Le DataSprint ResPaDon : une expérimentation interdisciplinaire autour de la constitution et de l'analyse de corpus issus des Archives de l'internet en lien avec le Web " vivant "

Audrey Baneyx, Dorothée Benhamou-Suesser, Eleonora Moiraghi

RES
PA
DON

Le DataSprint ResPaDon : une expérimentation interdisciplinaire autour de la constitution et de l'analyse de corpus issus des Archives de l'internet en lien avec le Web « vivant ».

Audrey Baneyx
médialab, Sciences Po

Dorothée Benhamou-Suesser
BnF

Eleonora Moiraghi
DRIS, Sciences Po

Colloque Humanistica 2022

18-21 mai 2022 Montréal (Canada)

Université de Lille · BnF Bibliothèque nationale de France · SciencesPo · Campus Condorcet Paris-Aubervilliers · GériCo · CollEx-Persée

- ➔ Le projet ResPaDon (2021-2023)
- ➔ Le *work package* 4 (2021-2022)
- ➔ Le data sprint comme expérimentation
- ➔ Les résultats du DataSprint ResPaDon

ResPaDon

ResPaDon : un réseau de partenaires pour développer et étendre l’usage des archives du web par les chercheurs

Réseau de Partenaires pour l’analyse et l’exploration de Données numériques

RES
PA
DON

Un constat initial

1

L’accès à la documentation, voire au patrimoine numérique national est un enjeu majeur pour la recherche.

2

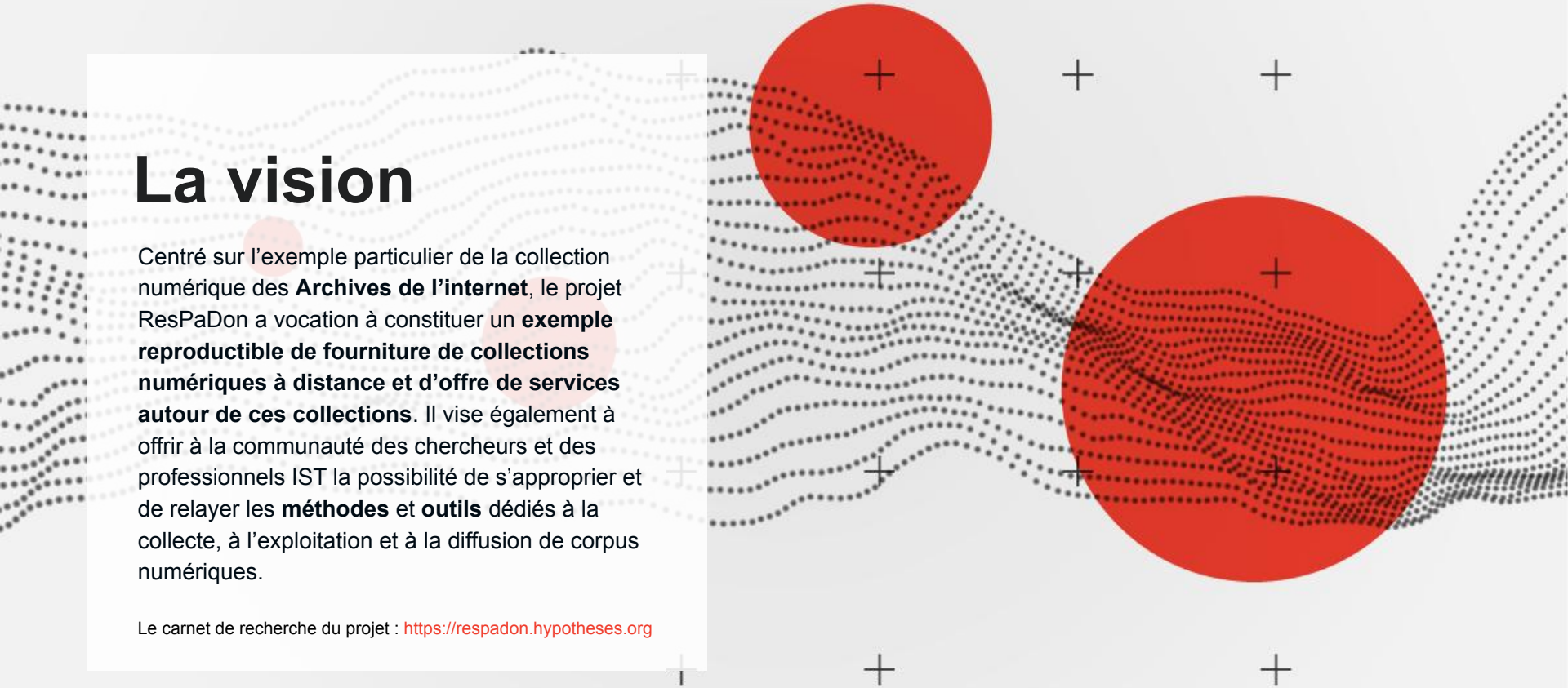
Depuis plus de 10 ans, les projets de collectes massives de documents numérisés ou nativement numériques ouvrent de nouvelles possibilités pour la recherche et permettent, voire demandent un renouvellement des usages et des méthodes d’exploitation de corpus documentaires.

Emmanuelle Bermès, Eleonora Moiraghi. **Le patrimoine numérique national à l’heure de l’intelligence artificielle**. *Revue Ouverte d'Intelligence Artificielle*, Volume 1 (2020), no. 1, pp. 89-109. DOI : 10.5802/roia.5. https://roia.centre-mersenne.org/articles/10.5802/roia.5

La vision

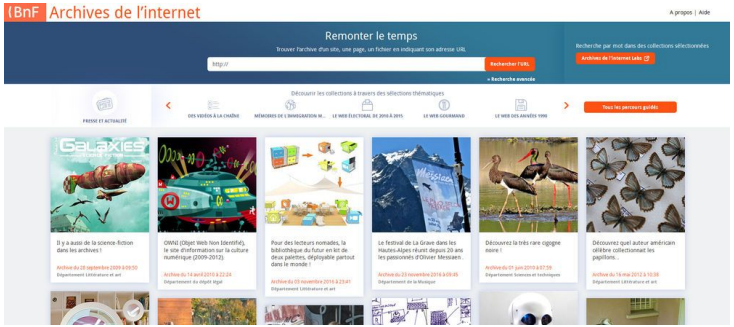
Centré sur l'exemple particulier de la collection numérique des **Archives de l'internet**, le projet ResPaDon a vocation à constituer un **exemple reproductible de fourniture de collections numériques à distance et d'offre de services autour de ces collections**. Il vise également à offrir à la communauté des chercheurs et des professionnels IST la possibilité de s'approprier et de relayer les **méthodes** et **outils** dédiés à la collecte, à l'exploitation et à la diffusion de corpus numériques.

Le carnet de recherche du projet : https://respadon.hypotheses.org



RES
PA
DON

Les collections d'archives web de la BnF (1996-2022)



L'**archivage du web** s'inscrit depuis 2006 dans le cadre de la mission de dépôt légal de la BnF.
Il porte sur le domaine **français.**
Les **collectes sont réalisées à l'aide d'un robot-logiciel**.
La collecte ne prétend pas à l'exhaustivité mais repose **sur un principe de représentativité**.

