RES
PA
DON

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

BILAN FINAL

Novembre 2023

Introduction

Ce document présente le bilan final du projet ResPaDon (Réseau de Partenaires pour l'analyse et l'exploration de données numériques). Le projet ResPaDon a débuté le 25 mars 2021, par la signature d'une convention entre l'Université de Lille et le GIS CollEx-Persée. Cette convention a été modifiée par un avenant prolongeant le projet jusqu'au 25 septembre 2023.

Rappel des objectifs du projet

Porté par l'Université de Lille et la Bibliothèque nationale de France (BnF) en partenariat avec le Campus Condorcet et Sciences Po, le projet ResPaDon se fonde sur une analyse des usages des archives du web et des corpus numériques constitués par la BnF pour expérimenter de nouvelles modalités d'accès et d'exploitation de ces collections numériques. Constitué autour d'un premier noyau d'établissements de l'enseignement supérieur et de la recherche, il s'appuie sur les initiatives prises par ces établissements en matière de services à la recherche pour favoriser le développement d'expérimentations.

Les objectifs du projet peuvent être ainsi synthétisés de la manière suivante :

- Analyser les usages actuels et potentiels des archives du web, et par extension des autres collections numériques mises à disposition par la BnF (2021).
- Expérimenter des dispositifs d'accès et des méthodes d'exploitation de ces collections (2022).
- En déduire des préconisations en matière de démarches d'accompagnement, de répartition des rôles entre acteurs, de compétences et d'outils nécessaires (2023).

Le projet est composé de 5 work packages (WP).

Intitulé du WP	Porteur	Objet principal	Livrables attendus
WP1. Enjeux stratégiques et préconisations	BnF Université de Lille	Production de préconisations opérationnelles sur l'organisation de l'accès aux archives de web en France	Rapport final Journée de clôture
WP2. Exploration des usages des archives du web	Université de Lille	Dresser une typologie des usages des archives du web par les chercheurs	Journée de lancement Typologie des usages Cas d'usages détaillés Colloque international
WP3. Expérimentation autour d'une	Université de Lille	Implémenter une capsule d'accès aux archives du	Spécifications techniques Implantation d'un prototype

capsule d'accès à distance	BnF	web dans l'emprise de l'Université de Lille	Bilan des tests du prototype et préconisations
WP4. Expérimentation autour de la constitution et du traitement de corpus web issus des archives de l'Internet, en lien avec le web vivant	Sciences Po	Rendre possible des logiques d'approche comparative entre le web vivant et les archives du web	Adaptation de Hyphe aux archives du web Datasprint Bilan de l'expérimentation et préconisations Guide de bonnes pratiques pour la comparaison des archives du web et du web vivant
WP5. Coordination, planification et communication	Université de Lille BnF	Actions de coordination, de planification et de communication autour du projet	Actions de formation Blog hypothèses

1. Les réalisations du projet

1.1. Chronologie des événements

Date	Evénement, initiative	Partenaires, publics mobilisés
2021		
Janvier	Première réunion du groupe d'organisation de la journée de lancement (WP1)	Equipe projet, GERiiCO
Janvier	1ère réunion du Comité de pilotage de ResPaDon	Copil
Février	Première réunion du groupe projet et lancement des work packages	Equipe projet
18 Mars	Réunion équipe Projet (communication)	Equipe projet
25 Mars	Signature de la Convention entre l'Université de Lille et CollEx-Persée	Université de Lille, CollEx Persée
Avril	Définition de l'identité visuelle de ResPaDon,	Equipe projet

	lancement du compte Twitter et de la liste de diffusion Renater	
17 Mai	**Journée de lancement en distanciel (WP1)**	**Public composé de chercheurs et de professionnels de l'information scientifique**
17 mai	Prise de poste de la coordinatrice du projet, Céline Ferjoux	
31 Mai	Réunion d'échange avec l'INA (visio)	Equipe projet, INA
du 19 Avril au 23 Juillet	Stage d'une étudiante (Master 1 information-documentation) sur les travaux scientifiques mobilisant les archives du web (WP2)	GERiiCO, Université de Lille
3 Juin	Réunion équipe Projet	Equipe projet
28 – 29 Juin	**Formation « en immersion » à la BnF (WP5)**	**Equipe projet, GERiiCO**
6 Juillet	Création du Carnet Hypothèses du projet (WP5)	Equipe projet
12 Juillet	1ère réunion des membres permanents du cycle d'ateliers (WP1)	ResPaDon (Copil, équipe projet) CollEx-Persée, INA, ADBU, BNUS
10 Septembre	COPIL	Copil
16 Septembre	Réunion Equipe Projet	Equipe Projet
1er Octobre	**Lancement du cycle d'ateliers, séance n°1, coordonnée par Emmanuelle Bermès (WP1)**	**Groupe de réflexion ResPaDon et chercheurs/experts invités**
14 Octobre	Webinaire n°1, restitution des travaux du Groupe de réflexion (WP1 et WP5)	Professionnels de l'information scientifique et chercheurs
17 Octobre	Réponse à l'appel à communication – « *Un patrimoine pour l'avenir, une science pour le patrimoine* » (WP5)	Equipe projet
18 Novembre	**Premier Datasprint de préfiguration au BnF Datalab (expérimentations autour de l'application Hyphe et des archives de**	**Equipe projet**

	l'Internet) (WP4)	
25 Novembre	Réunion Equipe Projet	Equipe projet
26 Novembre	**Cycle d'ateliers : séance n°2, coordonnée par Dorothée Benhamou-Suesser, Marie Cros et Eleonora Moiraghi (WP1)**	**Groupe de réflexion ResPaDon et chercheurs/experts invités**
16 Décembre	Webinaire n°2, restitution des travaux du Groupe de réflexion (WP1 et WP5)	Professionnels de l'information scientifique et chercheurs
2022		
Janvier	Réponses aux appels à communication pour les congrès LIBER et Humanistica	Equipe projet
26 janv	COPIL (bilan mi-projet)	Copil
Février	Rédaction d'un article pour l'événement « Un patrimoine pour l'avenir, une science pour le patrimoine » (Fondation des Sciences du Patrimoine)	Equipe projet
Mars	Invitation du projet ResPaDon pour une conférence plénière des Journées DHNord 2022 (MESHS)	Equipe projet
31 mars	Réunion Equipe Projet	Equipe projet
~~**25 mars**~~ **Reprogrammé le 10 juin**	**Cycle d'ateliers : séance n°3, coordonnée par Madeleine Géroudet et Arnaud Laborderie (WP1)**	**Groupe de réflexion ResPaDon et chercheurs/experts invités**
4-8 avril	**Datasprint ResPaDon au BnF Datalab (WP4)**	**Professionnels de l'information scientifique et chercheurs**
Mai	Début de la phase de tests utilisateur de la capsule d'accès à distance à l'Université de Lille (WP3)	Professionnels de l'information scientifique et chercheurs
13 mai	**Cycle d'ateliers : séance n°4, coordonnée par Amélia Laurenceau et Alexandre Faye (WP1)**	**Groupe de réflexion ResPaDon et chercheurs/experts**

		invités
2 juin	Webinaire n°4, restitution des travaux du Groupe de réflexion (WP1 et WP5)	Professionnels de l'information scientifique et chercheurs
9 Juin	Réunion Equipe Projet	Equipe Projet
23 juin	Webinaire n° 3 : restitution des travaux du Groupe de réflexion (WP1 et WP5)	Professionnels de l'information scientifique et chercheurs
27 Juin	Réunion Equipe Projet	Equipe Projet
12 juillet	**Cycle d'ateliers : séance n°5 coordonnée par Laurence Favier et Emmanuelle Bermès**	**Groupe de réflexion ResPaDon et chercheurs/experts invités**
8 septembre	Réunion préparation COPIL, point de suivi des WP, mobilisation événements 2023	Equipe projet
12 septembre	Webinaire n°5, restitution des travaux du Groupe de réflexion (WP1 et WP5)	Professionnels de l'information scientifique et chercheurs
19 septembre	Réunion du Comité de pilotage de ResPaDon	Copil
6 octobre	**Cycle d'ateliers : séance n°6, coordonné par Madeleine Géroudet et Dorothée Benhamou-Suesser**	**Groupe de réflexion ResPaDon et chercheurs/experts invités**
20 octobre	Webinaire n°6, restitution des travaux du Groupe de réflexion (WP1 et WP5)	Groupe de réflexion ResPaDon et chercheurs/experts invités
17novembre	**Atelier « enjeux juridiques » (suite) coordonné par Madeleine Géroudet et Arnaud Laborderie (WP1)**	**Groupe de réflexion ResPaDon et chercheurs/experts invités**
24 novembre	Réunion Equipe Projet	Equipe projet
25 novembre	Webinaire collectes et usages scientifiques du web électoral	chercheurs/experts invités

	Suivi – WP3 Capsule	
8 décembre	**Cycle d'ateliers : séance n°7, coordonnée par Eleonora Moiraghi et Anaïs Crinière-Boizet**	**Groupe de réflexion ResPaDon et chercheurs/experts invités**
2023		
16-17 janvier	**Formation « En immersion à l'Université de Lille »** **Retour d'expérience sur l'implémentation de la capsule à l'Université de Lille**	**Équipe projet élargie aux partenaires**
19 janvier	Webinaire n°7, restitution des travaux du Groupe de réflexion (WP1 et WP5)	Groupe de réflexion ResPaDon et chercheurs/experts invités
26 janvier	**Cycle d'ateliers : séance n°8**	**Groupe de réflexion ResPaDon et chercheurs/experts invités**
9 février	Réunion Equipe Projet	Equipe Projet
10 février	**Journée de restitution de résultats du datasprint à Sciences Po**	**Groupe de réflexion ResPaDon et chercheurs/experts invités et participants du datasprint**
7 mars	COPIL	
13 mars	**Journée d'étude : rencontre professionnelle à la BnF « faire réseau autour des archives du web : Perspectives du projet ResPaDon »**	**Professionnels de l'IST**
3-4-5 avril 2023	**Colloque international « Le web : source et archive »**	**Chercheurs**
25 mai	Réunion Equipe Projet	Equipe Projet
27 juin	COPIL « bilan scientifique »	Copil
11 sept	COPIL « bilan et perspectives »	Copil

1.2. Temps forts structurants du projet

1.2.1. Journée de lancement du 17 mai 2021

Initialement prévue comme un temps de lancement de la démarche d'enquête sur les usages des archives du web, la journée de lancement a été repensée à la demande du Conseil scientifique du GIS CollEx-Persée. Elle s'est articulée autour d'un double objectif : lancer le réseau autour du projet et réaliser un premier état de l'art sur la recherche sur les archives du web. Un comité d'organisation associant chercheurs de l'équipe GERiiCO et membres de l'équipe projet a été mis en œuvre dès janvier 2021 pour permettre la tenue de la journée le 17 mai de la même année.

Intitulé « Faire réseau autour des archives du web : usages et opportunités », le programme associait des temps de présentation des enjeux relatifs aux archives du web en France et à l'international, des retours d'usages par les chercheurs, des ateliers thématiques et un temps final sur la constitution d'un réseau autour des archives du web. L'événement a rassemblé à distance 26 intervenants (dont 16 chercheurs) et un public de 209 inscrits. Le programme de la journée a recueilli une large adhésion des inscrits puisque près de 60% ont un taux de présence supérieur ou égal à 75%. Le profil varié des participants témoigne de l'ancrage académique du projet : la manifestation a réuni chercheurs et enseignants chercheurs, professionnels de l'information scientifique (en bibliothèques ou fonctions supports dans des laboratoires), représentants de l'INA et des Bibliothèques dépôt légal imprimeur (BDLI, principalement des bibliothèques municipales qui accueillent des points d'accès aux archives du web.)

Comme le montrent les nombreuses réactions exprimées autour de cet événement de lancement, ce temps fort du projet a ouvert des perspectives de collaboration prometteuses avec de nombreux acteurs du domaine des archives du web (International Internet Preservation Consortium – IIPC, réseaux WARCNet et RESAW), et des humanités numériques (MeSHS Lille Nord de France, Maison Méditerranéenne des Sciences de l'Homme – MMSH).

La valorisation de cette journée a été l'occasion d'inaugurer le Carnet Hypothèses du projet avec la publication des enregistrements vidéo des interventions (https://ResPaDon.hypotheses.org/1). Cette réalisation a bénéficié de l'appui technique de la WebTV de l'Université de Lille.

1.2.2. Journées de formation « en immersion » à la BnF des 28 et 29 juin 2021

A la fin du mois de juin 2021, en raison de l'évolution de la situation sanitaire, la première action de formation du projet ResPaDon s'est tenue en semi-distanciel. Une partie de la formation a été réalisée en visio-conférence ; une journée complète a été maintenue en présentiel. La journée en présentiel a été l'occasion de réunir pour la première fois l'ensemble des partenaires du projet à la BnF (15 participants, en respect de la jauge sanitaire). Le programme de la formation a été conçu pour favoriser le partage d'une culture commune autour de la collection des archives du dépôt légal de l'Internet. Ces journées ont réuni des chercheurs et des professionnels de l'information scientifique venus des quatre établissements partenaires du projet.

Cette action de formation initiale a été identifiée comme pouvant constituer le modèle d'une future offre de formations sur les archives du web à destination d'une communauté plus large.

1. 1.2.3. Journées de formation « Immersion à la BU : créer la relation avec les chercheurs », Université de Lille, 16 et 17 janvier 2023

En miroir des journées réalisées à la BnF en début de projet, deux journées ont été organisées en janvier 2023, afin de proposer aux partenaires du projet ResPaDon une immersion dans les bibliothèques de l'Université de Lille, autour de la thématique de « la relation avec les chercheurs ». Elles ont réuni une quinzaine de participants venus des quatre établissements du projet. Ces journées ont proposé un aperçu de la variété des relations aux chercheurs et services développés au SCD de l'Université de Lille et ont aussi présenté un retour d'expérience sur l'expérimentation des capsules d'accès à distance aux collections des archives de l'internet. Des visites des bibliothèques ont également eu lieu.

Ces journées d'immersion ont été l'occasion d'échanges, de débats, et de partages des membres du projet entre eux, et avec l'ensemble des collègues du SCD mobilisés à cette occasion.

2. 1.2.4. Rencontres professionnelles « Faire réseau autour des archives du web. Bilan et perspectives du projet ResPaDon » le 13 mars 2023.

3. Une journée de présentations et d'échanges a été planifiée autour des résultats et des perspectives du projet. Elle a eu lieu à la BnF le 13 mars 2023 et a réuni un public de professionnels de l'information scientifique et technique.

4. La journée a été consacrée pour partie à des sessions de présentation et des ateliers autour de la fabrique des archives du web, de leurs enjeux et des modes d'exploration développés dans le cadre du projet ResPaDon. Les retours d'expérience sur les expérimentations menées pendant le projet ont donné lieu à des sessions dédiées. La journée s'est conclue par la présentation des préconisations du projet ainsi que par un temps d'échanges permettant de dessiner collectivement les perspectives ouvertes pour poursuivre les démarches engagées.

5. Les interventions de la journée ont été enregistrées et sont diffusées sur la chaîne Youtube de la BnF : https://youtube.com/live/Aq6xZ2vBY7k?feature=share

6. 1.2.5. Colloque international « Le Web : source et archive » du 3 au 5 avril 2023

Un colloque intitulé "*Le web : source et archive*" a eu lieu à l'Université de Lille du 3 au 5 avril 2023. Cet événement scientifique international a réuni une large communauté académique et les partenaires du projet ResPaDon pour s'interroger ensemble sur la place des sources issues du web dans la recherche et situer les pratiques d'archivage d'Internet dans des démarches et des questionnements pluriels. Le colloque a réuni 72 participants et a été diffusé en streaming sur la WebTV de l'Université de Lille.

