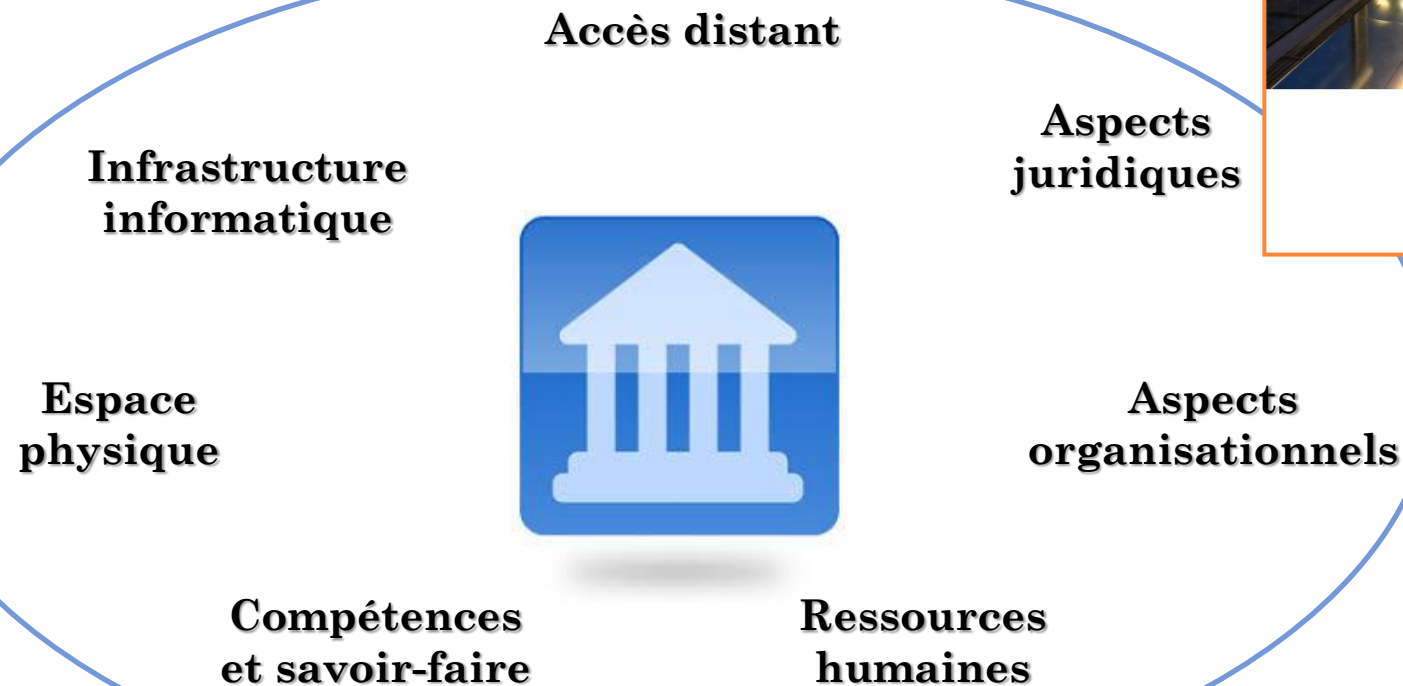
Etude réalisée par Adrien Nouvellet, Télécom Paristech



UNE GRANDE DIVERSITÉ DE SITUATIONS...



CONSTRUIRE L'AVENIR ?



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

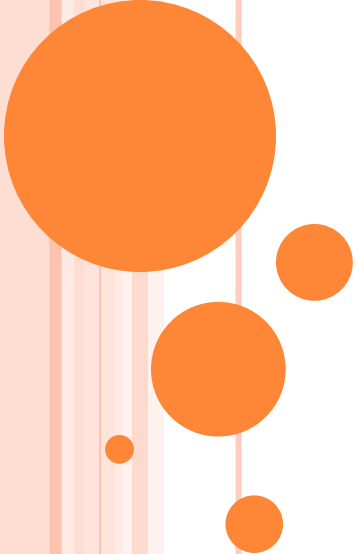
Offrir aux chercheurs, dans les
emprises de la Bibliothèque, des
outils de fouille et d'exploration de
textes et de données sur des corpus
numériques de la BnF

► Proposer des environnements
scientifiques et techniques
(plate-forme sécurisée, logiciels,
assistance d'experts...) pour explorer,
dans le respect des dispositions
réglementaires, les corpus numériques
de la BnF

CONTRAT D'OBJECTIFS
ET DE PERFORMANCE 2017-2021



(BnF) Bibliothèque
nationale de France



MERCI !

emmanuelle.bermes@bnf.fr

 **@figoblog**