La collection est **consultable** dans les salles de recherche des différents sites de la **BnF** et dans les **bibliothèques de dépôt légal imprimeur**.

Carnet de recherche Web corpora : https://webcorpora.hypotheses.org
Guide d'utilisation : https://www.bnf.fr/sites/default/files/2021-10/Guide_Archives_internet_BnF_2021_3.pdf
Parcours guidés : https://www.bnf.fr/fr/parcours-guides-archives-de-linternet
Jeux de données en accès ouvert : https://api.bnf.fr/fr/recherche?f%5B0%5D=sources:195
Bibliothèques de dépôt légal imprimeur donnant accès aux Archives de l'Internet : https://www.bnf.fr/fr/selection-partagee-et-acces-en-region-aux-archives-de-linternet

RES
PA
DON

Un réseau de partenaires

Réseau de Partenaires pour l'exploration et l'analyse de Données numériques

Projet de deux ans (2021-2023), co-porté par la BnF et l'Université de Lille, en partenariat avec Sciences Po et le Campus Condorcet

Deux laboratoires sont impliqués : GERiiCO (Université de Lille), médialab (Sciences Po)

Université de Lille

BnF Bibliothèque nationale de France

SciencesPo

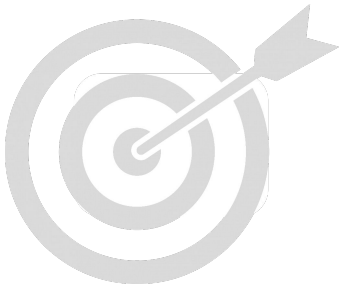
CAMPUS CONDORCET
PARIS - AUBERVILLIERS

GériiCO

CollEx-Persée

https://geriico.univ-lille.fr
https://medialab.sciencespo.fr
https://www.collexpersee.eu

Les objectifs du projet



1. **Analyser** les **usages** actuels et potentiels des **Archives de l'internet**, et par extension des autres collections numériques nationales.

2. **Expérimenter** des **dispositifs d'accès** et des **méthodes d'exploitation** de cette collection.

3. **Déduire** des **préconisations** en matière de démarches d'**accompagnement**, de **répartition des rôles** entre acteurs, de **compétences** et d'**outils** nécessaires.

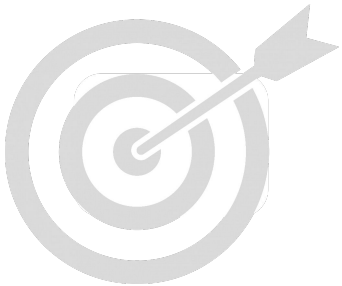
Les *work packages*

1. **Enjeux stratégiques et préconisations** / E. **Bermès** et M. **Géroudet**
2. Enquête sur les usages / L. **Favier (GERiiCO)**
3. Expérimentation d'une capsule d'accès à distance / M. **Cros** et S. **Aubry**
4. **Exploration du web vivant et archivé avec Hyphe** / E. **Moiraghi**
5. Communication / C. **Ferjoux**

Présentation des groupes de travail : https://respadon.hypotheses.org/113

WP4

Les objectifs du WP4



1. **Faire évoluer** l'application de constitution et curation de corpus web **Hyphe**, développée par le **médialab de Sciences Po**, pour permettre son utilisation sur les archives du web et la collection des Archives de l'internet de la BnF en particulier.
2. **Expérimenter et évaluer** la faisabilité d'un travail complémentaire entre web vivant et web archivé tel qu'à la BnF.
3. **Fournir** des **exemples** d'utilisation de Hyphe sur les archives du web et un **guide méthodologique** pour enquêter sur le web « vivant » et le web « archivé » par la BnF via cet outil.

Les sources et outils impliqués

- **Archives de l'internet** est l'**application** qui donne accès aux **collections de dépôt légal du web de la BnF**. Elle est fondée sur le **socle technique** de l'OpenWayback, originellement développée par Internet Archive. 1,6 Po de données sont ainsi accessibles via la recherche par URL.

- **Hyphe** est un **logiciel libre** conçu pour offrir aux chercheurs et étudiants un outil de création et nettoyage de corpus **web** reposant sur **un crawler orienté pour la recherche**. Les utilisateurs sont accompagnés par une **méthodologie** pour construire leur corpus web de manière à la fois granulaire et flexible avec des principes de curation simples.

1. Évolution du code de Hyphe

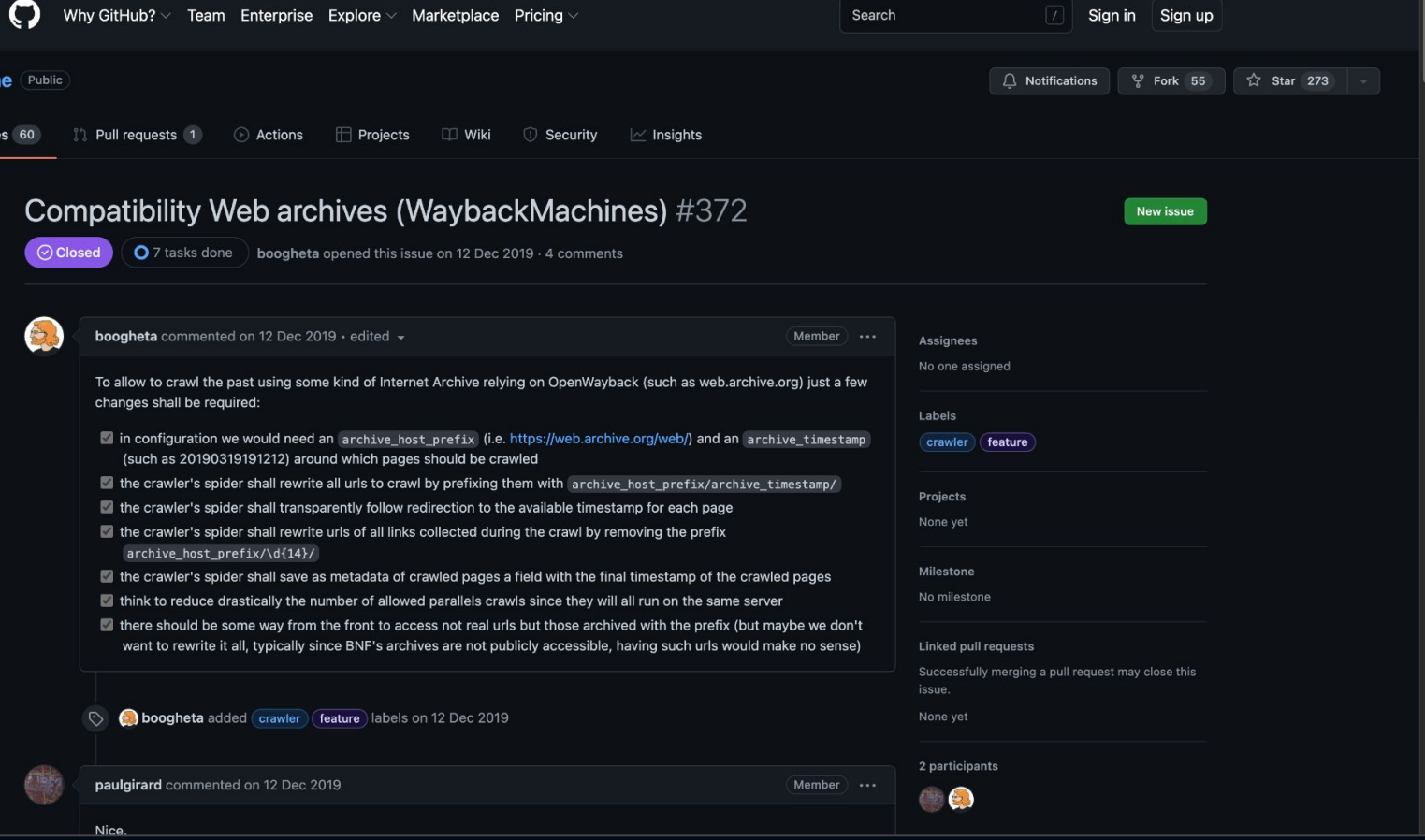
Évolution du **code** de Hyphe pour qu'il puisse fonctionner aussi bien sur le web que sur l'application Archives de l'Internet de la BnF et la wayback Machine d'Internet Archive.