Les propositions de communication portaient sur les thématiques suivantes reliées aux 3 axes principaux du colloque, décrits ci-dessous :

- Présentations de projets scientifiques mobilisant les corpus web et les humanités numériques
- Présentations de retours d'expérience de projets de recherche utilisant le web comme source : obstacles rencontrés, "success stories"

- Réflexions autour des pratiques professionnelles et académiques impliquant la collecte et la préservation de données issues du web
- Présentation d'expérimentations et de dispositifs favorisant l'accès aux archives numériques et aux corpus web
- Réflexions méthodologiques et épistémologiques sur les besoins d'accès à des données en ligne dans différentes disciplines et de préservation de ces données.

Les 3 axes principaux du projet étaient :

- Axe 1 : Le web à l'intersection de la mémoire et du savoir : enjeux épistémologiques
- Axe 2 : Politiques, pratiques et techniques archivistiques et archives web : du document aux corpus
- Axe 3 : Relations entre dispositif technique et données scientifiques : l'archive web en réseau

Le comité scientifique du colloque était composé de :

- Eléonore Alquier (INA, Dir. Adj. Data et technologies)
- Olivier Baude (Université Paris Nanterre, Modyco, Pr., Dir. TGIR Huma-Num)
- Emmanuelle Bermès (Maîtresse de conférence en ingénierie de la donnée et du document, Ecole nationale des chartes)
- Niels Brügger (Pr. Aarhus University, Pr., Head of WARCNet)
- Dominique Cardon (Sciences Po, Pr., Dir. scientifique du medialab)
- Marie Cornu (ISP, Dir. Recherche CNRS)
- Laurence Favier (GERiiCO, Université de Lille, Pr., Dir. Département de Sciences de l'information et du document)
- Madeleine Géroudet (SCD, Université de Lille, Rsp. du Département Services à la recherche et aux chercheurs)
- Abigail Grotke (Library of Congress, Ass. Head, Digital Content Management Section, IIPC's chair)
- Ian Milligan (University of Waterloo, Department of History, AP, Unleashed Archives Project)
- Laurent Romary (INRIA de Paris, Dir. de recherche, Dir. Culture)
- Philippe Useille (Univ. Polytechnique Hauts-de-France, Institut Sociétés et Humanités - ISH / Laboratoire de Recherche Sociétés & Humanités - LaRSH, MCF, responsable scientifique du pôle Humanités Numériques, MESHS Lille-Nord de France)

Une bibliographie sélective préparée par Odile Bracaval-Demarque (SCD, Université de Lille) pour réaliser une exposition documentaire lors du colloque est disponible via ce lien :

https://lilliad.univ-lille.fr/bibliographies/bibliographies/le-web-source-et-archive-bibliographie-selective

Le programme complet du colloque est en annexe 6 et est disponible sur le carnet Hypothèses : https://ResPaDon.hypotheses.org/files/2023/03/colloque_web_source_archive_04-2023.pdf

2. Bilan des work packages

WP1 : Enjeux stratégiques et préconisations

Pilotage du groupe : Emmanuelle Bermès (BnF) et Marie-Madeleine Géroudet (Université de Lille)

Participants : Dorothée Benhamou-Suesser (BnF), Amélia Laurenceau (Campus Condorcet), Arnaud Laborderie (BnF), Eleonora Moiraghi (SciencesPo)

Les initiatives menées dans le cadre du WP1 ont été structurantes pour l'ensemble du projet. Le WP1 a guidé la mise en œuvre d'une démarche stratégique et l'élaboration des préconisations finales du projet. Il s'est basé sur un groupe de réflexion qui s'est appuyé sur les analyses et expérimentations pour définir des préconisations opérationnelles pour la construction d'un réseau autour de l'accès aux archives du web.

Le groupe de réflexion était composé :

- De membres permanents représentant les partenaires du projet et des représentants des organisations suivantes : GIS CollEx-Persée, réseau des BDLI (représenté par la BNUS), INA, ADBU ;
- D'experts extérieurs, sollicités à titre ponctuel au titre de leur expertise et/ou de leur expérience sur les archives du web : chercheurs, ingénieurs, experts juridiques, organismes de formation, professionnels de l'IST...

Chaque atelier était coordonné par un ou plusieurs membres de l'équipe projet. Au total, 8 ateliers ont eu lieu sur une durée de 2 ans. Chaque atelier était composé d'un temps de présentations et d'échanges, puis d'un atelier destiné à la production de préconisations.

- 1er octobre 2021 « Accéder à un service tiers ou distant dans l'emprise d'un établissement » (Bibliothèque François Mitterrand)
- 26 novembre 2021 « Comprendre les archives du web : enjeux méthodologiques de la production à l'accès aux corpus » (Bibliothèque Richelieu).
- 10 juin 2022 « Enjeux juridiques liés à l'accès et l'exploitation des archives du web »
- 13 mai 2022 : « Opérateurs nationaux, acteurs de proximité : quelle complémentarité dans l'offre de services ? Quels rôles dans la relation aux chercheurs ? »
- 12 juillet 2022 : « Des usages aux services. A partir de l'enquête réalisée sur les usages, quelle offre de services définir ? »
- 6 octobre 2022 : « Formation et compétences pour les différents types d'acteurs : pratiques pédagogiques et didactique »
- 8 décembre 2022 : « Développer la co-construction des collectes »
- 26 janvier 2023 : « Développer de nouveaux usages : Bilan et perspectives »

Ces huit ateliers ont donné lieu à des restitutions publiques systématiques sous forme de webinaires qui ont réuni en moyenne une cinquantaine de participants en visio-conférence. Après une présentation synthétique d'une vingtaine de minutes, le webinaire animé par les coordonnateurs de l'atelier comportait un temps d'échanges avec les participants.

Au fur et à mesure des rencontres trimestrielles, les travaux du groupe de réflexion ont nourri les expérimentations conduites dans les autres work packages et ont permis d'élaborer les préconisations globales du projet et de sa suite.

Les webinaires sont accessibles depuis la page dédiée du carnet Hypothèses : https://ResPaDon.hypotheses.org/category/evenements/cycle-ateliers

Livrable :

Le livrable *Améliorer l'accès et l'exploitation des archives du web par les chercheurs : les 15 préconisations du projet ResPaDon* est en annexe 1.

WP2 : Étude des usages

Pilotage du groupe : Laurence Favier (laboratoire GERiiCO)

Participants : Antoine Henry, Ismaël Timimi, Joanna Casenave, Widad Mustafa El Hadi (GERiiCO), Marie Cros (Université de Lille), Alexandre Faye, Irène Bastard (BnF) Amélia Laurenceau (Campus Condorcet)

Le WP2 a cherché à identifier, caractériser et analyser les usages des archives du web par les chercheurs pour éclairer les services que les professionnels des bibliothèques pourraient offrir en la matière. Plus largement, il a aussi travaillé à saisir l'évolution des pratiques scientifiques (dont l'accès aux sources du Web fait partie) en relation avec celle des services des bibliothèques.

L'étude s'est employée à identifier et caractériser le type de source web qui intéresse les chercheurs, les modes de collecte qu'ils engagent, de constitution de corpus qu'ils construisent à partir de ces sources et leur traitement à des fins scientifiques. Les besoins peuvent aller de la simple consultation d'archives du web, l'extraction et/ou le traitement de corpus avec l'utilisation de méthodes et logiciels spécifiques à l'utilisation de corpus à visée d'analyse comparative ou comme objet d'expérimentation de technologies innovantes.

Dans cet objectif, le groupe de travail a recueilli et analysé plusieurs types de données :

- Les pratiques des chercheurs qui collectent des données issues du web, qu'il soit vivant ou archivé, ont été étudiées notamment via une campagne de huit entretiens qualitatifs.

- Une expérimentation pédagogique a été menée avec des étudiants en master 1 qui ont effectué des recherches sur les archives du web.

- Une analyse rétrospective de 20 projets de chercheurs basés sur les archives du web et accompagnés par les services de la BnF entre 2011 et 2021 a été menée. Cette étude a permis d'établir une typologie multidimensionnelle permettant de caractériser ces projets.

Livrable :

Le livrable *Rapport final du groupe de travail : Analyse des usages des archives du web dans le cadre du projet ResPaDon* est en annexe 2.

WP3 : expérimentation d'une capsule d'accès à distance

Pilotage du groupe : Sara Aubry (BnF) et Marie Cros (Université de Lille)

Participants : Dorothée Benhamou-Suesser, Arnaud Laborderie, Antoine de Sacy, Jean-Philippe Moreux, David Benoist, Marie Carlin, Adoté Chillloh, Lionel Micault (BnF), Jennifer Morival (Université de Lille)

L'expérimentation prévue dans le WP3 visait à tester la mise en œuvre d'une capsule sécurisée permettant d'explorer et de fouiller les archives du web depuis les emprises du service commun de documentation de l'Université de Lille. Les échanges au sein du groupe de travail ont permis de mener une réflexion organisationnelle, technique et juridique permettant l'implantation de ce dispositif.

Une convention d'application juridique entre la BnF et l'Université de Lille a été mise au point afin de garantir la sécurisation des corpus d'archives web mis à disposition des équipes de recherche dans les capsules.

Sur la base de briques de solutions existantes et en tenant compte des besoins spécifiques exprimés au sein du WP3, les équipes de la BnF ont développé et consolidé les solutions techniques permettant d'installer des dispositifs d'accès distant aux archives du web à l'Université de Lille. (lien dsi lille ?) En plus d'un accès semblable au dispositif de consultation existant en DBLI, un corpus spécifique a été constitué autour de la thématique des élections présidentielles et législatives de 2002, auquel a été adjoint un ensemble d'outils de fouille et de visualisation de données. L'interface développée par la BnF regroupe et adapte différents outils développés par la communauté internationale des archives du web (SolrWayback, Archives Unleashed Toolkit, adaptation de Jupyter Notebooks) en un même parcours usager, avec pour objectif de donner un aperçu de différentes approches méthodologiques et outils permettant de travailler sur un corpus d'archives web.

En parallèle, les échanges au sein du WP3 entre la BnF et l'Université de Lille ont permis de définir, de qualifier et de prototyper les ressources documentaires et les outils de médiation

nécessaires à la compréhension et à l'utilisation des archives et de la capsule par les chercheurs. (Cette démarche s'est faite en articulation avec les opérations menées dans le cadre du BnF-DataLab, et également en lien avec le travail des correspondants à la BnF qui documentent les collections.) Des « médiateurs archives du web » ont été identifiés au SCD de Lille afin d'accompagner les chercheurs lors de leurs tests. Des formations dédiées ont été créées par les équipes de la BnF pour former ces médiateurs.

Les équipes de l'Université de Lille ont organisé les modalités d'installation de la capsule et d'accueil des chercheurs, qui ont été conviés à venir explorer les corpus et outils mis à leur disposition. Des parcours de test, développés par le WP3, leur ont été proposés, pour explorer dans une première phase les archives du web, puis dans une seconde phase pour découvrir la collection élections 2002 de manière plus approfondie avec le recours aux outils de fouille et de visualisation de données. Le parcours de test comprenait également une phase d'évaluation qui prenait la forme d'un court entretien mené par le médiateur à la fin de chaque rendez-vous. Des sessions de test, notamment avec des étudiants de master ont également eu lieu.

Les travaux conduits dans le cadre du WP3 ont permis d'expérimenter la configuration et l'installation d'un dispositif d'accès à distance aux collections des archives du web, et de mettre à l'épreuve en conditions réelles toutes les dimensions d'une telle installation. Le livrable du WP3 dresse un bilan complet de l'expérimentation et propose une série de recommandations dans l'optique de consolider, développer ce dispositif.

Livrable :

Le livrable *L'expérimentation capsule au cœur du projet ResPaDon : bilan et recommandations* est en annexe 3.

WP4 : Web vivant

Pilotage du groupe : Eleonora Moiraghi (DRIS SciencesPo)

Participants : Antoine de Sacy, Marie Carlin, Sara Aubry (BnF), Audrey Baneyx, Benjamin Ooghe (médialab, SciencesPo)

L'expérimentation prévue dans le WP4 visait à tester de nouvelles manières d'explorer les archives du web en utilisant les outils développés par le médialab de SciencesPo pour répondre à des questions de recherche portant sur une thématique transverse au web vivant et au web archivé.

En 2021, des études, tests et évolutions ont été réalisés sur Hyphe, un logiciel de constitution et de curation de corpus web, et l'application Archives de l'internet qui permet de consulter les archives du web de la BnF, pour rendre possible l'utilisation de Hyphe sur le web archivé (voir documentation technique). Ces évolutions permettent également l'utilisation de Hyphe avec la Wayback Machine d'Internet Archive.

L'utilisation de Hyphe sur les archives du web a d'abord été testée lors d'un mini-Datasprint qui a été organisé le 18 novembre 2021. Il préfigurait un événement de plus grande ampleur, le Datasprint ResPaDon, qui s'est déroulé pendant cinq jours du 4 au 8 avril 2022 dans les espaces BnF Datalab situés en bibliothèque de Recherche de la BnF.

Cet événement a donné l'opportunité à quatre équipes pluridisciplinaires, associant chercheurs, ingénieurs, designers et professionnels de l'information scientifique et technique, de tester de nouvelles pratiques et d'explorer de nouvelles pistes méthodologiques. Les quatre expérimentations menées et leurs résultats sont décrits sur un site web dédié qui a été développé à la suite de l'événement. Les travaux de groupe ont permis également de dégager quelques réflexions générales d'ordre méthodologique pour étudier de manière complémentaire les archives du web et le web vivant à des fins de recherche. Ces réflexions font l'objet d'un document librement accessible sur le même site web.

Livrables :

- Site web « Explorer les archives du web avec Hyphe » (retour d'expérience et travaux du datasprint): https://ResPaDon.medialab.sciencespo.fr/
- Réflexions méthodologiques pour étudier de manière complémentaire les archives du web et le web vivant à des fins de recherche : https://drive.google.com/file/d/11jfRnaZFum0eRPetxLb8G0JwuoOBFtps/view
- Logiciel libre Hyphe et ses évolutions pour fonctionner sur le web archivé : https://github.com/medialab/hyphe/issues/372

WP5 : Coordination du projet, planification et communication

Coordination du projet : Céline Ferjoux (Université de Lille)

Le rôle du projet ResPaDon dans la constitution d'un réseau autour de l'accès aux archives du web implique la définition d'une stratégie de communication, qui permette aux chercheurs et aux acteurs de l'information scientifique et technique de recevoir des informations régulières sur le projet, son impact et ses suites. Cette stratégie se compose de :

- La mise en œuvre d'événements ouverts à l'ensemble de la communauté : journées de lancement, webinaires de restitution des ateliers ; journée professionnelle ; colloque
- Une démarche de communication sur le web et les réseaux sociaux ; création et alimentation du carnet hypothèses, création du compte twitter ResPaDon_Projet, création d'une liste de diffusion sur Renater
- La présentation des résultats du projet dans des événements et des publications scientifiques.

Démarche de communication sur le web et les réseaux sociaux

Dès avril 2021, une identité visuelle a été définie et les premiers canaux de communication ont été créés : compte Twitter @ResPaDon_Projet et liste de diffusion ResPaDon@groupes.renater.fr (10 newsletters, 128 abonnés en 2023).

Ces premiers canaux ont été complétés d'un carnet de recherches sur Hypotheses.org, lieu privilégié pour la diffusion de l'ensemble des initiatives du projet. Le Carnet Hypothèses a été créé le 6 juillet 2021 et inauguré avec la publication des enregistrements vidéo de la journée de lancement, le 31 août. S'y trouvent les avancées du projet, les résultats, les captations des webinaires, la captation vidéo des événements principaux.

Sur Twitter, l'audience du projet bénéficie d'une visibilité étendue à la communauté scientifique nationale et internationale. Le compte @ResPaDon_Projet dénombre 682 abonnés en 2023.

Communications scientifiques

Le projet ResPaDon a donné lieu à un certain nombre de communications scientifiques.

➢ Communications dans des colloques et manifestations scientifiques

- "Building a network of partners to develop the use of web archives : the ResPaDon project" in Colloque "Un patrimoine pour l'avenir, une science pour le patrimoine", 15, 16 mars 2022, Fondation des Sciences du Patrimoine, https://ResPaDon.hypotheses.org/578

- Benhamou-Susser, Dorothée, Cros, Marie. « La plongée au coeur du web porte un nom : ResPaDon ». Amue, la collection numérique, 20 avril 2022. https://www.amue.fr/fileadmin/amue/systeme-information/documents-publications/la-collection-numerique/amue-collection-numerique_20.pdf

- Benhamou-Suesser, Dorothée, Morival, Jennifer. « ResPaDon : Expanding Research Use of French Web Archives », IIPC Web Archiving Conference 2022: Session 10: Researching Web Archives: Tools and Access. Conférence en ligne hébergée par la Library of Congress, mai 2022. https://www.youtube.com/watch?v=bTt9H3S2qXg.

- Baneyx, Audrey, Benhamou-Suesser, Dorothée, Moiraghi, Eleonora. « 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 » », Colloque Humanistica 2022. Montréal, mai 2022. https://sciencespo.hal.science/hal-03688620.

- Bermès, Emmanuelle, Favier, Laurence, Géroudet, Marie-Madeleine. « Collaborer entre chercheurs et bibliothécaires autour des archives du web : le projet ResPaDon", in Colloque "Travailler en humanités numériques : collaborations, complémentarités et tensions", juin 2022, MESHS Lille Nord https://publi.meshs.fr/page/dhnord2022...dhnord.---3

- Faye Alexandre. « Faciliter les usages des archives du Web. Enjeux de l'accompagnement sur des données culturelles », Rencontres 2022 de la donnée culturelle : panorama et cas d'usage. Paris, INHA, décembre 2022. https://youtu.be/zvwWoDmmpE0v

- Aubry, Sara, Benhamou-Suesser, Dorothée, Morival, Jennifer. « Developing New Academic Uses of Web Archives Collections : Challenges and Lessons Learned from the experimental Service deployed at the University of Lille during the ResPaDon Project », IIPC Web Archiving Conference 2023. Hilversum, mai 2023. https://www.youtube.com/watch?v=gw5h-QmmBiM.

- Aubry, Sara, Baneyx, Audrey, Bermès Emmanuelle, Favier Laurence, Faye, Alexandre, Géroudet, Marie-Madeleine, Ooghe Tabanou, Benjamin. « A network to develop the use of web archives: three outcomes of the ResPaDon project ». In RESAW 2023 - Exploring the Archived Web during a Highly Transformative Age. Marseille, 2023. https://resaw2023.sciencesconf.org/434920.

➢ Communications dans le cadre du projet

- Journée scientifique d'ouverture, : « Faire réseau autour des archives du web, usages et opportunités », 17 mai 2021. https://ResPaDon.hypotheses.org/1
    - Holownia, Olga, Tuleu, Benoît. "Les archives du web : perspectives croisées pour un état des lieux".
    - Bonah, Christian. "Créer des archives web avec la BnF et l'INA dans le cadre d'un projet de recherche : l'exemple de Bodycapital"
    - Greffet, Fabienne. "De la démocratie en numérique. L'analyse des activités affiliées aux candidats à l'élection présidentielle française de 2002 à 2017 à partir des archives du web"
    - Cartier, Emmanuel. "Détecter et suivre les évolutions lexicales dans les archives du web : le projet Néonaute"
    - Gebeil, Sophie. "Les archives du Web comme source pour une histoire des pratiques mémorielles depuis les années 2000"
    - Brugger, Niels. "Studying web archives across borders: the case of the WARCnet network"
    - Beaudouin, Valérie, Cote, Christian, Genin, Christine. "Littératures numériques et expression de soi durant le confinement"
    - Barnabé, Fanny, Montembault, Hugo, Alvarez, Julian, Dor, Simon. "Quelle place pour le web dans les études sur le jeu vidéo ?"
    - Conduire des recherches à partir du web vivant ou comment faire face à la volatilité des contenus sur le web :
        - Venturini, Tommaso. "D'un Web des documents à un Web de flux : l'oralité secondaire des médias numériques"
        - Ooghe-Tabanou, Benjamin. "Méthodes, outils et limites pour collecter et analyser le web et les réseaux sociaux"
        - Levrier, Guillaume. "Des communautés militantes en ligne qui disparaissent : cartographier le web mourant"
        - Blanchard, Gersende. "Retour d'expériences de recherches menées sur le web vivant: quels défis et quels enseignements."

- Bermès, Emmanuelle, Cardon, Dominique, Groudiev, Stéphanie, Schafer, Valérie. “Comment faire réseau autour des archives du web ?”

- Journée de restitution Datasprint du 10 février 2023

  - Mazoyer, Béatrice, Plique, Guillaume, De Sacy, Antoine, Tosetto, Cristina, Wiatrowski, Clara. « Cartographie de la critique en ligne des arts du spectacle ».

  - Bellony, Leslie, Brioudes, Guillaume, Degrange, Isabelle, Faye, Alexandre, Jacomy, Alexis, Locoh-Donou, Kevin, Sala, Caroline. « Crise de la COVID 19 : positionnement des acteurs du web par rapport aux institutions »

  - Sara Aubry, Sara, Greffet, Fabienne, Heude, Cyril, De Mourat, Robin, Ooghe-Tabanou, Benjamin. « Structuration des communautés politiques autour des candidats aux élections présidentielles (exemple du candidat Jean-Luc Mélenchon) »

  - Benhamou-Suesser, Dorothée, Girard, Paul, Levrier, Guillaume, Morival, Jennifer, Pehlivan, Zeynep. « La notion de « génome » dans les archives électorales BnF »

- **Webinaire ResPaDon : « Collectes et usages scientifiques du web électoral », 25 novembre 2022.**

  - Benhamou-Suesser, Dorothée, Géroudet, Marie-Madeleine. « Présentation de la capsule d’accès à distance des Bibliothèques de l’Université de Lille ».

  - Crinière-Boizet, Anaïs, « 2002-2022 : 20 ans de collectes du web électoral »

  - Copin, Isabelle, Duffes, Laurence, Soulé-Sandic, Catherine. « Sélectionner des sites web à archiver : une pratique collaborative » (Table ronde)

  - Copin, Isabelle, Barbedet, Mathilde. « Valoriser les archives à travers des « parcours guidés » ».

  - Greffet, Fabienne. « Déléguer la communication politique. Les internautes, porte-parole des candidats à l’élection présidentielle, 2002-2017. »

  - Levrier, Guillaume, « Caractériser les représentations politiques du vivant « génomique » dans les archives du web français ».

- **Journée d'étude : « Faire réseau autour des archives du web : Perspectives du projet ResPaDon » – 13 mars 2023, BnF, Paris**

    - Désos-Warnier, Catherine. "Faire réseau autour des archives du web : bilan et perspectives du projet ResPaDon – chronique de la JE du 13 mars à la BNF". DLIS. https://dlis.hypotheses.org/6345

    - Nyffenegger, Isabelle, Colas, Alain, Roche, Julien. « Conférence d'ouverture »

    - Bermès, Emmanuelle, Géroudet, Marie-Madeleine, « Le projet ResPaDon : retour sur l'origine et les coulisses d'un projet d'envergure nationale »

    - Tybin, Vladimir, Benhamou-Suesser, Dorothée. « Les archives du web : présentation et enjeux pour la recherche »

    - Baneyx, Audrey, Moiraghi, Eleonora, Greffet, Fabienne, « Le vivant et les archives : l'expérience du DataSprint ResPaDon »

    - Aubry, Sara, Cros, Marie. « Embarquez dans la capsule : retour sur l'expérimentation d'un accès distant »

    - Bermès, Emmanuelle, Géroudet, Marie-Madeleine, « Les préconisations du projet »

    - Gebeil, Sophie, Groudiev, Stéphanie, Miura, Grégory, Mussou, Claude, Tuleu, Benoît, « ResPaDon : et la suite ? » (Table-ronde)

    - Schafer, Valérie. « Continuer à faire réseau : le regard d'une chercheuse sur le projet ResPaDon »

- **Colloque international : « Le web : source et archive », 3, 4, 5 avril 2023, Université de Lille.**

    3 avril 2023 : Relations entre dispositif technique et données scientifiques : l'archive web en réseau

    - Colot, Olivier, Nyffenegger, Isabelle, Favier, Laurence, Bermès Emmanuelle, Géroudet, Marie-Madeleine. « Discours d'ouverture »

    - Milligan, Ian. « Interpreting the web : the critical role of historical context in web archival research »

    - Cote, Christian. « Méthodologie pour l'élaboration d'un corpus et d'une archive du web littéraire francophone »

    - Bouillard, Léna, Jaworska-Kaska, Alicja. « Les sources de l'étude du web militant » (Table ronde)

    - Tosetto, Cristina. « Cartographie de la critique en ligne dans les arts du spectacle : entre approche synchronique et diachronique »

- Shen, Shiming, Kergosien, Eric, Treleani, Matteo. « Du corpus télévisuel au corpus web à l'aide de l'outil visuel automatique : méthodes du projet CROBORA »

4 avril 2023 : Le web à l'intersection de la mémoire et du savoir : enjeux épistémologiques

- Muller, Caroline, Clavert, Frédéric. « Le goût de l'archive numérique et les archives du web ».
- Bonah, Christian, Lellinger, Solène, Sala Caroline. « Des émissions culinaires aux vidéos conseils : transformation des recettes audiovisuelles de la télévision au web et la question du faire à manger « digital ».
- Gebeil, Sophie. « Web vivant et web archivé, aux sources de l'histoire nativement numérique ».
- Favier, Laurence, Henry, Antoine, Bastard, Irène, Faye Alexandre. « Etudier les usages des archives du web. »
- Bert-Erboul, Clément, Clémencin, Grégoire, Finez, Jean, Dahmani, Amira, Demars, Brice, Macaud, Amélie. « Recherches et méthodes mobilisant les archives du web » (Table ronde).
- Casenave, Joana, Favier, Laurence. « Exploration des archives du Web par un public étudiant: contribution à l'analyse critique des sources ».
- Gebeil, Sophie, Miura, Grégory. « Enjeux épistémologiques et didactiques des sources web » (Table ronde)

5 avril 2023 : Politiques, pratiques et techniques archivistiques et archives web : du document aux corpus

- Bachimont, Bruno. « Entre inscription éphémère et donnée pérenne : peut-on archiver le Web au-delà de son enregistrement ? Quelques remarques méthodologiques et critiques ».
- Benhamou-Suesser, Dorothée, Cros, Marie, Hadjimanolis, Gwladys, "Explorer les archives de l'internet à l'Université de Lille : regards croisés sur un dispositif expérimental au service des chercheurs »
- Castex, Lucien, Mallet-Poujol, Nathalie, "Question(s) de droit(s)" (Table ronde)
- Monjour, Servanne, Sauret, Nicolas, « Archiver le web littéraire. Défis méthodologiques et conceptuels »
- Gabrysiak, Louis, Gensburger, Sarah, Severo, Marta. « Collectes du confinement et archives du web : exploration croisée des archives de BNF et de l'INA »
- Faye, Alexandre, Pailler, Fred, Aubry, Sara, Silvestre de Sacy, Antoine, Schafer, Valérie, « Harlem Shake à la BnF … À la recherche d'un phénomène viral dans les archives du Web ».

- Bermès, Emmanuelle, Géroudet, Marie-Madeleine. « Le cycle d'ateliers ResPaDon : bilan et préconisations »

A paraître en avril 2024 : un numéro spécial des *Cahiers du Numérique* consacré au colloque ResPaDon « Le web : source et archive » dont l'appel à publication court du 13 novembre 2023 au 15 janvier 2024 :
https://lcn.revuesonline.com/revues/23/LCN_le_web_source_et_archive.pdf

ANNEXES

Les livrables du projet

Annexe 1 : *Améliorer l'accès et l'exploitation des archives du web par les chercheurs : les 15 préconisations du projet ResPaDon*

Annexe 2 : *Rapport final du groupe de travail : Analyse des usages des archives du web dans le cadre du projet ResPaDon*

Annexe 3 : *L'expérimentation capsule au cœur du projet ResPaDon : bilan et recommandations*

Annexe 4 : Le livrable WP4

- Site web « Explorer les archives du web avec Hyphe » (retour d'expérience et travaux du datasprint): https://ResPaDon.medialab.sciencespo.fr/
- Réflexions méthodologiques pour étudier de manière complémentaire les archives du web et le web vivant à des fins de recherche : https://drive.google.com/file/d/11jfRnaZFum0eRPetxLb8G0JwuoOBFtps/view
- Logiciel libre Hyphe et ses évolutions pour fonctionner sur le web archivé : https://github.com/medialab/hyphe/issues/372

Les programmes des événements conclusifs

Annexe 5 : Programme de la journée d'étude : « Faire réseau autour des archives du web : Perspectives du projet ResPaDon » – 13 mars 2023, BnF, Paris

https://respadon.hypotheses.org/files/2023/02/programme13mars_web.pdf

Annexe 6 : Programme du colloque international : « Le web : source et archive », 3, 4, 5 avril 2023, Université de Lille

https://respadon.hypotheses.org/files/2023/03/colloque_web_source_archive_04-2023.pdf