2. Expérimentation sous la forme d'un data sprint au BnF DataLab

Organisation d'un data sprint de 5 jours réunissant des ingénieurs, des chercheurs et des personnels IST pour **tester et déterminer** si des logiques d'**approche comparative entre web « vivant » et web « archivé »** sont possibles.

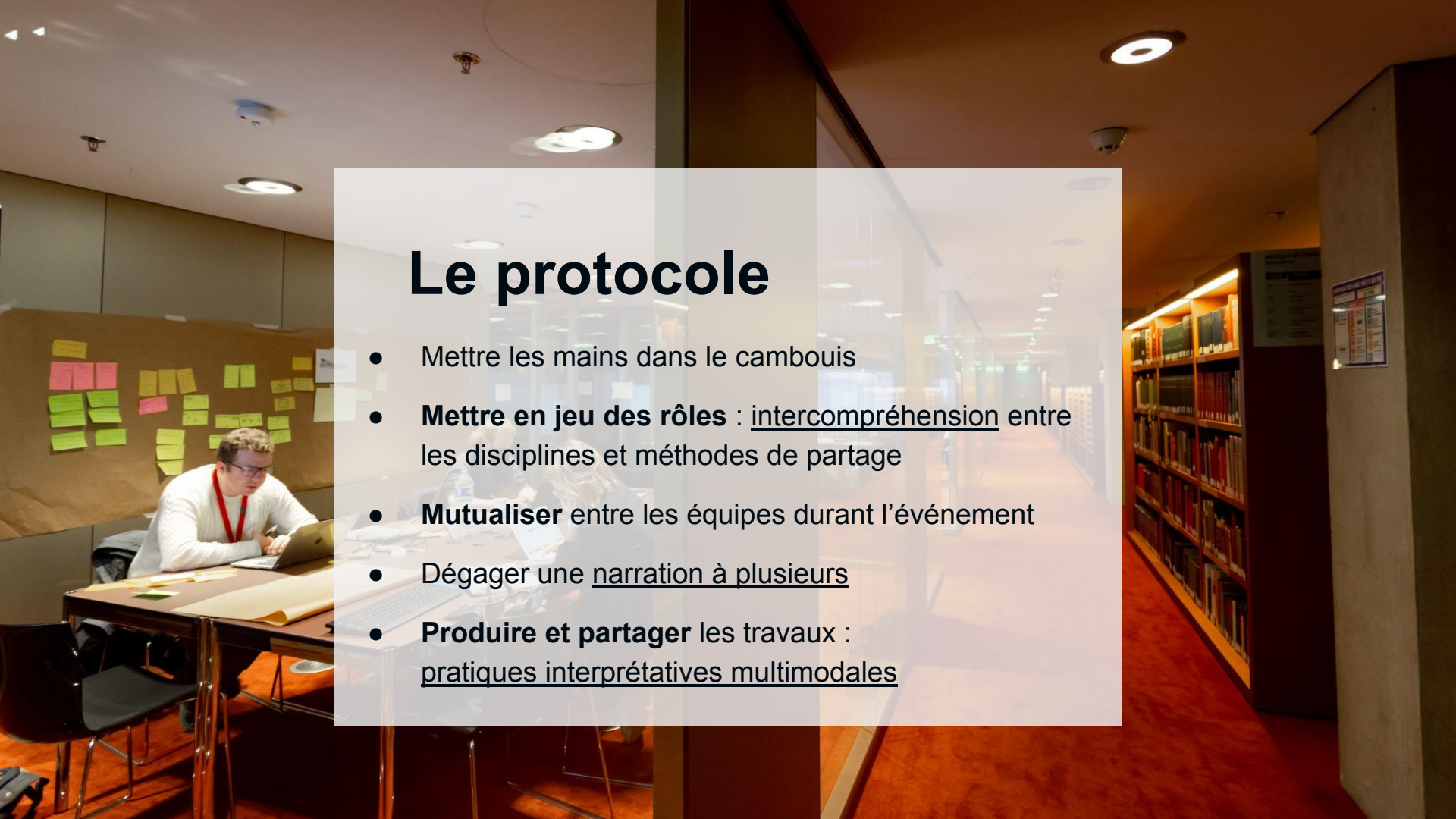
3. Diffusion des résultats sur un site web

Développement d'un site web restituant les résultats du DataSprint :
- description des analyses et mise à disposition des **visualisations des données**
- **guide méthodologique** pour l'exploration du web « vivant » et du web « archivé » par la BnF avec Hyphe.


Why GitHub?
Team
Enterprise
Explore
Marketplace
Pricing
Search
Sign in
Sign up
medialab / hyphe
Public
Notifications
Fork 55
Star 273
Code
Issues 60
Pull requests 1
Actions
Projects
Wiki
Security
Insights
Compatibility Web archives (WaybackMachines) #372
New issue
Closed
7 tasks done
boogheta opened this issue on 12 Dec 2019 · 4 comments
boogheta commented on 12 Dec 2019 • edited
Member
To allow to crawl the past using some kind of Internet Archive relying on OpenWayback (such as web.archive.org) just a few changes shall be required:
in configuration we would need an archive_host_prefix (i.e. https://web.archive.org/web/) and an archive_timestamp (such as 20190319191212) around which pages should be crawled
the crawler's spider shall rewrite all urls to crawl by prefixing them with archive_host_prefix/archive_timestamp/
the crawler's spider shall transparently follow redirection to the available timestamp for each page
the crawler's spider shall rewrite urls of all links collected during the crawl by removing the prefix archive_host_prefix/\d{14}/
the crawler's spider shall save as metadata of crawled pages a field with the final timestamp of the crawled pages
think to reduce drastically the number of allowed parallels crawls since they will all run on the same server
there should be some way from the front to access not real urls but those archived with the prefix (but maybe we don't want to rewrite it all, typically since BNF's archives are not publicly accessible, having such urls would make no sense)
boogheta added crawler feature labels on 12 Dec 2019
paulgirard commented on 12 Dec 2019
Member
Nice.
Assignees
No one assigned
Labels
crawler
feature
Projects
None yet
Milestone
No milestone
Linked pull requests
Successfully merging a pull request may close this issue.
None yet
2 participants