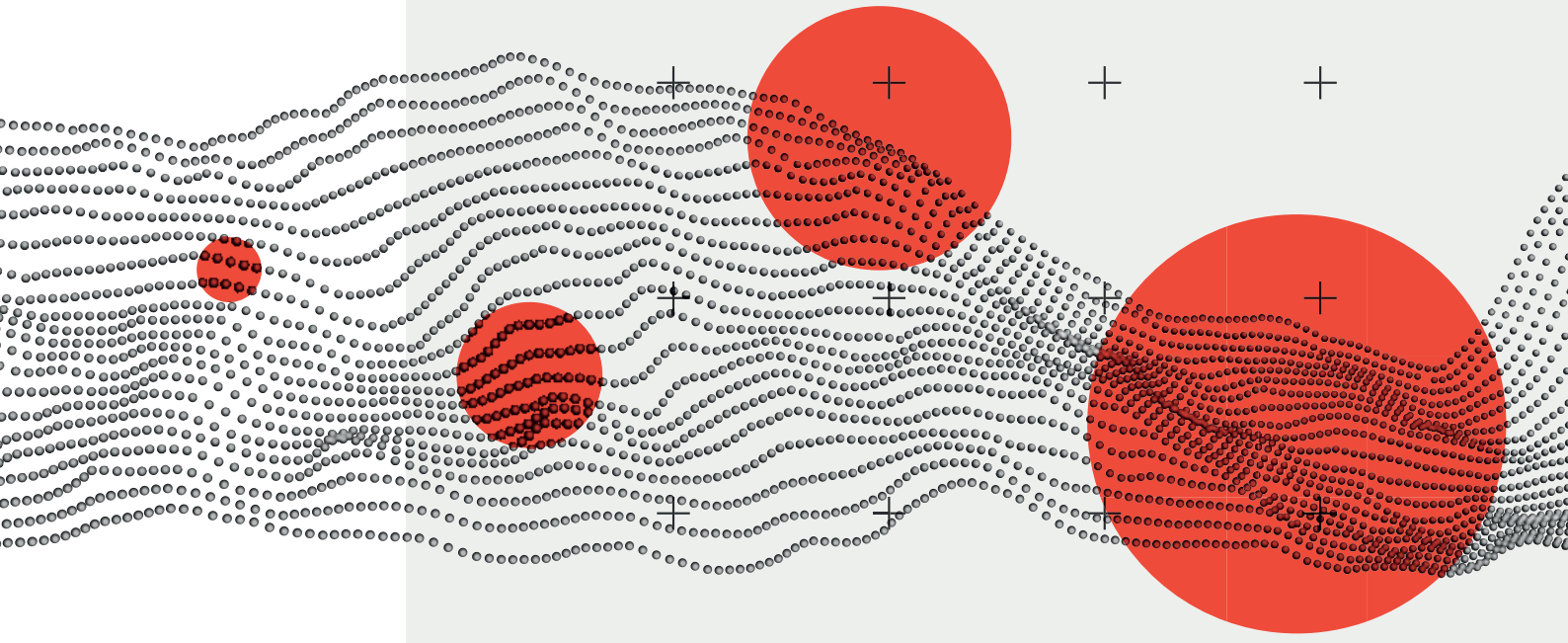
Annexe 1 :

Améliorer l'accès et l'exploitation des archives du web par les chercheurs : les 15 préconisations du projet ResPaDon

AMÉLIORER L'EXPLOITATION DES ARCHIVES DU WEB PAR LES COMMUNAUTÉS DE RECHERCHE

Les 15 préconisations du projet ResPaDon

RES
PA
DON



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

De 2021 à 2023, le projet ResPaDon, Réseau de Partenaires pour l'exploitation et l'analyse de Données numériques a réuni chercheurs et professionnels de l'information autour de l'amélioration de l'accès et de l'exploitation des archives du web conservées à la Bibliothèque nationale de France (BnF) et à l'Institut National de l'Audiovisuel (INA). Un cycle d'ateliers a été mis en œuvre sur toute la durée du projet dans l'objectif d'établir un ensemble de préconisations à proposer aux producteurs et aux utilisateurs des archives.

Ces préconisations se focalisent sur le développement des usages des archives du web par les communautés de recherche. Elles abordent les archives du web par le prisme de leur usage et de leur potentiel d'exploitation. Elles constituent un ensemble de mesures idéales, qu'il s'agit de lire en tenant compte du statut juridique actuel des collections constituées par dépôt légal. Elles représentent un futur possible, dont la réalité dépendra des moyens effectivement disponibles et de la mobilisation de l'ensemble de la chaîne des acteurs, des producteurs du web aux communautés de recherche.

Ces préconisations sont composées de 5 grands principes et de 15 actions.

En raison de la nature particulière du web, l'étude scientifique des contenus qui y circulent implique la fabrication d'une archive

Les archives du web ont vocation à constituer une source de la recherche parmi d'autres

Les publics doivent pouvoir être autonomes dans l'exploitation des archives du web

Un réseau national associant chercheurs et bibliothécaires constitue un catalyseur essentiel pour développer l'exploitation des sources web

La médiation des sources web par des acteurs pluriels implique le développement de nouvelles compétences

PRINCIPE 1.
En raison de la nature particulière du web, l'étude scientifique des contenus qui y circulent implique la fabrication d'une archive.

1. Soutenir la définition et la transmission des méthodes d'études des sources web au service de la recherche.

Qui ? Enseignants-chercheurs et professions en soutien, réseaux internationaux (RESAW, WARCnet)

Quand ? Processus déjà engagé, à poursuivre et à accompagner

Les enseignants-chercheurs engagés dans les réseaux (RESAW, WARCnet...) développent et partagent des méthodologies d'études des sources web. Il est important de soutenir le développement de méthodes et des concepts épistémologiques associés, de renforcer l'émergence d'une communauté de pratiques en matière d'enseignement et d'intégrer les méthodes d'études des sources web dans les programmes de master. La collaboration internationale est une clé pour la réussite de cette action.

2. Normaliser la méthodologie de création, de documentation et de citation d'une archive web.

Qui ? Consortium IIPC, TC 46 «Information et documentation» de l'ISO

Quand ? Long terme

La conception d'une norme de haut niveau définissant les caractéristiques d'une archive web et les modalités de leur documentation et de leur citation doit faciliter la compréhension, l'exploitation et la réutilisation des sources web par les chercheurs. Elle crée un terreau commun pour des méthodologies partageables.

3. Intégrer les sources web dans le développement de la culture et de la littératie numérique des étudiants, chercheurs et professionnels.

Qui ? Opérateurs de formation initiale et continue, acteurs de la formation doctorale

Quand ? Moyen terme

La connaissance des sources web, parmi lesquelles les archives collectées par la BnF et l'INA, doit s'inscrire dans la culture numérique acquise par

les étudiants, les chercheurs et les professionnels. Il s'agit d'inscrire ces sources dans les démarches existantes pour le développement de la littératie numérique dans l'enseignement supérieur et la recherche.

PRINCIPE 2.
Les archives du web ont vocation à constituer une source de la recherche parmi d'autres.

4. Inscrire les sources web, archives et web vivant, dans l'évolution des pratiques de recherche et dans l'ouverture des processus et résultats de la recherche.

Qui ? BnF, INA, communauté de l'enseignement supérieur et de la recherche (ESR), dont acteurs de la science ouverte

Quand ? Moyen terme

L'ouverture d'un dialogue entre producteurs des sources web, acteurs de l'archivage web et acteurs de l'ouverture de la science doit permettre l'identification de pistes d'actions communes pour faciliter l'exploitation et la réutilisation des sources web.

5. Favoriser l'exposition des métadonnées et la mise en place de mécanismes de découvrabilité des archives du web.

Qui ? BnF, INA

Quand ? Moyen terme

La découverte des archives du web doit être facilitée par des dispositifs techniques. L'ouverture des métadonnées des archives du web doit permettre d'effectuer directement sur le web une recherche sur les URL archivés sans accès aux contenus des sites.

6. Permettre la découverte des archives du web à partir d'une navigation ou d'une exploration du web vivant.

Qui ? BnF, INA, médialab de Sciences Po

 Quand ? Moyen terme

Le renforcement du lien entre le web vivant et les archives du web facilite la découverte des archives du web et la mise en œuvre de démarches multi-sources. Avec cette action, il s'agit de rendre possible la vérification de la présence d'un site dans les archives du web depuis le web vivant et de pérenniser la connexion du logiciel Hyphe, développé et maintenu par le médialab de Sciences Po, aux archives du web.

PRINCIPE 3.
Les publics doivent pouvoir être autonomes dans l'exploitation des archives du web.

7. Faciliter l'accès et la réutilisation des archives du web en faisant évoluer les conditions règlementaires actuelles.

Qui ? Ministère de la Culture

Quand ? Court terme. Cette action représente un prérequis.

L'évolution des conditions réglementaires actuelles doit faciliter l'exploitation des archives du web à des fins de recherche. A minima, cette évolution réglementaire doit rendre possible le déploiement de points d'accès aux archives du web dans tout établissement de l'enseignement supérieur et de la recherche et de faciliter la réutilisation des contenus de l'archive web à des fins de recherche.

8. Déployer et pérenniser des capsules d'accès aux archives du web dans des établissements de l'enseignement supérieur et de la recherche.

Qui ? BnF, Directions des systèmes d'informations (DSI) et structures documentaires des établissements engagés

Quand ? Moyen terme. L'investissement technique initial constitue un prérequis.

Le déploiement et la pérennisation d'un dispositif de capsules d'accès aux archives du web dans les établissements de l'ESR implique un investissement technique initial pour l'amélioration du dispositif. Une collaboration entre la BnF et les DSI des établissements de l'ESR doit permettre de définir les conditions techniques nécessaires pour une installation et une maintenance facilitées du dispositif de capsule.

9. Faciliter l'enrichissement collectif d'un bac à sable à usage pédagogique et de recherche en accès ouvert.

Qui ? BnF, communauté de l'enseignement supérieur et de la recherche

Quand ? Moyen terme

L'ouverture d'un corpus à des fins pédagogiques et de recherche doit faciliter la prise en main des archives du web par des chercheurs et des étudiants. Elle implique de développer une logique de clearance juridique pour rendre accessible directement sur le web une collection cohérente. Après l'ouverture d'un bac à sable par la BnF, les acteurs du réseau devront pouvoir contribuer à l'enrichissement de l'outil.

PRINCIPE 4.
Un réseau national associant chercheurs et professionnels de l'information et des bibliothèques constitue un catalyseur essentiel pour développer l'exploitation des sources web.

10. Réunir autour de nœuds bien identifiés les acteurs intéressés à l'exploitation de la source web et à la production de son archive : laboratoires de recherche, bibliothèques universitaires et structures documentaires, MSH, bibliothèques municipales…

Qui ? Bibliothèques universitaires et structures documentaires, laboratoires de recherche, MSH, bibliothèques municipales, acteurs de la formation

Quand ? Première identification des partenaires intéressés à court terme. Déploiement des nœuds à moyen terme.

Ces nœuds auraient vocation à se fonder sur un territoire ou une thématique. La première étape consiste à définir le périmètre de coopération de ces nœuds et leurs modalités de fonctionnement, avant de les déployer et les faire vivre.

11. Développer au sein de ces nœuds un ensemble d'activités de soutien à l'accès et à l'exploitation des sources web : offre de sensibilisation et de formation, collectes partenariales et accès distant aux archives du web.

Qui ? BnF, INA, bibliothèques universitaires et structures documentaires, laboratoires de recherche, MSH, bibliothèques municipales

Quand ? Moyen terme

Un nœud pourra se construire sur tout ou partie de ces activités, sur la base d'un territoire ou d'une thématique. La présence d'un accès distant aux archives du web dans les établissements contributeurs d'un nœud ne doit pas être une condition pour la participation au réseau. Au contraire, il est privilégié une diversité de modalités de participation.

12. Mettre en œuvre une co-animation du réseau par les nœuds et les acteurs nationaux : documentation mutualisée, rencontres régulières, échanges de pratiques…

Qui ? Etablissements partenaires des nœuds, acteurs nationaux (BnF, INA, Huma-Num, CollEx-Persée)

Quand ? Moyen terme

La compréhension partagée des sources web et de leurs utilisateurs est au cœur de ce réseau. Le principe de co-animation permet de concevoir le réseau comme un lieu de mutualisation et d'échanges réciproques. La connaissance des méthodes de constitution de l'archive et la compréhension des usages et des besoins des chercheurs sont au cœur de ce réseau.

PRINCIPE 5.
La médiation des sources web par des acteurs pluriels implique le développement de nouvelles compétences.

13. Développer les compétences de médiateurs au sein des nœuds du réseau.

Qui ? BnF, INA, acteurs de la formation initiale et continue, acteurs du réseau

Quand ? Moyen terme

Chaque nœud du réseau doit identifier plusieurs médiateurs chargés d'accompagner la découverte des sources web et leur exploitation à des fins de recherche. La construction de ce rôle de médiateur nécessite que la formation au web et à ses archives trouve toute sa place dans les formations initiales et continues des professionnels des bibliothèques et de l'information. En plus de l'acquisition de cette culture commune par tous les professionnels, le développement des compétences des médiateurs requiert la mise en œuvre régulière de sessions spécifiques de formation.

14. Développer les compétences des ingénieurs appelés à accompagner les projets de recherche autour des sources web

Qui ? BnF, INA, acteurs du réseau, acteurs de la donnée, acteurs de la formation initiale et continue

Quand ? Moyen terme

L'accompagnement de projets de recherche autour du web et de ses archives par les ingénieurs nécessite le développement de compétences spécifiques sur les méthodes d'analyse des sources web et sur les outils et logiciels disponibles. Le développement de compétences de niveau 2 par les ingénieurs s'inscrit dans une démarche plus large de développement des métiers de la donnée et de compétences associées à ces métiers. Il s'agit d'accompagner le développement de compétences par des cursus spécifiques de formation.

15. Développer les collectes partenariales associant chercheurs et professionnels de l'information et des bibliothèques, pour favoriser l'acquisition de ces nouvelles compétences.

Qui ? BnF, acteurs du réseau

Quand ? Moyen terme

Les collectes partenariales sont au cœur de la démarche de développement d'une culture commune autour des archives du web. La participation au processus de collecte permet de mieux comprendre la méthode de fabrique de l'archive. La co-construction de collectes entre les producteurs actuels des archives du web et les acteurs du réseau à venir permet ainsi de concevoir des corpus adaptés aux pratiques de recherche et de développer une culture partagée de l'archive web.

CONTACT :
respadon@univ-lille.fr

Annexe 2 :

Rapport final du groupe de travail :

Analyse des usages des archives du web dans le cadre du projet ResPaDon

Analyse des usages des archives du web dans le cadre du projet ResPaDon

Rapport final du groupe de travail

Coordination Laurence FAVIER

Équipe :

Alexandre Faye (Responsable des projets de recherche, Service du dépôt légal numérique, Bibliothèque nationale de France)

Irène Bastard (Chef de projet Public et usages, Délégation à la stratégie et à la recherche)

Amélia Laurenceau (Directrice du Département Soutien à la recherche, Humathèque, Campus Condorcet)

Marie Cros (Chargée de projets "données", Département Services à la recherche et aux chercheurs, Université de Lille, SCD)

Laurence Favier (Professeure en SIC, Université de Lille, GERiiCO Université de Lille)

Joana Casenave (Maître de conférences en SIC, Lab. GERiiCO Université de Lille)

Antoine Henry (Maître de conférences, Université de Lille, GERiiCO Université de Lille)

_____Novembre 2023_____

Table des matières

I. PRESENTATION DES TRAVAUX

Le groupe de travail cherche à établir une étude des usages des archives du web par les chercheurs et, secondairement, par les étudiants, pour établir une typologie de services que les professionnels des bibliothèques pourraient leur offrir. Plus largement, il contribue à saisir l'évolution des pratiques scientifiques (dont l'accès aux sources du web fait partie) en relation avec celle des services des bibliothèques.

Les études menées dans ce groupe s'emploient à identifier et caractériser le type de source web qui intéresse les chercheurs, les modes de collecte qu'ils engagent, de constitution de corpus qu'ils construisent à partir de ces sources et leur traitement à des fins scientifiques ou d'enseignement. Les besoins peuvent aller de la simple consultation d'archives du web, à l'extraction, au traitement de corpus avec l'utilisation de méthodes et logiciels spécifiques à l'utilisation de corpus à visée d'analyse comparative ou comme objet d'expérimentation de technologies innovantes, à l'archivage pérenne des matériaux recueillis.

Dans cet objectif, le groupe de travail recueille et analyse des données visant à saisir différents points de vue sur des cas types d'utilisation d'archives du web. La recherche collecte quatre catégories de données :

- Des données issues de chercheurs et recueillies par entretien
- Des données issues de la BnF émanant des projets auxquels elle a contribué avec les chercheurs porteurs de projets ;
- Un retour d'expérience pédagogique sur l'usage des archives du web en BDLI à Lille, complété par celles de la Wayback Machine et de l'INA
- Des données bibliographiques signalant des projets de recherche traitant d'archives du web

En 2021, la collecte de données de catégorie 2 a permis l'élaboration d'une première typologie de projets reposant sur l'expérience des équipes du Dépôt légal Numérique de la BnF (Irène Bastard et Alexandre Faye).

Parallèlement à l'analyse des usages de chercheurs du point de vue des institutions, une campagne d'entretiens avec des chercheurs (données de catégorie 1) menés par GERiiCO et le Campus Condorcet (Laurence Favier, Antoine Henry, Amélia Laurenceau) relate l'analyse que les chercheurs font de leurs propres pratiques au sein de leurs projets alors qu'ils ne connaissent pas ou n'utilisent pas nécessairement les services du Dépôt Légal Numérique de la BnF.

Une expérience pédagogique a également été menée avec des étudiants de master 1 à qui l'on a demandé un travail impliquant la recherche d'archives du web. Une promotion de Master 1, sous la direction de Joana Casenave et Laurence Favier, a effectué des recherches sur une liste de sujets prédéfinis et a interrogé les ressources électroniques de la bibliothèque municipale de Lille (BDLI) ainsi qu'Internet Archives et l'INAthèque.

Concernant les données bibliographiques, un stage de deux mois et demi, réalisé par une étudiante du Master 1 en Information et documentation de l'Université de Lille sous la direction de Laurence Favier, a été financé à la fin de l'année universitaire 2020-21, afin de contribuer à la recherche bibliographique (recherche de travaux publiés utilisant les archives du web).

II. PRINCIPAUX APPORTS DU GROUPE DE TRAVAIL CONSACRÉ À L'ANALYSE DES USAGES (WP 2)

A Contribution de GERIICO et du Campus Condorcet : restitution d'entretiens, expériences pédagogiques

A.1. Les pratiques des chercheurs : résultats de la campagne d'entretiens (L. Favier, A. Henry, A. Laurenceau)

Huit entretiens ont été menés, dont nous restituons une partie ci-après.

Nous n'avons pas considéré que l'utilisation du seul web français était une condition pour sélectionner nos interlocuteurs, car nous nous intéressons avant tout à leur démarche générale eu égard aux archives du web.

Dans tous les cas, les chercheurs collectent des données issues du web, qu'il soit vivant (pages anciennes demeurant accessibles au moment de la recherche et pages récentes) et/ou « mort » (pages et sites anciens n'étant plus accessibles et nécessitant de consulter des ressources spécifiques, les archives de la BnF, de l'INA, d'Internet Archives, pour y avoir accès). Il est à peu près constant que la recherche ne se limite quasiment jamais au seul web, mais concerne tout l'écosystème numérique en ligne et ouvert : sites et réseaux sociaux.

Le résultat de cette collecte est un corpus et non une sélection issue de fonds d'archives, ce qui modifie l'approche méthodologique. Mais la notion elle-même de *corpus* doit être interrogée, car elle n'est pas la même selon les disciplines scientifiques et l'utilisation des matériaux du web implique de la faire également évoluer. C'est le statut accordé aux contenus rassemblés, leurs liens et leur contextualisation qui sont en question et donc la manière de les organiser (en fonction de la thématique, de la provenance, des dates, des producteurs, etc.). Nous y reviendrons dans le rapport final. L'aspect multidisciplinaire de toutes les recherches menées avec les matériaux issus du web que nous avons pu repérer contribue, en même temps que le statut des matériaux rassemblés, à réinterroger cette méthodologie des corpus.

Par ailleurs, nous constatons que le travail scientifique reposant sur des archives du web ne concerne pas seulement le sujet de l'étude du web et de son histoire. Les sujets étudiés par les chercheurs que nous avons rencontrés dépassent très largement le cadre d'un intérêt des archives sur le web.

B.1.1 Sciences politiques : révoltes et conflits dans le monde arabe à l'ère numérique

Le projet de recherche

Le projet est conduit par un enseignant-chercheur de l'EHESS titulaire d'une chaire thématique intitulée « Révoltes et conflits dans le monde arabe à l'ère numérique ». Le projet sur lequel a porté l'entretien a donné lieu à un ouvrage : "Syrie, une nouvelle ère des images. De la révolte au conflit transnational".

La recherche porte sur les usages de la vidéo par les manifestants et ensuite les combattants dans le cadre du conflit en Syrie. Le chercheur, résidant en Syrie au moment du déclenchement de la guerre en 2011 a repéré un « activisme de l'image » de la part des protagonistes du terrain qui n'ont pas forcément de formation à l'image, mais qui vont filmer les vidéos d'une certaine manière, et recueillir ainsi les témoignages de victimes de la répression, des défections de soldats, de la propagande, des

hommages aux martyrs, *etc.* Il y a toute une culture de l'image combattante de la vidéo en particulier qui s'est développée dans le sillage de ces événements. Il faut arriver à répondre aux questions : qui fait ces vidéos ? Pourquoi ? Comment peut-on les classer ? Et surtout, comment caractériser ces écritures audiovisuelles ?

La collecte des données commence sur le terrain syrien alors que le chercheur concerné y habite et se poursuit en France. Elle s'étale sur une durée de 11 ans.

Le projet se veut fondamentalement multidisciplinaire. Il relève des sciences politiques (analyse d'un contexte de guerre et de protestation). C'est en même temps une recherche qui appartient à l'histoire du temps présent. Elle relève aussi de l'anthropologie visuelle puisqu'elle s'intéresse à la manière dont des groupes spécifiques utilisent l'image pour se représenter, pour s'exprimer. Enfin elle est concernée par les *digital studies* en ce que la culture de l'image étudiée (les vidéos) appartient à la culture numérique en réseau, ce qui implique non pas seulement des outils de collecte et de traitement, mais aussi une logique d'acteurs à comprendre.

Objets collectés

Ce sont des vidéos issues de YouTube, téléchargées quotidiennement, et qui forment à ce jour un ensemble d'environ 8 000 vidéos conservées.

À partir de ces vidéos, il s'agit ensuite de mener une enquête pour essayer de savoir qui a filmé, qui apparaît à l'image et éventuellement ensuite, de retrouver les personnes

Les vidéos sont des archives vivantes au moment où elles sont collectées, mais la possibilité de les retrouver comme archives constituées est un enjeu. Cette recherche vise à la fois à construire des archives et à en récupérer là où il y en a. Mais elle n'a pas donné lieu à une recherche dans des dépôts d'archives existants (archives historiques).

Méthodes et outils

Cette collègue travaille seule, de manière artisanale selon ses propos, sans s'appuyer sur une culture numérique particulière. La première difficulté est celle des formats et les outils comme les contenus sont instables et condamnés à l'obsolescence. Il y a des vidéos de 2011 faites sur de petits téléphones portables qui ne sont pas encore des smartphones. Elles ne sont plus lisibles par des logiciels de lecture et ne sont plus téléchargeables. On est face à l'urgence de la sauvegarde.

Les outils utilisés sont YouTube Downloader HD et You Tube Free Converter. YouTube Downloader HD est un logiciel gratuit, dénué de publicités et très simple à utiliser pour télécharger des vidéos YouTube sur son ordinateur. Cependant, l'outil propose peu de fonctions d'export ni de file d'attente pour télécharger plusieurs vidéos à la fois. You Tube Free Converter se télécharge automatiquement depuis YouTube et s'installe sur le navigateur (Firefox en l'occurrence).

Les besoins de techniques et méthodes d'archivage sont très importants pour pouvoir réutiliser les vidéos.

La stratégie de collecte débute par une recherche avec des mots-clés en arabe. Elle permet de trouver une vidéo « point de départ » pour en trouver d'autres. L'étape suivante est d'aller sur la chaîne qui l'a diffusée, ce qui permet de repérer des chaînes (des chaînes d'activistes) et non plus des vidéos comme « éléments » singuliers à sauvegarder. Une autre stratégie de recherche est de faire une recherche par événement et de procéder par « épaississement ». Dans tous les cas, une vidéo seule ne peut rien dire. Les vidéos sont toujours mises en résonance avec d'autres vidéos. Le chercheur insiste : « l'important c'est vraiment la contextualisation ».

De plus, les vidéos sont souvent anonymes. C'est une difficulté supplémentaire à cause du contexte sécuritaire et guerrier. La date de mise en ligne est primordiale. Le nom du compte utilisateur qui a mis en ligne la vidéo n'est pas forcément la personne qui a filmé. Les vidéos sont reprises, circulent de nouveau. Donc après, si on veut vraiment faire une recherche pointue, il faut rechercher la première vidéo en émettant l'hypothèse que c'est la vidéo source et nous n'en sommes même pas sûrs.

La question de l'authenticité, de la véracité, se pose tout le temps. On ne peut jamais savoir si une image a vraiment été tournée en Syrie ou pas. Évidemment, on reconnaît des lieux, etc. Mais à la limite, dans des déclarations où les gens ont des masques, qu'est-ce qui nous prouve que cette image a été tournée en Syrie ? À chaque fois, on ne s'appuie sur pas sur une image, mais sur des images.

Le but du chercheur est de savoir comment les acteurs font, et ils en sont très conscients, pour attester de la véracité de leurs images. C'est quelque chose que l'on retrouve partout et quand ce souci n'existe pas c'est un signe. Par exemple, les groupes djihadistes ne contextualisent jamais leurs images. Ce n'est pas leur question. Eux sont dans un temps déjà eschatologique, ils sont dans une autre chronologie.

Le classement est effectué par lieu, par date et par catégorie avec des sous-catégories. Pourtant la recherche dans le corpus (pur retrouver une vidéo) est difficile.

Soutien technique, soutien externe

Le chercheur n'a pas de formation numérique particulière. Elle s'est forgée une méthodologie et une expertise technique qui se sont inventées sur le terrain.

Elle n'a pas utilisé les services du pôle audiovisuel de l'EHESS, car il est orienté sur la production d'image et non sur la collecte et le traitement d'images.

En revanche, elle est en lien avec le pôle audiovisuel de la BnF, car elle est membre du projet ANR SHAKK « De la révolte à la guerre en Syrie : conflits, déplacements, incertitudes » (CéSor-EHESS/BnF/Ifpo/Iremam[1] : « L'ANR associe le Centre d'études en Sciences sociales du religieux (CéSor), l'Institut français du Proche-Orient (Ifpo), le département de l'audiovisuel de la Bibliothèque Nationale de France (BnF**)** et l'Institut de Recherches et d'Études sur les Mondes Arabes et Musulmans (Iremam) l'équipe est composée de neuf chercheurs – anthropologues, politistes, historiens et linguistes – d'un ingénieur de recherche, spécialiste des Humanités numériques, ainsi que d'un conservateur spécialiste des archives audiovisuelles ».[2]

La relation avec la BnF se fait sur l'archivage. Il s'agit de permettre de pérenniser les archives collectées, de les mettre à disposition d'un public restreint de chercheurs, de constituer ainsi une mémoire de cette guerre : « on veut que cette mémoire-là, elle soit quelque part », qu'on ne puisse pas contester ce qui a eu lieu.

Le chercheur fait part de ses besoins en termes de formation : formation au numérique pour l'aider dans ces recherches, mais aussi formation à « l'écosystème du web », et au droit à l'image.

En matière de droit à l'image, le problème essentiel est celui de la réutilisation des images pour faire des produits secondaires, par exemple de faire un film qui les reprenne. Ces images sont-elles la propriété de qui ? De YouTube ? Il y a des films qui se sont faits entièrement avec des images de YouTube.

Enfin, bien que participant à un projet ANR, le chercheur souhaiterait échanger plus largement avec d'autres chercheurs, car cette collaboration pourrait permettre d'avancer beaucoup plus vite.

[1] https://shakk.hypotheses.org/.

[2] Source : Hypothèses.

B.1.2 Sciences politiques : analyse comparée internationale (européenne) sur le rôle des médias numériques dans les dynamiques du conflit politique au sein des démocraties contemporaines (2 projets)

Projet

Ce chercheur, membre du Centre d'études européennes et de politique comparée, est porteur ou co-porteur de deux projets :

- « What Do 'the People' Want? Analysing Online Populist Challenges to Europe »[3]
- Les mobilisations d'extrême droite européenne[4]

Il s'agit de recherches comparées mobilisant des équipes et des moyens importants pour des recherches en sciences humaines.

Objets collectés

Il s'agit dans tous les cas d'archives vivantes. La temporalité intéressant ce chercheur est celle qui date de la naissance du web, c'est-à-dire des années 1990 à aujourd'hui.

Les données collectées sont issues du web, des réseaux sociaux et d'entretiens avec des militants et des activistes, principalement des mouvements conservateurs et des partis d'extrême droite.

Il s'agit d'une part de *web tracking data* (traces d'activité d'un individu) permettant de connaître les types d'information que les gens cherchent sur Internet (par des journaux classiques, par exemple, ça peut être le Monde ou le Figaro, mais aussi par des sites d'informations moins classiques). Cela représente 5 à 6000 individus européens. D'autre part, la collecte de données provient de données textuelles issues de la surveillance de comptes des réseaux sociaux. Sur les réseaux sociaux sont récupérés des textes, des statuts, des commentaires. Les données secondaires (« like », partage, emojis) sont utilisées pour des analyses quantitatives de réseaux.

Méthodes et outils

Plusieurs méthodes sont utilisées conjointement : l'analyse de réseau, l'analyse de contenu (manuelle) sur les contenus textuels et l'intelligence artificielle (machine learning) pour obtenir une catégorisation automatique des contenus basée sur l'analyse manuelle.

Soutien technique, soutien externe

L'équipe mobilisée est pluridisciplinaire : historiens, sociologue du numérique, politistes, informaticiens, mathématiciens. Des spécialistes travaillent uniquement sur la visualisation des données massives, des mathématiciens sur la fabrique d'indicateurs, par exemple sur la dispersion de l'attention.

Il est intéressant de constater que, dans ce type de projet, de nouveaux acteurs deviennent fournisseurs de données : il ne s'agit pas seulement des acteurs de l'industrie numérique (les GAFAM propriétaires des plateformes où s'agrègent les *big data*), mais aussi d'intermédiaires reconnus par les institutions qui financement les projets telles l'Union européenne, des sociétés privées, dont la mission

[3] https://www.sciencespo.fr/centre-etudes-europeennes/en/node/27002.html

[4] https://www.sciencespo.fr/research/cogito/home/ce-que-les-droites-extremes-europeennes-partagent-sur-les-reseaux-sociaux/

est de collecter et d'anonymiser des données personnelles pour des traitements ultérieurs et qui peuvent aussi contribuer à leur traitement. On peut citer comme exemple Netquest.