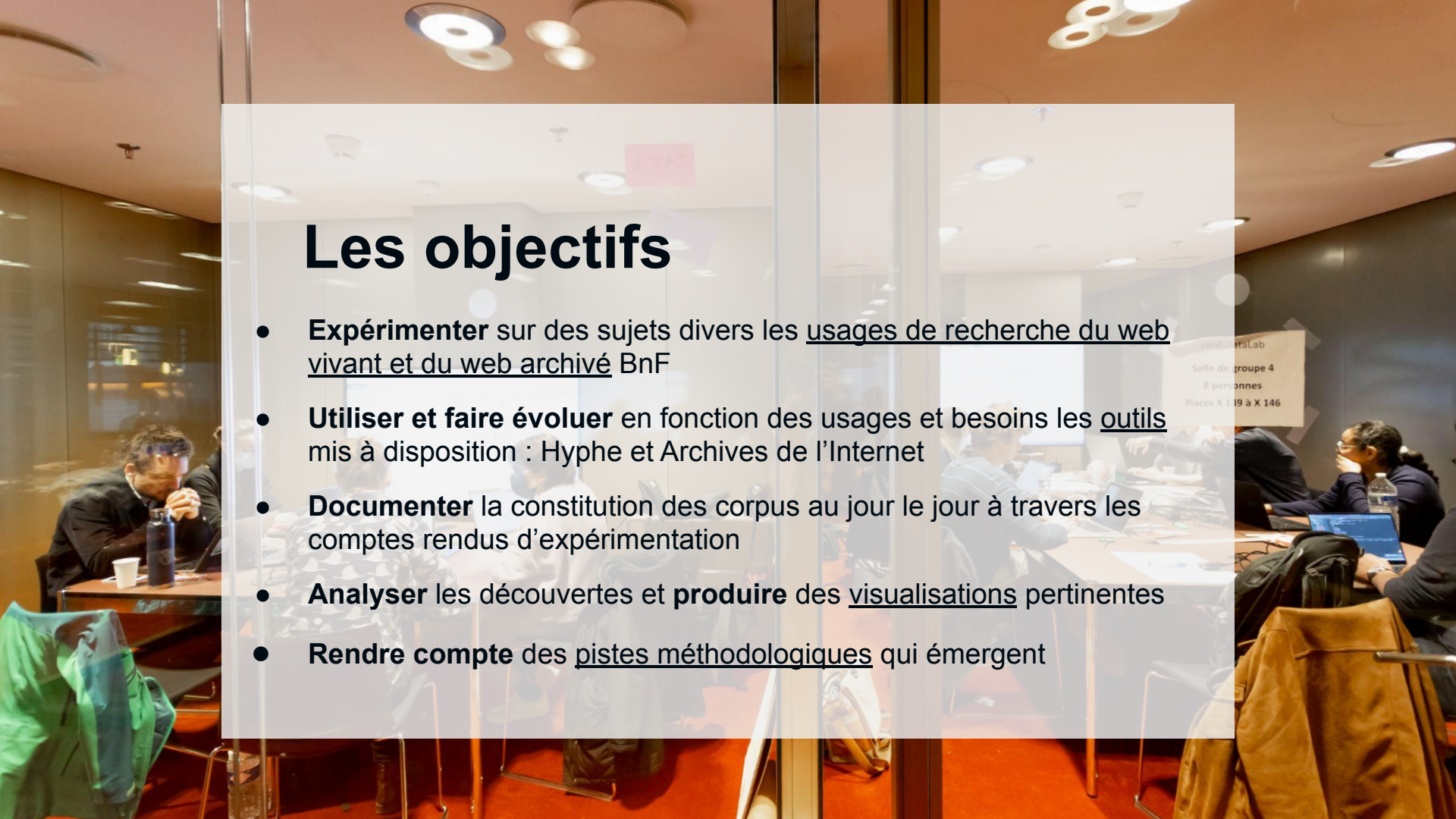
DataSprint

4 - 8 avril 2022, BnF DataLab



Le protocole

- Mettre les mains dans le cambouis
- **Mettre en jeu des rôles** : intercompréhension entre les disciplines et méthodes de partage
- **Mutualiser** entre les équipes durant l'événement
- Dégager une narration à plusieurs
- **Produire et partager** les travaux : pratiques interprétatives multimodales



Les objectifs

- **Expérimenter** sur des sujets divers les usages de recherche du web vivant et du web archivé BnF
- **Utiliser et faire évoluer** en fonction des usages et besoins les outils mis à disposition : Hyphe et Archives de l'Internet
- **Documenter** la constitution des corpus au jour le jour à travers les comptes rendus d'expérimentation
- **Analyser** les découvertes et **produire** des visualisations pertinentes
- **Rendre compte** des pistes méthodologiques qui émergent

résultats

Groupe 1

Crise de la Covid-19. Positionnement des acteurs du web par rapport aux institutions

Leslie Bellony, Guillaume Brioudes, Isabelle Degrange, Alexandre Faye, Alexis Jacomy, Kevin Locoh-Donou, Caroline Sala

RES
PA
DON

Sujet

- étudier le **positionnement des acteurs du web par rapport aux institutions** en s'intéressant à leurs liens (sites institutionnels, collectivités territoriales, établissements de santé, presse et plateforme médicale, presse généraliste et locale, association, syndicats, etc.) ;
- interroger d'éventuelles **recompositions du champ**

Démarche

- **approche chronologique** sur 3 grandes périodes : 2020 / 2021 / 2022
- exploitation des **métadonnées** produites par les sélectionneurs lors de la **collecte (api.bnf.fr)**
- important travail de **qualification** des acteurs pour obtenir des graphes pertinents (outils pour nettoyer et interfacer : OpenRefine et scripts Python)
- **test de nombreux outils pour la visualisation** : Nansi, Graphololy, Sigma, D3JS

Points saillants

- mise au point d'un **nouveau type de visualisation pour comparer des graphes entre eux en reposant sur le maintien de noeuds fixes d'une période à une autre**
- se servir de la cartographie comme un point d'entrée dans les archives et faciliter l'appréhension de la collection


2020
2021
2022
Designer : Alexis Jacomy

Groupe 2

Pour une cartographie de la critique en ligne des arts du spectacle.

Cristina Tosetto (chercheuse associée au laboratoire CLARE de l'Université de Bordeaux Montaigne), Béatrice Mazoyer, Guillaume Plique, Clara Wiatrowski, Antoine De Sacy

Sujet

- comprendre dans quelle mesure les **sites de critique des arts** du spectacle parviennent à créer des **communautés en ligne**
- analyser les **liens potentiels entre ces sites de critique et les acteurs institutionnels**
- analyser l'**impact de la pandémie** et les restructurations potentielles qu'elle a pu impliquer entre les différents acteurs de ce réseau (acteurs individuels, publics, privés, institutionnels)

Démarche

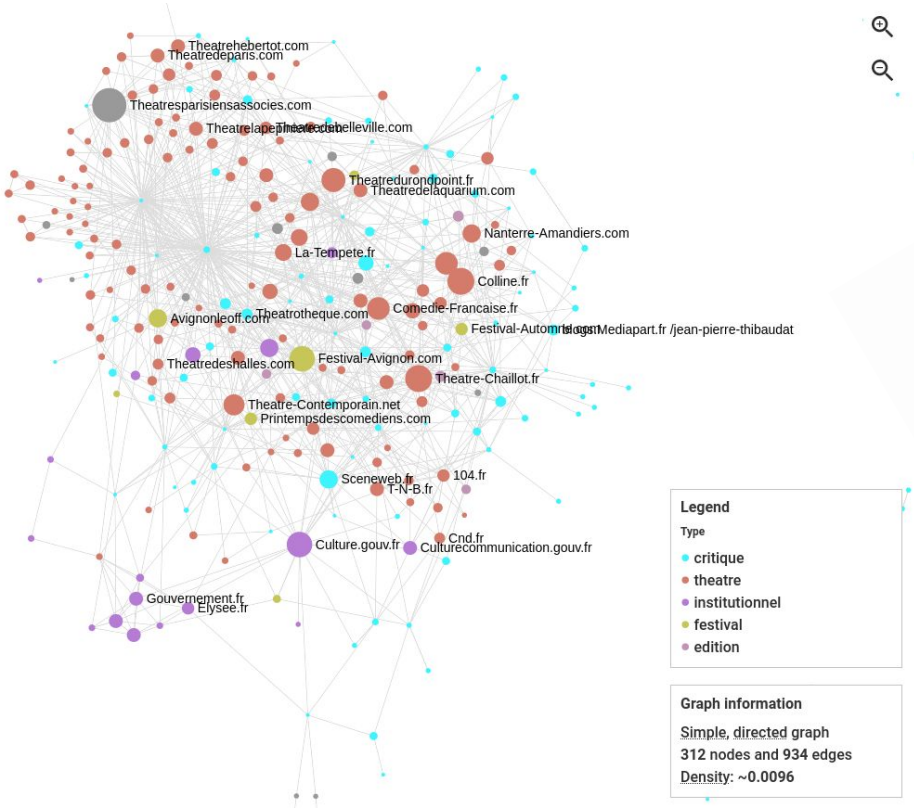
- fenêtres temporelles retenues : n-0 (2020-2022), n-1 (2018-2019), et n-2 : 2016-2017
- **approche synchronique** : interroger et représenter les relations entre les différents types d'acteurs (critiques académiques, sites étatiques, théâtre privés versus publics ; sites institutionnels, théâtres, critiques)
- **approche diachronique** cherche à percevoir les évolutions, renforcement de liens, évolution des noeuds ou sites centraux

Points saillants

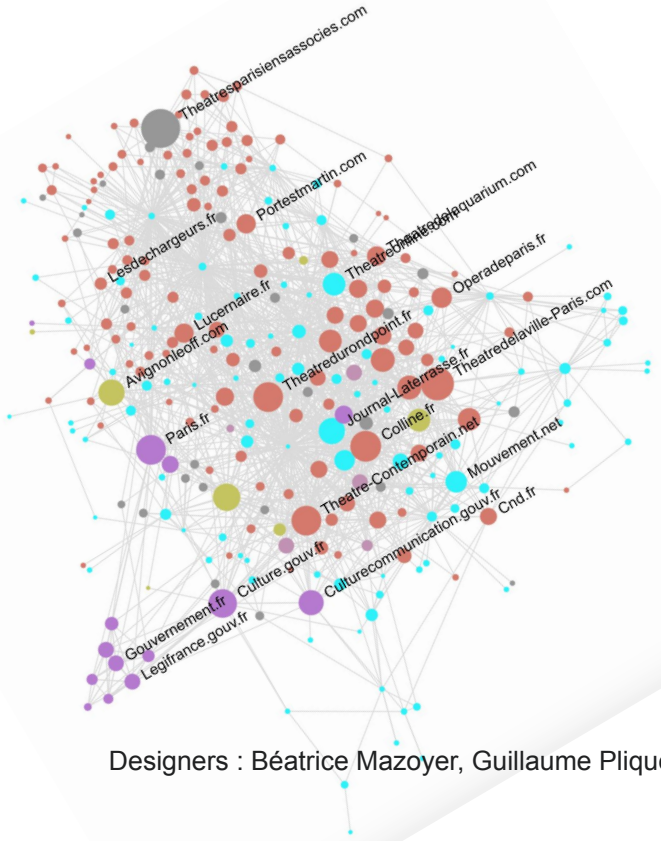
- analyses synchroniques très poussées en testant **plusieurs hypothèses**, avec des **visualisations graphiques distinctes**

RES
PA
DON

N0 (période covid -> 2022-01-31)



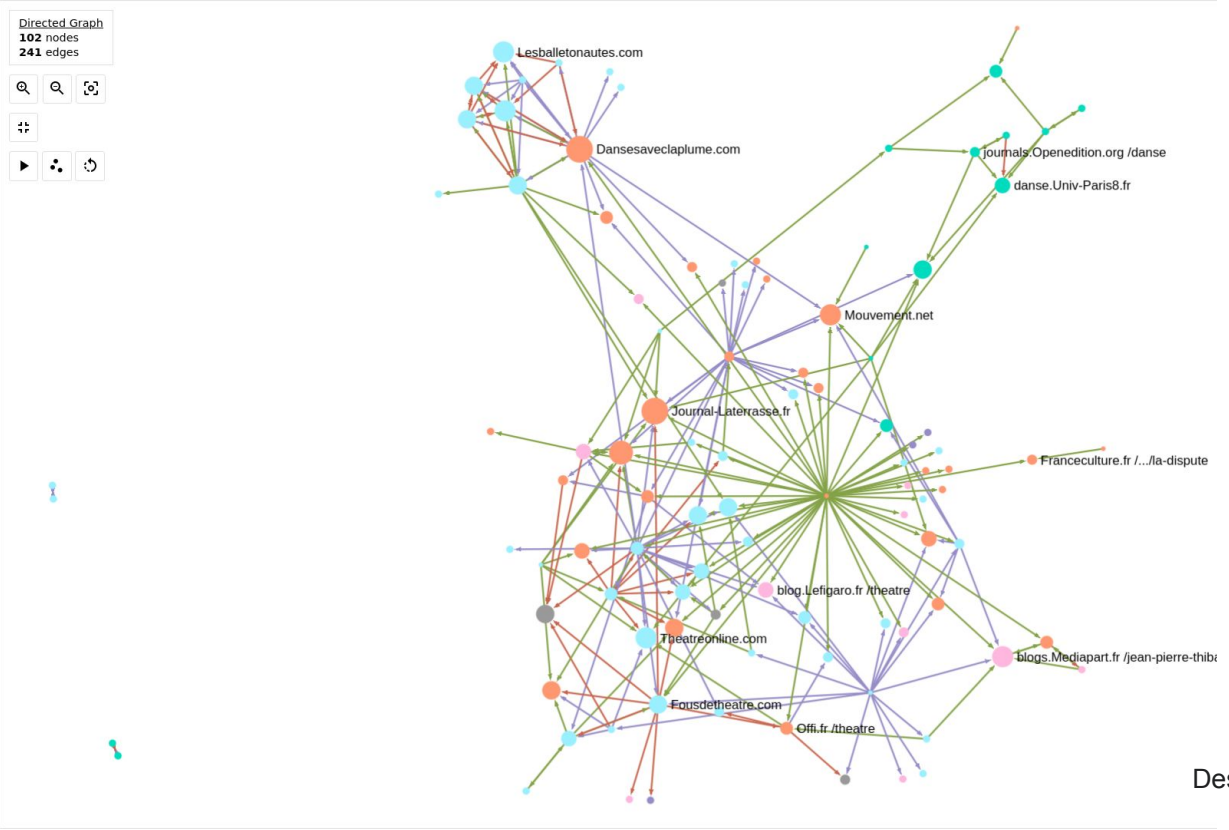
N-1 (02/02/2018 - 31/01/2020)



Designers : Béatrice Mazoyer, Guillaume Plique

RES
PA
DON

période n-1 - n-2 - liens entre critiques



Node colors

Nature attribute as a category:

- individu
- professionelle
- academique
- presse
- a decider
- ...