B.1.3 Littérature et histoire contemporaine : Récits d'inceste (publiés en France) 1986-2019

Projet de recherche

Il s'agit d'une recherche doctorale financée par des fonds ANR. Le chercheur est rattaché au laboratoire Pléiade de Paris 13 et l'IRIS par son co-directeur de recherche. Les deux directeurs de recherche sont historiens. Le projet ANR concerné est DERVI (Dire, Entendre, Restituer les Violences Incestueuses). Il se présente ainsi :
« À la croisée des études socio-historiques, juridiques et anthropologiques, la recherche pluridisciplinaire DERVI renouvellera la connaissance de l'inceste en se focalisant de manière novatrice sur sa divulgation et sur sa dimension empirique. DERVI étudiera ce moment primordial du dévoilement de l'inceste à différentes échelles (familles, réseaux institutionnels, médias), en divers contextes (familial, judiciaire, administratif, médiatique, littéraire), suivant des modalités (révélation, dénonciation, témoignage, signalement, détection, récits, « affaires ») et avec des répercussions (indignation, indifférence, déni, scandale) variables. Qui dit ou montre quoi ? À partir de quand ? Qui entend quoi ? Qui en rend compte ? Quelles sont les modalités du dévoilement de l'inceste ? Comment est-il accueilli ? Les réponses à ces questions aideront à comprendre ce qui permet ou empêche de dire, entendre et restituer l'inceste aujourd'hui. DERVI associera les terrains ethnographiques longs et les approches historiques, fondées sur l'analyse qualitative et quantitative (…) ».[5]

La recherche doctorale porte sur les récits autobiographiques. La doctorante souligne là aussi le caractère fondamentalement pluridisciplinaire de la recherche concernée : elle vient de la sociologie, a fait des études de lettres qui lui ont permis de devenir professeur de Lettres dans le secondaire et finit par faire une thèse dirigée par des historiens. Elle souligne les difficultés dans la constitution d'une méthodologie acceptable par ses directeurs de thèse, par exemple, relativement à l'idée même de corpus.

Objets collectés

Il s'agit de repérer des récits autobiographiques, principalement imprimés et d'étudier la révélation d'un problème qui devient public par la diffusion de ces récits dans la période récente et de leur réception dans l'espace public.

Ces récits sont difficiles à identifier. Leur identification est presque plus importante que le traitement qui en sera fait. Le catalogue de la BnF ne permet pas de les repérer de manière satisfaisante et il est nécessaire de les identifier de manière indirecte : par les sites d'associations de victimes, par de multiples associations relatives à l'inceste, par la presse et la télévision (consultation des archives) d'où l'intérêt des archives de l'INA même si certaines d'entre elles ne sont accessibles qu'après l'avoir été sur YouTube. La recherche implique de trouver des archives historiques autant que des archives vivantes. Les associations changeant de nom, la consultation des archives du web permettra de retrouver les anciens sites et de trouver les nouveaux noms. Les réseaux sociaux sont le moyen d'identifier la réception de ces récits (tweets, posts Instagram et Facebook).

[5] Source : https://dervi.hypotheses.org

Méthodes et outils

Tous les matériaux trouvés sont signalés, stockés dans Zotero.

Pour identifier ces récits et leur réception, une observation participante est nécessaire et doit être couplée à la recherche bibliographique. Cela suppose de créer des comptes dans Twitter par exemple.

Mais il a également été utile de s'inscrire dans des réseaux sociaux littéraires comme Babélio, car l'indexation permet de trouver ce qu'on ne peut pas identifier dans un catalogue professionnel tout en analysant la réception de ces récits. La consultation de sites de vente en ligne d'ouvrages comme Leboncoin permettent de trouver des livres sur le sujet ayant servi à des étudiants d'écoles de formation au travail social : la recherche sur ce type de site vient compléter ka collection de récits.

Une volonté de trouver des méthodes de valorisation numérique de ce corpus a également été soulignée : par exemple comment faire une sorte de site musée des récits d'inceste (forme déjà vue sur d'autres données) ?

Soutien technique

Le soutien technique est nécessaire pour la numérisation d'ouvrages ou d'éléments susceptibles de faire partie du corpus d'étude afin d'envisager des traitements automatisés ultérieurs. Mais il l'est également pour traiter le corpus recueilli.

Une formation offerte par l'infrastructure de recherche HumaNum a permis à la doctorante de faire l'apprentissage d'Iramuteq, ce qui n'a pas été possible par l'école doctorale. Elle l'a aussi initié à Nakalona, pack logiciel pour créer des bibliothèques numériques. Elle a également bénéficié des services d'une autre infrastructure de recherche, PROGEDO[6], car résidante en Pays de la Loire bien qu'elle effectue une thèse à Paris.

La BnF aurait été un partenaire souhaité, car le corpus est d'intérêt public et mériterait d'être mis à la disposition des chercheurs.

Un besoin de soutien juridique est là aussi déterminant pour en savoir davantage sur la fouille de texte, le droit de reproduction. Il y a des récits non publiés, des données sensibles dont il faut savoir comment les ouvrir à d'autres recherches.

Enfin, le problème de l'archivage pérenne de ces corpus si difficilement constitués est une question cruciale à laquelle les chercheurs ne sont pas formés.

B.1.4 Recherche sur l'histoire du web

Nous regroupons ici trois entretiens. Trois projets de recherche sont concernés.

A) wikimaps.io

Projet de recherche

Ce projet exploratoire contribue aux initiatives visant à rendre Wikipédia plus accessible et intelligible, en valorisant et simplifiant l'exploitation de son potentiel. En particulier, il s'attache à valoriser les traces de la consultation des pages, comme autant d'actes élémentaires qui engagent l'attention collective et nous informent sur ce qui compte. Plus spécifiquement, en considérant les langues, le temps et l'espace, il est ainsi possible de mieux caractériser les régimes d'attention et la globalité

[6] www.proged.fr

relative des faits sociaux (attentats, élections, théories du complot, controverses, événements sportifs ou culturels...).

Le projet bénéficie du soutien financier du fonds national Suisse.

Objets collectés

Les fichiers logs des pages Wikipédia (archives du web) qui ont publiquement accessibles.

Méthodes et outils

Développement d'outils spécifiques (avec du Python, NodeJS, MariaDB, D3JS et WebGL) pour collecter les logs, les analyser et les représenter sous forme graphique.

Soutien technique

Il mobilise des compétences multiples grâce à la création d'une équipe pluridisciplinaire par l'Université de Lausanne et grâce au soutien d'un partenaire externe, Wikimédia, pour traiter les données.

B) Évolution des standards du web et loi sur la protection des données à caractère personnel

(histoire de la proposition (infructueuse) du standard *Do Not Track* et élaboration du Règlement Général de Protection des Données)

Projet de recherche

Il s'agit d'une thèse de doctorat qui porte sur l'analyse des débats et controverses qui ont eu lieu dans les arènes d'élaboration des normes et lois relatives à la notion de « donnée à caractère personnel » (notamment *Do Not Track* et le RGPD). Pour cela, il se base sur une analyse des discours et des thématiques qui sont inscrits dans les textes d'acteurs impliqués W3C, Commission européenne, OCDE, lobbyistes ... Le projet est réalisé dans le cadre d'une thèse de doctorat (financement de thèse par le laboratoire).

Objets collectés

Ils sont composés des pages web institutionnelles, documents publiés émanant de l'Union européenne ou d'acteurs extérieurs participant indirectement au débat (notes publiées des lobbyistes et autres groupes extérieurs à l'Union européenne). Les sources sont, en particulier, les archives du W3C et celles de la Commission européenne disponibles sur le web.

Méthodes et outils en fonction des objectifs poursuivis

Un script a été développé spécifiquement pour faire une analyse textométrique destinée à repérer le mécanisme de la prise décision publique selon le modèle *Advocacy Coalition Framework*.

C) ResisTIC

Les résistants du net. Critique et évasion face à la coercition numérique en Russie, ANR-17-CE26-0020 (financé par l'ANR et soutenu par le CitizenLab, Université de Toronto).

Projet de recherche

Ce projet réalise une enquête inédite sur les résistances en ligne en Russie pour mettre à jour des pratiques sociales et des techniques de contournement des contraintes en ligne peu connues. Il a l'ambition de contribuer, au-delà du cas russe, aux réflexions sur les reconfigurations du politique à l'épreuve des techniques de la communication dans le monde contemporain.

Voici ce qu'indique une publication issue du projet :

Pursuing the autonomisation and "sovereignisation" of their national Internet (RuNet) since the early 2010s, authorities in the Russian Federation are establishing increasingly stricter regulations on Internet innovation and practices. Since 2018, the team of the ResisTIC (Criticism and circumvention of digital borders in Russia) *project explores how different actors of the RuNet resist and adapt to the recent wave of authoritarian and centralizing regulations. One of the project's primary objectives is to explore the extent to which control and circumvention strategies are embedded in, and conducted by means of the infrastructure of the RuNet. This special issue provides a detailed overview of the different strands of research undertaken by the ResisTIC project team at the crossroads of digital sovereignty, data and infrastructure. Articles by the project team are entwined with contributions by specialists based in Russia and worldwide*. (Daucé F., Musiani F. 2021. « Infrastructure-embedded control, circumvention and sovereignty in the Russian Internet: An introduction »[7].

L'entretien a été mené avec une spécialiste de sociologie de l'innovation, chargée de recherche CNRS et qui a participé au projet Web90 dirigé par Valérie Schaffer et bénéficié du soutien technique et des ressources de la BnF. Ce projet est décrit par l'équipe DLN de la BnF dans le présent document.

Objets collectés

Les « objets » collectés sont des sites et des blogs, des entretiens dans le cadre du web vivant.

Méthodes et outils

Si la chercheuse a participé au projet Web90, elle n'a pas mobilisé l'expertise acquise sur les archives du web pour traiter de ce projet consacré au web russe. L'équipe du projet se compose de chercheuses et chercheurs en sciences humaines et sociales. Le besoin d'appui, notamment technique, est moins pressant que dans le cadre de Web90, car il n'y a pas de besoins en infrastructure de *crawl* et d'archivage.

Ces recherches de 2022 sur le web avec des matériaux du web montrent en comparaison avec l'étude menée par la BnF en 2011 (« Etude sur les archives de l'Internet : rapport d'analyse ») que :

- Le web est ancré comme terrain de recherche, il n'est plus seulement un début de terrain ou un espace pour identifier de premières pistes. Il sert ainsi aujourd'hui à mettre en exergue des tendances, des événements, des pratiques, mais aussi des influences.
- Pour les chercheurs interrogés (ils ne sont pas historiens contrairement à ceux de 2011), il semble tout à fait normal de mobiliser des sources/contenus/terrains issus du web. Celui-ci est tout à fait légitime pour des travaux de recherche.
- Le fait que le web est mouvant (des ressources apparaissent et disparaissent), contrairement au traitement des archives historiques en général, semble intégré dans les recherches. L'usage des archives du web (notamment par la Wayback machine) est plébiscité.

[7] https://firstmonday.org/ojs/index.php/fm/article/view/11685

- La question de la sélection des données et de l'échantillonnage (la représentativité d'un corpus) continue à se poser.

A.2.Expériences pédagogiques (J. Casenave, L. Favier, A. Henry)

Enseignement sur l'histoire des médias vus par le web

Deux des chercheurs interrogés à propos des projets sur l'histoire du web ont également signalé un usage des archives du web pour l'enseignement.

Le chercheur impliqué dans Wikimaps.io s'appuie sur les archives du web pour son enseignement afin d'illustrer l'évolution des « unes » des journaux en ligne (passage d'une conception pour les humains à une conception pour les outils de crawl).

Dans un autre cas, le chercheur portant le projet sur les standards du web, aurait souhaité mettre en œuvre une démarche pédagogique (niveau Licence) d'analyse comparative entre les pages du début du web et le web actuel pour illustrer à la fois les évolutions technologiques, mais aussi les évolutions dans les pratiques du graphisme sur le web.

Retour d'expérience (J. Casenave)

Dans le cadre d'un cours assuré conjointement par Joana Casenave et Laurence Favier aux étudiants de Master 1 Information-Documentation de l'Université de Lille, il a été décidé d'aborder le domaine des archives du web. Sur les 20 heures que compte le cours, une séquence de 10 heures a été consacrée à cet objet d'étude.

L'objectif était triple :

- faire découvrir aux étudiants les archives du web ;
- les amener à s'interroger sur les caractéristiques externes et internes de ces archives et réfléchir ainsi aux notions de document et d'archive qu'elles convoquent ;
- les diriger et les accompagner dans un travail personnel d'analyse de ces archives.

Organisation du cours

Le cours sur les Archives du web a été organisé en deux séquences :

Une première séquence du cours a été dédiée à la présentation des archives du web et des typologies qu'elles recouvrent, des problématiques de collecte et de conservation qui leur sont propres, ainsi que des enjeux de médiation qu'elles soulèvent : introduction historique, modes de collecte et critères de sélection, spécificités des archives du web (instabilité des contenus et dimensions temporelles dans les corpus d'archives du web – contenu original, accumulation, transformations), métadonnées et description des archives du web, stockage et modalités d'accès aux archives du web.

Une seconde séquence du cours a été dédiée à l'analyse des archives du web. À partir de notions ayant trait à la diplomatique et à l'archivistique, plusieurs points ont été abordés : critique externe et critique interne des documents, authenticité, origine et auctorialité, datation et typologie documentaire, originalité et spécificités, contextualisation, archivage et préservation. L'objectif était de fournir aux étudiants des éléments analytiques afin qu'ils puissent élaborer une grille d'observation de ces archives, dans le cadre d'un travail de recherche personnel.

La grille analytique documentaire proposée par Louise Gagnon-Arguin, Sabine Mas et Dominique Maurel (2019) a également été présentée aux étudiants, afin qu'ils puissent l'enrichir et l'adapter à l'analyse des archives du web. Cette grille comprend initialement 11 sections permettant d'analyser

tout document de manière globale, en tenant « compte des différents aspects sous lesquels un document peut être considéré dans un contexte administratif et archivistique », avec l'objectif de « dégager les caractéristiques propres [du document], tout en le situant dans l'ensemble auquel il appartient » (Gagnon-Arguin, Mas, Maurel, 2019, p. 43). Voici les 11 sections de cette grille : contexte de création, définition, contenu, conditions de validité, fonctions, conservation, responsable, documents et dossiers reliés, informations complémentaires, lois et règlements, et bibliographie.

Les étudiants se sont basés sur l'ensemble de ces éléments pour élaborer leur propre grille, adaptée à leur corpus d'observation.

Enfin, l'ensemble de la promotion a été reçue à la section Patrimoine de la Médiathèque Jean Lévy, par le bibliothécaire en charge du poste de consultation BnF des Archives du web. Les étudiants ont ainsi pu découvrir le fonctionnement de ce poste de consultation, avec la possibilité d'effectuer des recherches libres dans la collecte large, ainsi que d'explorer les collectes ciblées et les parcours de navigation qui y sont associés.

Travail des étudiants

Dans le cadre de ce cours, les étudiants ont réalisé un travail personnel de recherche qui a été évalué par les deux enseignants-chercheurs responsables de ce cours. À partir d'une liste de sujets qui leur ont été proposés (liste en annexe), les étudiants ont réalisé une recherche documentaire ayant pour finalité de repérer et collecter des documents provenant à la fois du web vivant et des archives du web. Les étudiants ont ensuite effectué une analyse comparative de ces deux types de documents, à partir de la grille analytique élaborée en cours.

En sus, il leur a été demandé de mener, sur le sujet choisi, une étude bibliographique, afin d'identifier, le cas échéant, des articles scientifiques s'appuyant sur des sources documentaires provenant du web (vivant ou archives du web).