Node sizes

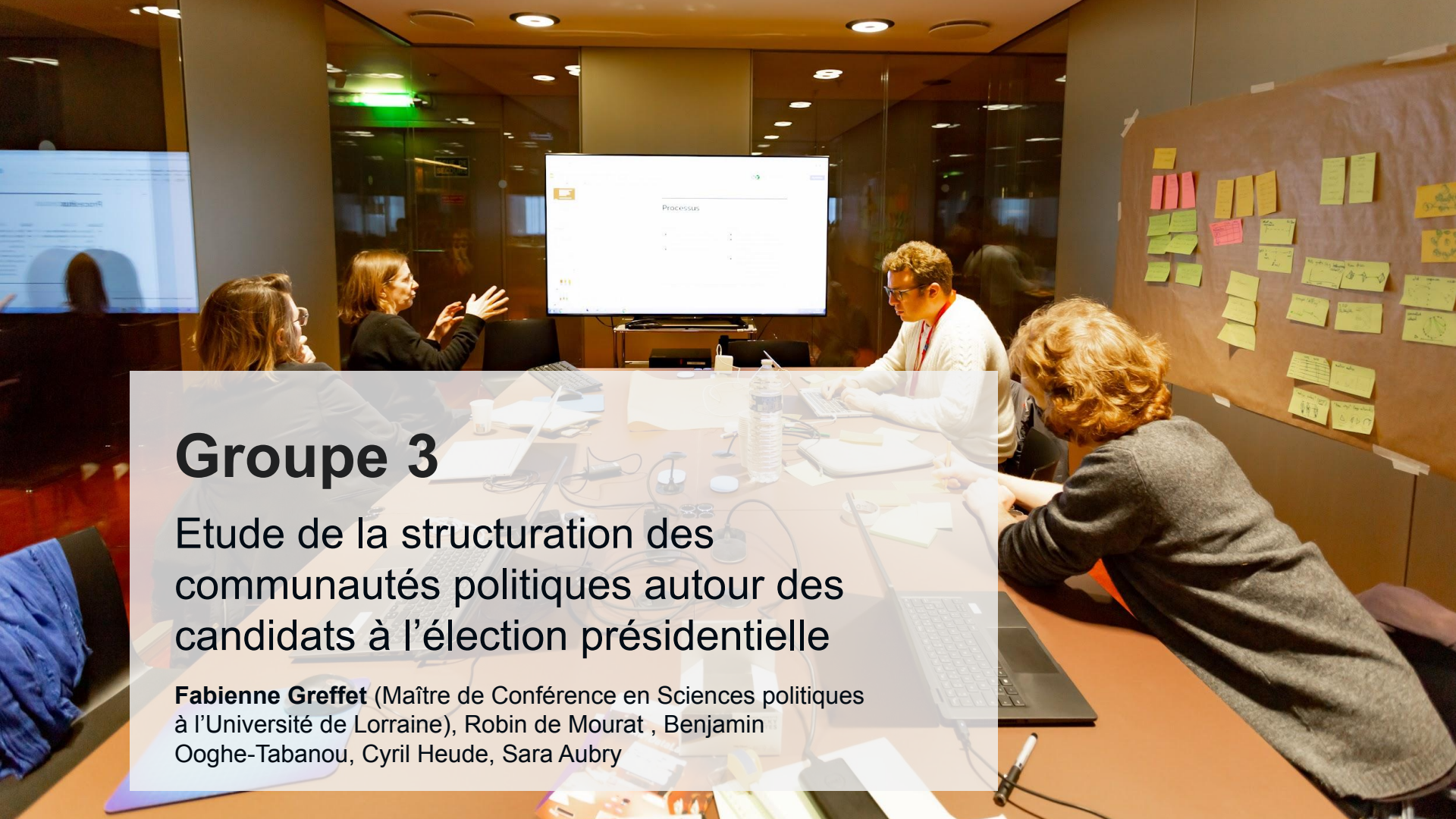
node_size kwarg (scaled to 3-15 px)

Edge colors

time attribute as a category:

- n-1
- n-2
- all
- ...

Designers : Béatrice Mazoyer, Guillaume Plique



Groupe 3

Etude de la structuration des communautés politiques autour des candidats à l'élection présidentielle

Fabienne Greffet (Maître de Conférence en Sciences politiques à l'Université de Lorraine), Robin de Mourat , Benjamin Ooghe-Tabanou, Cyril Heude, Sara Aubry

Sujet

- observer les **transformations des communautés politiques de soutien aux candidats à l'élection présidentielle**, à travers l'**exemple particulier** des communautés de soutien à **Jean-Luc Mélenchon** en 2012, 2017, 2022
- étudier les **évolutions des formes éditoriales** que prennent ces divers sites de soutien