Les étudiants ont travaillé en binôme et ont, pour la partie « Archives du web », consulté le site Internet archives.org, le poste de consultation BnF de la médiathèque Jean Lévy, ainsi que le poste de consultation INA du SCD de l'Université de Lille.

Remarques sur le travail des étudiants

À la lecture du travail des étudiants, voici quelques remarques. Il nous semble intéressant de signaler les points de difficulté rencontrés par les étudiants, qui font certainement écho à des difficultés plus globales que pourrait rencontrer le public intéressé par les Archives du web. Ces points pourraient alors faire l'objet d'une attention particulière de la part de l'équipe RESPADON.

Difficultés rencontrées par les étudiants :

- Difficulté à repérer les contours des archives du web et à le différencier du web vivant, dans deux situations :
    - lorsque les pages web archivées par la BnF ou par Internet Archives sont encore disponibles sur le web vivant,
    - lorsque les pages disponibles sur le web vivant comportent une date de création ou de rédaction éloignées temporellement de notre époque.
- Difficulté à comprendre la différence typologique entre les archives du web collectées par la BnF et les sections « archives » disponibles sur certains sites web, notamment sur les sites des organes de presse.
- Difficulté à intégrer la différence entre une publication web et une mise à disposition sur le web de documents numérisés ou PDF. Ainsi, certains étudiants font un amalgame entre un

article scientifique publié sur le web, accessible par les bases de données bibliographiques et une page web.

- Difficulté à choisir des éléments rigoureux d'analyse des archives du web, dont la typologie varie grandement d'un document à un autre, et qui se situent au croisement de plusieurs domaines : archivistique, documentation, bibliothéconomie.
- Difficulté à mener une analyse sur un site web archivé lorsque la collecte d'archives du site a été réalisée uniquement en surface (page d'accueil archivée, mais peu – ou pas du tout – de pages disponibles au-delà de la page d'accueil) : problématique de la constitution des corpus d'archives du web.

Les étudiants ont par ailleurs relevé plusieurs éléments au cours de leur travail :

- l'importance des questions de méthodologie d'analyse, nécessaires à une appréhension efficace de ces ressources documentaires du web. Un cadre méthodologique combinant des notions d'archivistique, de diplomatique, de philologie et d'analyse documentaire paraît ainsi souhaitable.
- les questions éthiques soulevées, dans le domaine des Archives du web, par les modes de collecte et les usages possibles de ces archives.
- les enjeux liés à la patrimonialisation du web, à une époque où les publications web constituent une part essentielle de nos mémoires collectives et individuelles.

Le repérage d'articles scientifiques se basant sur des sources provenant du web a également fait l'objet de recherche de la part des étudiants. Les résultats obtenus montrent que les articles faisant référence à des archives du web sont rares, et le cas échéant, il s'agit généralement de pages web archivées par Internet Archives. De plus, les étudiants ont tendance à considérer toute référence, dans un article scientifique, à une page web datée de plusieurs années (des tweets, un discours publié sur le web, une page de blog) comme étant l'utilisation d'une archive du web, et ce même lorsque la page web citée est encore disponible sur le web vivant.

Ce travail sur les Archives du web a été très apprécié des étudiants, qui ont eu ainsi l'occasion de découvrir de manière approfondie un pan méconnu de la documentation web. Cette expérience pédagogique met également en avant les besoins de formation des étudiants sur cet objet documentaire, dans l'optique d'une appréhension globale des spécificités de ces archives du web et des enjeux qu'elles soulèvent.

B. Contribution de la BnF : typologie des projets de recherche autour du DLN

20 ans d'Archives du web, 20 projets

La Bibliothèque nationale de France a testé la collecte automatique de publications web à des fins de conservation pour la première fois en 1999. Cette mission patrimoniale s'est trouvée confirmée et renforcée par l'instauration du dépôt légal de l'Internet (loi DADVSI de 2006 et son décret d'application de 2011). La gestion des Archives du web repose sur des équipes documentaires et techniques, qui organisent la collecte, la conservation, l'accès à ces collections nativement numériques. L'activité de collecte se répartit entre la collecte large du domaine de premier niveau ".fr", menée avec un objectif de représentativité, et des collectes ciblées appuyées sur des sélections menées par un réseau de

correspondants internes et externes en vue d'assurer un archivage plus complet et régulier de certaines pages ou sites web.

Cette entreprise est loin d'être anodine, puisqu'elle fait entrer une forme de collection inédite au sein de la bibliothèque, pour laquelle les possibilités et freins techniques restent intrinsèquement liés aux usages à venir. La collection des Archives du web se construit pas à pas, en tenant compte des caractéristiques du web à une époque, en explorant les possibilités de stockage et indexation en fonction des moyens alloués. Depuis le début de cette collection, les interactions du DLN avec les chercheurs visent à s'assurer que les choix opérés dans cette construction sont cohérents avec les besoins de la communauté de recherche. Cette attention aux usages relève toutefois d'une certaine utopie : il a fallu que la communauté scientifique elle-même se lance dans ces projets inédits de recherche recourant aux matériaux numériques, et les usages à venir ne peuvent être totalement prédits… C'est donc au terme d'une certaine accumulation de cas et projets qu'un retour d'expérience sur les démarches menées peut être entrepris. Des bilans de projets étaient bien sûr mis en place à chaque occasion, mais c'est une perspective multi-projets que l'on cherchera à déployer ici, en revenant sur 20 projets esquissés ou réalisés au fil des 20 années des Archives du web.

B.1. De l'enquête de 2011 à l'analyse de 2021

Pour démarrer ce regard rétrospectif, rappelons aussi qu'une précédente enquête avait été conduite en 2011 par la DSG auprès d'une quinzaine de chercheurs pour explorer leur retour sur les archives du web. À cette occasion, il était retenu dans un premier temps la difficulté à identifier des interlocuteurs sur le DLN, l'intérêt des chercheurs pour le concept d'archives de l'Internet, mais leurs incertitudes quant aux usages et conditions d'usages associés. Le renouvellement de la démarche d'étude, 10 ans après, témoigne de l'évolution des prises à ces questions : d'une part, il est possible d'avoir un retour historique plus consistant sur les projets réalisés ou abandonnés ; d'autre part, ce questionnement se place dans un contexte plus large, avec les acteurs de l'ESR comme les BU. En effet, cette analyse est cette fois-ci menée dans le cadre du projet RESPDADON, projet financé par le GIS Collex-Persée afin de déployer plus largement les accès aux Archives de l'Internet sur l'ensemble du territoire. Ce projet, coordonné par l'Université de Lille, la BnF, Condorcet et Sciences Po, intègre des actions à de multiples niveaux : techniques bien sûr, mais aussi juridiques ou organisationnels, grâce à des réflexions sur les éléments de formation et d'information nécessaires au développement des usages. Laurence Favier gère, avec le laboratoire GERIICO, un workpackage visant à une meilleure compréhension des usages actuels et potentiels des archives du web. C'est à ce chantier que contribue ce rapport proposant une analyse des projets entrepris avec la BnF. Il complète les autres travaux du workpackage proposant respectivement une analyse sociologique, à partir d'une dizaine d'entretiens, sur l'usage des matériaux web dans les pratiques de recherche contemporaine, un retour d'expérience de l'INA, ainsi qu'une analyse des publications sur le sujet. Si les retours d'expérience sur les projets entrepris avec la BnF constituent une source de réflexions certaines pour le projet RESPADON, ces éléments doivent être resitués dans le cadre plus large des pratiques de recherche. Le DLN ne peut en effet se développer sans tenir compte de l'écosystème de l'ESR, à la fois sur sa structuration (incitation à la recherche par projet et à la pluridisciplinarité, organisation des équipes projet avec des titulaires et contractuels présents sur un temps limité, suivi de performance sur la valorisation, modalités de publications, etc.), mais aussi sur sa sociologie et construction (émergence de sujets et méthodes, profils et trajectoires des chercheurs, etc.). En cela, l'articulation entre ce rapport et l'étude menée par l'équipe GERiiCO permettra de percevoir plus largement l'environnement du DLN.

B.2. Méthode : une typologie basée sur le retour d'expérience de la BnF

Dans ce rapport-ci, nous allons spécifiquement nous concentrer sur les projets ayant interagi avec le DLN. L'objectif est de comprendre les caractéristiques des projets conduits et d'esquisser une typologie de ces démarches. Ce propos permettra d'une part au DLN de mieux identifier les points d'attention des projets se présentant et d'autre part d'illustrer dans des communications aux chercheurs ne connaissant pas le DLN des modalités de travail possibles, sans singulariser un sujet ou une équipe en particulier. La démarche retenue s'est attachée à identifier des descripteurs des projets conduits, de manière unitaire, avant de regrouper les projets ayant des caractéristiques communes et de voir comment décrire de manière complémentaire ces projets « idéaux-typiques ». Cette méthode ne vise pas à « révéler » des mécanismes de projets ou de liens particuliers, mais à faire un état des lieux plus large et complet afin de caractériser des éléments de distinctions.

Dans la mesure où beaucoup de matériaux existent sur les projets réalisés, avec des interactions et des interventions régulières des chercheurs impliqués (dans des productions *ad hoc*, à l'occasion des 20 ans du DL web ou de la journée de lancement Respadon), nous avons mobilisé cette matière grise pour décrire les projets. Cette description a été enrichie par une réunion de discussion avec les équipes du DLN actuelles, Alexandre Faye, Dorothée Benhamou-Suesser, ainsi que Alexandre Chautemps et Sara Aubry qui ont une perspective historique plus longue. Nous n'avons pas reconduit d'entretiens spécifiques avec les acteurs, ni les « anciens » du DLN ni les chercheurs eux-mêmes à ce stade. C'est donc des éléments issus de la mémoire et ce qui reste des projets qui servent de matériau. Il serait probablement utile de revenir avec les chercheurs ayant participé à des projets sur les descripteurs et la typologie, pour vérifier que nous n'avons pas omis un pan de description ou des éléments qui paraissent déterminant dans le déroulé des projets, et que la typologie fait sens dans les associations et distinctions qu'elle fait émerger. Cette typologie pourrait en effet rencontrer les réflexions épistémologiques sur ces nouveaux champs de recherche, comme celles engagées par Sophie Gebeil (Website Story, 2021). Si cette démarche n'a pas pu être entreprise du fait des délais de production du rapport pour Respadon, d'autres cadres de travail du projet permettront peut-être de rebondir sur ce rapport et d'en créer une lecture commune pour les livrables du projet.

B.3. Perspective historique : l'évolution du cadre de travail

Enfin, pour mener cette démarche, il convient de redonner le cadre historique des projets qui seront intégrés dans la typologie. En effet, si la BnF a depuis 2006 mis en place une équipe dédiée au DLN, l'offre et le contexte n'ont plus rien à voir aujourd'hui avec ce qui était proposé à l'époque. Ainsi, un récit « historique » paraît nécessaire pour introduire les 20 projets qui seront décrits par la suite.

À l'origine des premières coopérations mises en place autour du DLN, figurent trois démarches qui ont toutes conduit à une amélioration des pratiques du DLN, que ce soit en matière de collecte ou d'amélioration des outils d'accès aux collections.

Dans une première phase, ce sont essentiellement les actions de collecte et de préservation de contenu en ligne, qui ont porté les premières coopérations. Le laboratoire PACTE (unité mixte de recherche du CNRS, de l'Université Grenoble Alpes et de Sciences Po Grenoble) a ainsi initié une coopération ponctuelle avec le DLN en 2007 à l'occasion de la collecte de l'élection présidentielle, participé également à la collection Jeux olympiques de 2012 et souhaité bénéficier des métadonnées de la collecte électorale 2013 pour compléter son propre jeu de données (un ensemble de 2 millions de tweets collectés dans le cadre du projet Trielec. Pour répondre à ce dernier besoin, le DLN a fourni un listing de ses seeds ou urls de départ sélectionnées pour la collecte électorale). La base de travail du DLN sert d'inventaire, souvent complémentaire, des ressources à explorer.

Les deux autres premières collaborations se sont nouées suite à des rencontres personnelles lors d'événement professionnel : il s'agit du congrès de l'AAF qui s'est tenu à Angers du 20 au 22 mars 2013 pour le « centre des archives du féminisme d'Angers » et d'un atelier du livre ayant eu lieu à la BnF le 30 novembre 2015 pour « l'association pour l'autobiographie et le patrimoine autobiographique » et son président Philippe Lejeune. Ces deux institutions sont d'ailleurs citées comme les premiers contacts du DLN, sans doute car il continue de contribuer régulièrement aux collectes projets « Journaux personnels » et « Mouvements sociaux ». Ces acteurs ne cherchaient pas à produire une recherche sur le matériau, mais à s'assurer de sa collecte et de sa patrimonialisation. Ces démarches résultent de contacts initialement informels entre des agents de la BnF et ces institutions, autrement dit d'un jeu d'opportunités, ce qui montre la difficulté à rentrer dans les archives du web sans l'aiguillon d'un professionnel.

Si les collaborations orientées vers la collecte et l'archivage ont bien fonctionné lors de cette première phase, la difficulté de manipulation des ressources et surtout les contraintes d'accès constituent déjà un frein, d'autant que les chercheurs (Fabienne Grefft, Thierry Vedel) avaient souhaité faire travailler des étudiants sur les contenus web archivés.

La deuxième phase de collaborations avec le DLN correspond à l'accueil de projets plus structurés, dont l'objectif premier est l'analyse du contenu des collectes et qui, à partir de 2015, vont conduire aux développements de nouveaux outils d'accès par le DSI (application Internet Labs, parcours guidés) ou amener le DSI et le DLN à travailler avec les ingénieurs de recherche de l'équipe projet. L'intérêt croissant des chercheurs pour la fouille de textes et de données a ainsi motivé l'inscription par la BnF à son plan quadriennal de recherche 2016-2019 du projet CORPUS, afin de préfigurer un service de fourniture de corpus numériques à destination de la recherche, qui lui a permis d'engager ces nombreuses collaborations.

Le premier projet clé de cette période a été porté par la sociologue Valérie Beaudouin, qui avait pris langue avec le DLN afin d'archiver des matériaux de travail sur les blogs des écrivains. Dans la foulée de la discussion, elle a élaboré un projet soutenu par le labex « Les passés dans le présent » sur les archives de la Grande Guerre et piloté par la DSG. La collaboration avec Zeynep Pehlivan, ingénieure de recherche, et le DLN a conduit à deux projets successifs et imbriqués : une cartographie du web de la Grande Guerre et une analyse de la circulation et réappropriation des sources historiques numérisées issues de Gallica. C'est donc le fait que les AW constituent un ensemble cohérent ou vue du web à un instant donné qui s'avéra central dans ce projet. Le travail porta concrètement sur une analyse des liens entre les sites, c'est-à-dire des métadonnées qui rendent compte de la structuration des réseaux sur le web. Toutefois, la cohérence du corpus s'avéra critiquable dès qu'il s'agissait de mettre en place une analyse diachronique : la multiplication des liens durant la période du centenaire de la Grande Guerre correspondait-elle à une intensification des interactions sur le web ou relevait-elle d'une amplification de la collecte ?