Démarche

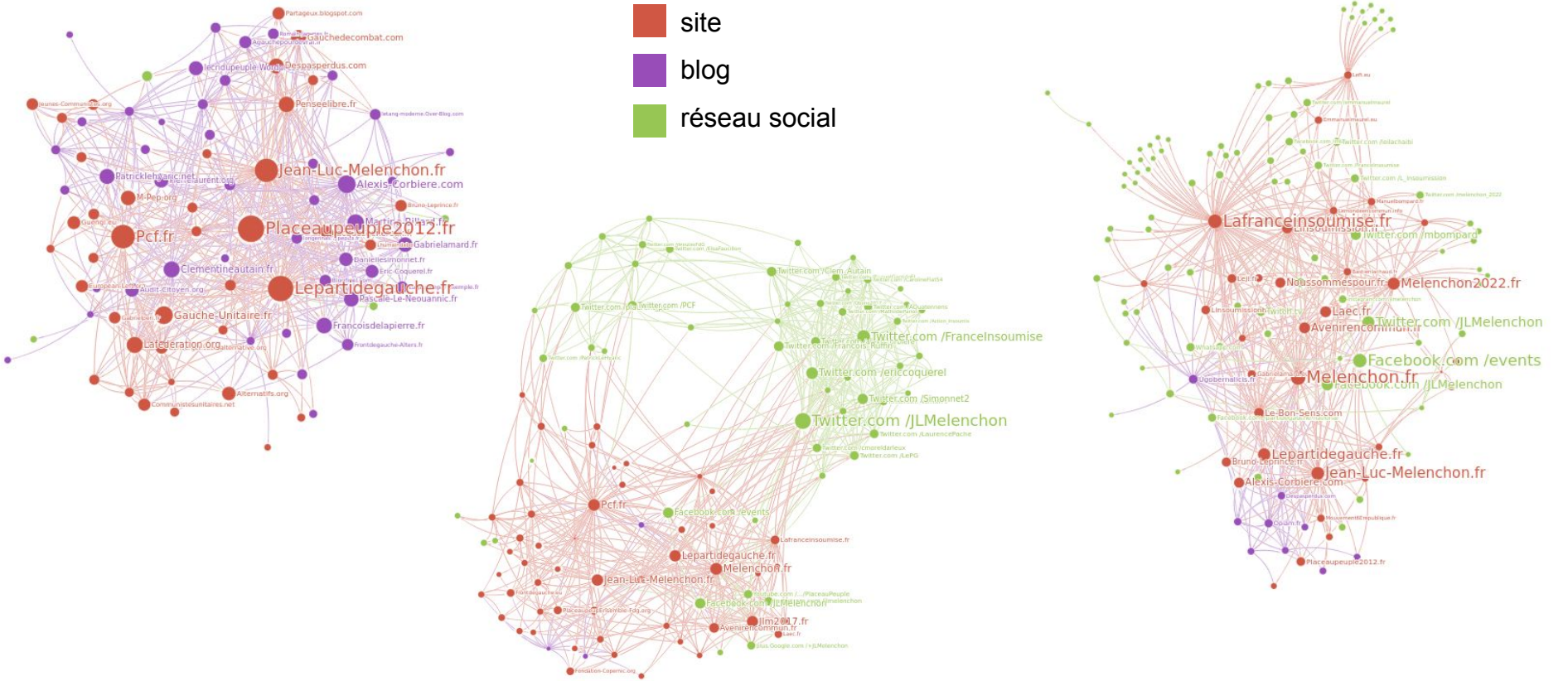
- **démarche longitudinale** : crawl de trois corpus dans trois fenêtres temporelles distinctes, centrées respectivement autour des élections présidentielles de 2012, 2017, 2022, en partant à chaque fois des mêmes types de ressources : **pages wikipedia**
- mise au point d'une **méthodologie de comparaison des trois graphes résultants**

Points saillants

- **choix de points d'entrée du crawl** stable sur les différentes périodes
- besoin de **visualiser l'évolution d'un graphe dans le temps** en distinguant les **sites en commun** sur ces trois périodes des sites spécifiques
- création d'une **méthodologie** et d'un **outil de visualisation**

Présence des entités entre 2017 et 2022

Designer : Robin de Mourat



Évolution des formes éditoriales dans la communauté entre 2012 et 2022

Designer : Robin de Mourat

Groupe 4

Le concept de génome dans le discours politique français

Guillaume Lévrier (doctorant au CEVIPOF de Sciences Po), Zeynep Pehlivan, Paul Girard, Jennifer Morival, Dorothée Benhamou-Suesser

Sujet

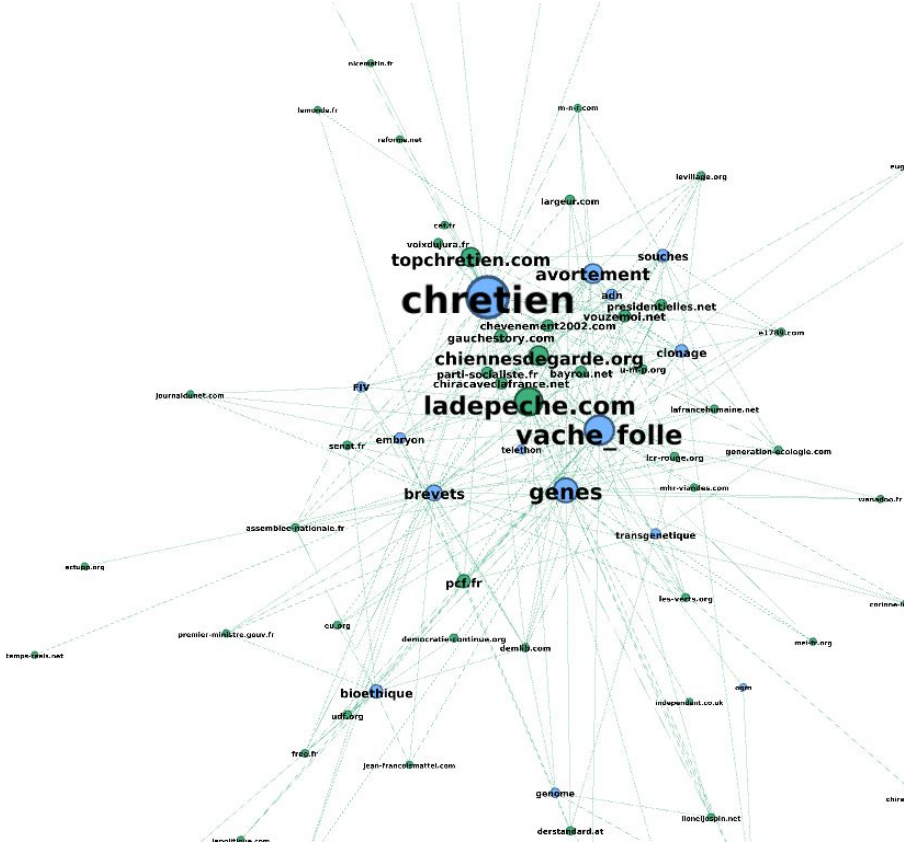
- étudier les **sources web qui publient des contenus autour de la génomique** (génom* et Crispr) sur une période longue, et en particulier les contenus politiques, et les types d'acteurs qui mobilisent ces mots : académiques, médié, politiques…
- A l'issue d'une **exploration liminaire sur 4 collections indexées en plein texte** (Incunables 1996-2000 ; Actualités (2010-2014) et Elections 2002, le corpus Elections 2002 après analyse dans Hyphe, est retenu comme le plus pertinent, et le sujet recentré sur l'analyse de la place du génome dans le discours politique

Démarche

- La démarche a conjugué des **requêtes et recherches de syntagmes sur l'index Solr plein texte** de la collection Elections 2002, de la **fouille de texte avec TF-IDF** grâce à l'interrogation du plein texte des pages, et l'utilisation de Hyphe pour repérer les **liens entre les sites** et étendre le corpus ("prospection")
- Les **métadonnées** accompagnant la collection 2002 ont été importées dans Hyphe pour enrichir le corpus

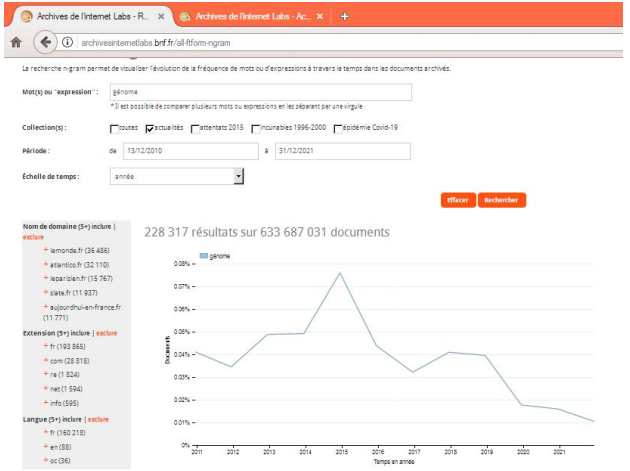
Points saillants

- Combiner **une approche fouille de texte et une approche d'étude des communautés**
- Le choix des périodes a été en partie guidé par la **disponibilité des collections indexées en plein texte**



telethon,bayer,bioethique,clonage,gene,chretien, embryon,ogm,brevets,adn, vachefolle,eugenisme,souches,transgenique,FIV avortement…

	Corpus 1	Corpus 2	Corpus 3	Corpus 4
Recherche	*génom**	*génom**	*CRISPR*	*CRISPR*
Source	Incunables	Élections 2002	Actualités	Actualités
Dates	1996-2000	2000-2004	2010-2014	2015-nov2018
Nb Hits Solr	8713	9974	156	21 213
URL uniques	5 427	2 603	29	4 777
Web Entités	728	113	20	274
Nb domaines	720	94	13	156
Exploration Observable	https://observablehq.com/d/fa865a0af1dbfbe2	https://observablehq.com/d/3a700e1dd589d9f6	https://observablehq.com/d/b6bfb547b6d6f6cc	https://observablehq.com/d/2138ab099b729dcd
Mode	Domain	Domain	Subdomain	Domain
Depth premier crawl	1	0	0	0



Les approches méthodologiques

Groupe 1 : Hyphe comme outil heuristique permettant de visualiser la constitution des collections, leurs faiblesses et leurs manques / ou « que nous apprend Hyphe sur la collection COVID-19 de la BnF ? »

Groupe 2 : comparaison de trois corpus web archivés, étude des évolutions d'un champ et combinaison d'une approche synchronique et diachronique

Groupe 3 : comparaison entre deux corpus du web archivé et un corpus du web vivant (trois coupes temporelles pour étudier la structuration des communautés politiques sur le web à partir de points d'entrée fixes)

Groupe 4 : complémentarité entre fouille de texte et Hyphe

Des questions transverses

- Importance de la connaissance de la fabrique des collections et de leurs biais. **Qu'observe-t-on / que représente-t-on ?**
- **Comment rendre compte, visualiser les évolutions entre web archivé et web vivant, ou entre web archivé sur deux fenêtres temporelles** (par exemple représenter les évolutions des réseaux) **?**
- **Comment interfacer les outils et les inscrire dans un processus d'analyse de corpus itératif ?** scripts de nettoyage, d'interfaçage

Les intérêts pour ResPaDon

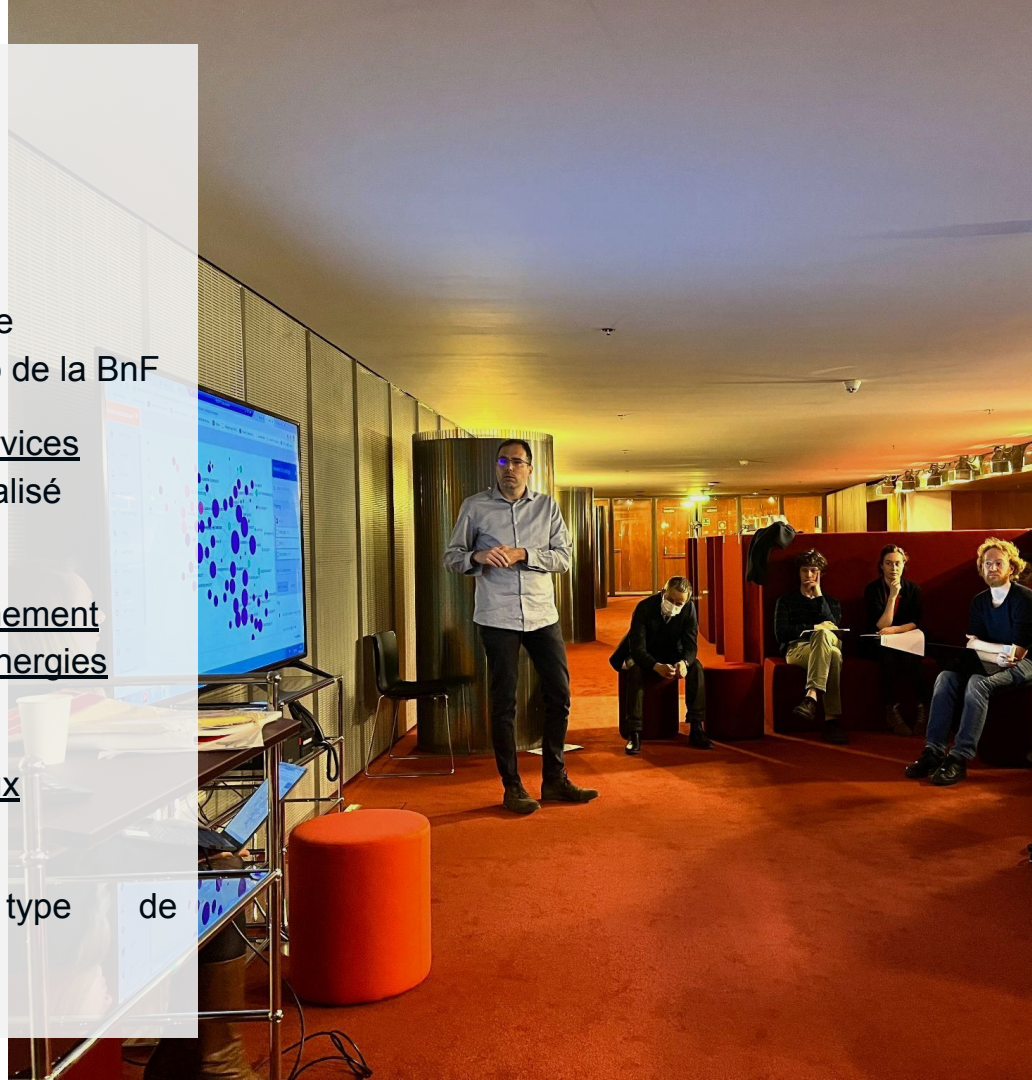
Fournir des éléments pour améliorer le workflow de constitution et documentation des archives du web de la BnF

Intégrer Hyphe au workflow et au catalogue de services du BnF DataLab et des « capsules » (accès délocalisé aux archives du web)

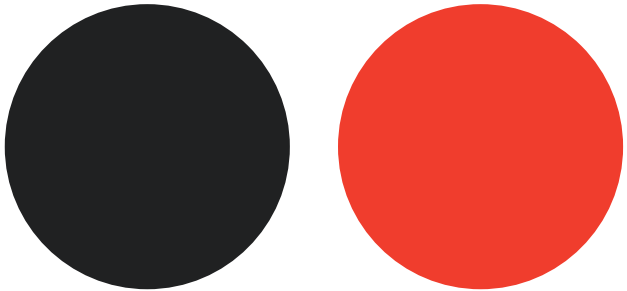
Voir émerger de nouvelles modalités d'accompagnement des chercheurs pour les acteurs de l'IST et des synergies entre différentes communautés de recherche

Identifier des problèmes techniques et de nouveaux besoins autour des archives du web

Éprouver des hypothèses méthodologiques sur ce type de sources



Contacts



COLLEXPERSEE.EU/PROJET/RESPADON

GROUPES.RENATER.FR/SYMPA/INFO/RESPADON

RESPADON@UNIV-LILLE.FR

@Respadon_Projet

médialab : audrey.baneyx@sciencespo.fr
BnF : depot.legal.web@bnf.fr