Néonaute est également un projet qui a associé un chercheur, Emmanuel Cartier, linguiste, et un ingénieur de recherche, Loïc Galland. Il répondait à l'appel à projets « Langues et numérique 2017 » lancé par la Délégation à la langue française et aux langues de France (DGLFLF). Le travail portait cette fois sur une analyse textuelle des contenus, afin d'identifier le vocabulaire émergeant du web. L'équipe projet a mis en place un process permettant d'explorer les contenus de la collecte Actualité, de les analyser et d'exporter le résultat sous forme de segments textuels. Les outils mis en place à cette occasion visaient à explorer les contenus et non pas les métadonnées. Le projet Néonaute a pu bénéficier des apports de deux autres projets Web90 et ASAP, qui ont conduit au développement de l'application Internet Labs et à un travail d'indexation des fichiers WARC permettant une recherche par mot dans des collections définies (collection Web90 et Attentats).

Durant cette même période, les projets Web90 et ASAP pilotés par la chercheuse Valérie Schafer correspondent à une démarche plus historique, qui exploite en même temps qu'elle critique les archives web comme matériau de recherche. Le projet ANR Web90 a réuni une équipe

pluridisciplinaire de 8 chercheurs qui ont pu participer au développement de l'application Internet Labs et aux travaux d'indexation menés par le DSI. Cette équipe n'a pas développé de process ou d'analyse automatisés, mais s'est appuyée sur des méthodes traditionnelles d'analyse, de contextualisation des captures, aboutissant à la publication d'un parcours guidé sur le web des années 90. Celui-ci constitue à son tour une porte d'entrée très riche sur les archives web de cette période. Dans ces travaux, les modalités de constitution de l'archive et le caractère patrimonial de l'archive web sont au cœur des réflexions.

De ce point de vue, le travail de Sophie Gebeil a poursuivi ce travail méthodologique. Accueillie dans le cadre du dispositif « chercheuse associée » alors qu'elle produisait sa thèse sur « La fabrique numérique des mémoires de l'immigration maghrébine sur le web français (1999-2014) », la chercheuse a mis en œuvre un cadre réflexif sur le matériau et les méthodes, qui intègre et interroge également d'autres sources externes au web pour étudier le web (entretien avec des producteurs, reconstitution de l'itinéraire de contenu non nativement numérique passé sur le web, grille d'analyse d'un site…etc). Comme pour l'équipe Buzz-F, le travail donna lieu à la publication d'un parcours guidé.

Ces deux démarches, une première orientée vers l'analyse automatisée et une seconde vers l'approche critique historique, ne sont pas à opposer en ce sens que des outils et des préoccupations sont partagés. Elles montrent, l'une et l'autre, une certaine maturité et font apparaître l'importance des interactions entre les chercheurs et les équipes de la BnF (DLN et DSI), et mettent à jour une certaine complémentarité pour des sujets et contenus pourtant différenciés (sciences politiques, sociologie, linguistique, histoire).

Depuis 2020, une nouvelle phase s'est ouverte caractérisée par la formalisation des dispositifs d'accompagnement. La mise en place du BnF DataLab a été l'occasion de structurer une offre de services autour des archives du web, construite à partir des expériences précédentes, et apportant une plus grande lisibilité aux actions de soutien à la recherche proposées par l'établissement. Une proposition de collecte à la demande, une aide à la fouille de données et un service d'extraction de métadonnées et données ont été conçus pour répondre aux principaux types de besoins identifiés. Le projet ResPaDon, officiellement lancé au printemps 2021, se donne pour objectif de mettre en place des capsules expérimentales d'accès aux archives du web dans l'environnement de l'ESR. Toutes ces dynamiques contribuent à poser un nouveau cadre de travail et à ouvrir une nouvelle phase.

Dans le même temps, lors de la crise sanitaire de la Covid-19 et en particulier lors du premier confinement, la BnF et son réseau externe de partenaires territoriaux (réseau des BDLI) ont constitué une archive web de la Covid-19, qui a suscité l'intérêt d'équipes de recherche de SHS n'ayant encore jamais travaillé sur ce type de matériau. Ces contacts, s'ils ont parfois donné lieu à des propositions de projet intéressantes, ont rarement débouché sur une réalisation soit faute de financement, soit du fait de conditions juridiques d'exploitation des données jugées trop contraignantes, ou bien encore, car les équipes de la BnF considéraient que la faisabilité du projet n'était pas acquise pour des raisons techniques, méthodologiques ou de calendrier.

Les projets initiés à partir de 2020 montrent une nouvelle évolution par rapport à la phase des années 2016-2019 : ils mobilisent de nouveaux outils développés par des partenaires de la BnF (SolrWayback de la Bibliothèque royale du Danemark, Hyphe du Medialab de Sciences Po), les API des grandes plateformes du web (blogspot, Twitter) et accordent une place prépondérante à l'image et aux contenus audiovisuels.

Le projet Bodycapital (ERC Advanced Grant BodyCapital) est porté par le chercheur Christian Bonah du laboratoire SAGE (Université de Strasbourg), historien des sciences et spécialiste de l'analyse des productions audiovisuelles. Christian Bonah a mené un premier travail de recherche avec l'Ina portant sur une collecte de tweets réalisée à partir de l'API Tweeter. L'équipe projet s'est associée aux équipes de la BnF (DLN, DSI, Département des collections-Sciences et Techniques) pour produire une collecte à la demande sur le thème de l'alimentation. Les fichiers produits ont été livrés à la BNUS, partenaire du projet qui a mis en place une infrastructure de travail pour le déploiement de l'outil SolrWayback.

Cet outil propose des fonctionnalités de recherche, de visualisation et d'export avancées par rapport aux outils standard de consultation de la BnF et en particulier une recherche par image très attendue par l'équipe projet.

Le projet Lifranum vise à élaborer un corpus représentatif des productions littéraires francophones nativement numériques et à faciliter son usage dans différents champs disciplinaires en proposant un outil de recherche sous la forme d'une plateforme web. Les fonctionnalités de recherche sont enrichies par des annotations produites grâce à un ensemble de traitements innovants. Outre les métadonnées et données dérivées, la plateforme web permettra de récupérer des valeurs sur la proximité stylistique des textes générées par une intelligence artificielle. Bénéficiant d'un financement ANR, le projet est remarquable par son interdisciplinarité. Il est coordonné par le chercheur en littérature Gilles Bonnet du laboratoire MARGE (Université Jean-Moulin - Lyon 3) et mobilise également le laboratoire en informatique ERIC (Université Lumière - Lyon 2). La BnF est co-responsable de la phase consacrée à la collecte des contenus web, pour laquelle elle apporte une aide technique scientifique (présentation du robot Heritrix, montée en compétence de l'équipe). L'outil Hyphe est utilisé pour une exploration guidée des réseaux de liens, ainsi que pour l'ajout et la création d'annotations qui viendront enrichir les données de la plateforme et la connaissance des collectes BnF.

Dans ce panorama historique, la troisième phase de projets témoigne d'une certaine maturité des recherches menées autour des archives du web.

Le point de départ des projets peut ainsi aussi bien être une collecte, qu'un jeu de métadonnées ou l'exploitation des possibilités offertes par les outils développés par le DLN et la communauté. Plusieurs équipes-projets posent désormais la question de l'articulation entre les analyses qualitative et quantitative, autrement dit entre l'analyse automatisée et l'approche critique historique. C'est le cas du projet Buzz-F porté par la chercheuse Valérie Schafer et lauréat de l'appel à projets BnF DataLab. Ces problématiques de recherche conduisent à un renouvellement méthodologique, qui interroge la notion de représentativité d'un document d'archives. L'approche par échantillonnage (aléatoire ou construit), la datavisualisation, l'annotation des corpus sont des méthodes proposées pour répondre à ce défi (projets Lifranm, Bodycapital, Buzz-F). Si les nouveaux outils SolrWayback et Hyphe permettent en principe aux chercheurs de gagner en autonomie dans leur exploitation des collections ou d'être plus présents dans la préparation des collectes et la production de l'archive, ils requièrent aussi une montée en compétences des équipes projets, la mise en place d'infrastructure et une coordination plus soutenue pour s'assurer que les connaissances et les choix scientifiques soient totalement partagés.

Le tableau suivant reprend 20 projets de cette histoire des recherches autour des archives du web à la BnF en reconstruisant ces trois grandes phases et en intégrant également les projets qui n'ont pas abouti : regarder les projets abandonnés ou non retenus amène à dessiner le périmètre rendu possible par le DLN et les conditions permettant de mettre en œuvre ces recherches, conditions qui ne sont pas toujours réunies (côté DLN ou côté recherche).

Illustration 1 : Périmètre - 20 projets décrits pour esquisser une typologie

REF	Projets	Sujet	Chercheur(euse)	État
Phase 1 : 2012-2014				
1	PACTE-Grenoble	Élection présidentielle et Jeux Olympiques sur le web	Fabienne Greffet Jean-Marc Francony	Terminé
2	Trielec- Internet en campagne	L'élection présidentielle sur le web et les RSN	Thierry Vedel Sciences Po	Terminé
Phase 2 : 2015-2019				
3	Le web de la Grande Guerre	Les archives de la grande guerre	Valérie Beaudouin Zeynep Pehlivan	Terminé
4	web 90	Histoire du web et des médias	Valérie Schafer Équipe de 8	Terminé
5	ASAP	Archives et archivages du patrimoine nativement numérique face aux attentats	Valérie Schafer Équipe	Terminé
6	Chercheuse associée S. Gebeil	Le web comme espace de mémoire des migrations	Sophie Gebeil	Terminé
7	Néonautes	Vocabulaire du web : repérage de l'usage des néologismes et de la féminisation des noms de métier	Emmanuel Cartier Loïc Galland	Terminé
Phase 3 : Depuis 2020				
8	Chercheur associé A. Delaporte	La mémoire des animaux de guerre lors des commémorations du centenaire de la 1ère guerre mondiale	Alexander Delaporte	Terminé
9	Lifranum	Littérature francophone numérique	Christian Cote	En cours
10	Bodycapital	Images et audiovisuels sur le web liés à l'alimentation au regard des questions de santé et bien-être	Christian Bonah	En cours
11	Covid et Migrations, Maison française d'Oxford	Images des migrations et analyse de la représentation des parcours à partir d'un corpus web	Thomas Lacroix	Abandonné
12	MigraChiCovid	La diaspora chinoise et les conditions de vie des migrants chinois durant la crise sanitaire de la Covid	Simeng Wang	Abandonné
13	FakeNews et IA	Traitement de l'information et vérification des sources (IA et images truquées)	Xavier Fresquet	Abandonné
14	C. Tosseto	Critique théâtrale	Cristina Tosseto	Non retenu
15	Projet Terro (en partenariat avec l'Ina et les AN)	Réception médiatique des procès (analyse lexicale des RSN et de la presse)	Pascal Plas	En cours

16	Saskia Huc-Hepher	Communautés religieuses francophones de Londres (projet de collecte inclusive)	Saskia Huc-Hepher	Abandonné
17	web mémoire	Collectes et mémoires de la Covid sur le web	Marta Severo et Sara Gesbinger	Démarrage
18	Et ma dose ?	Acception et opposition aux nouveaux traitements thérapeutiques durant la crise sanitaire (analyse des images mobilisées)	Solène Lellinger	Non retenu
19	BUZZ-F	La viralité sur le web des Harlem Shake aux memes de la covid-19	Valérie Schafer (projet lauréat de l'AAP BnF DataLab)	En cours
20	Chercheur associé A. de Forges de Parny	Raconter l'histoire sur YouTube : les nouvelles formes d'écriture numérique de l'histoire	Arthur de Forges de Parny	En cours

B.4. Analyse des projets sur les AW à partir de 4 caractéristiques

Même si l'historique souligne l'évolution méthodologique des projets et des dispositifs d'accompagnement mis en place par les chercheurs et la BnF au cours du temps, on peut s'attacher à trouver les points communs entre ces différents projets. Notre approche dans cette seconde partie consistera à positionner les principaux projets sur des axes d'analyse, afin de systématiser les comparaisons et d'esquisser ainsi une typologie des projets.

Les caractéristiques que nous nous proposons d'analyser sont les suivantes :

- Les profils des chercheurs et des équipes
- Les types des ressources utilisées
- Les méthodes mises en place et les outils utilisés
- Les résultats de recherche

Pour chaque caractéristique, plusieurs modalités descriptives sont envisagées, même si l'esquisse vise à retenir une modalité qui semble soit plus différenciante, soit plus synthétique que les autres.

B.5. Les profils des chercheurs et des équipes

Pour décrire les profils des chercheurs et des équipes qui ont conduit des projets avec le DLN, on s'est d'abord interrogé sur les disciplines universitaires de rattachement, puis sur l'organisation côté équipe de recherche et notamment les compétences « techniques » mobilisées. C'est en fait deux caractéristiques qui paraissent importantes à retenir pour comprendre les attentes en termes d'accompagnement : l'inscription de la recherche sur les archives du web dans un projet d'équipe ou dans une recherche individuelle, et la place des compétences techniques dans les projets entrepris.

La « discipline » d'une recherche pourrait paraître centrale dans l'appréhension d'un projet, à la fois du côté académique et du côté de la BnF : les départements de la direction des collections de la BnF, répartis-en 14 grands champs disciplinaires, sont impliqués auprès du DLN et du DSI dans l'évaluation, l'instruction et l'accompagnement des projets, par exemple lorsque des collectes sont prévues dans le cadre du projet. La participation des départements doit aussi contribuer à la valorisation des résultats de recherche. Ainsi, l'ancrage disciplinaire évoque un référentiel de pratiques et d'activités qui semble utile pour chaque recherche.

Dès les prémices des projets recourant au DLN, toutes les disciplines semblent pouvoir faire des Archives du web un matériau. L'historique des projets montre la variété des disciplines représentées au fil du temps, même si beaucoup embarque une démarche historiographique. Les disciplines identifiées par projet sont ainsi :

- science politique (phase 1 : projet Trielec Internet en campagne)
- histoire, dont l'histoire des médias et du web (phase 2 : Grande Guerre, Web90, ASAP, Sophie Gebeil ; phase 3 : Bodycapital, Arthur de Forges)
- sociologie (phase 2 : Grande Guerre ; phase 3 : MigraChiCovid, Saskia Huc-Hepher)
- linguistique (phase 2,: Neonautes ; phase 3 : Alexander Delaporte)
- littérature, arts du spectacle (phase 3 : Lifranum, Cristina Tosetto, web mémoire)
- sciences, dont histoire des sciences (phase 3 : Bodycapital, Et ma dose?)

Ainsi, l'augmentation du nombre de projets au fil du temps ne semble pas démentir le caractère universel des archives du web.

On pourrait se demander si certaines disciplines seraient plus avides des archives du web que d'autres. Par exemple, la science politique serait-elle la première consommatrice de ce matériau archivistique, puisque les contenus disparaissent du fait de l'agenda politique ? Il n'y a pas d'éléments corroborant cette idée voire, au contraire, les projets en science politique sont plutôt éclipsés au cours des dernières phases : soit que les chercheurs de cette discipline privilégient le web vivant, soit qu'ils aient d'autres appuis à la recherche et d'autres réservoirs de sources comme l'INA ou les ressources de Sciences Po. En conclusion de ce point sur la discipline de rattachement des chercheurs qui mènent des projets avec le DLN, soulignons que les équipes connaissent le terrain avant de plonger dans le matériau des archives du web : tou(te)s les chercheur(euse)s avaient une familiarité et une expertise antérieure sur le sujet qu'il(elle)s renouvellent en le regardant « par le biais » des archives du web. L'expertise du champ de recherche est donc première et doit se coupler avec une expertise documentaire et une expertise technique liées à la manipulation de ce matériau singulier.

En effet, derrière le rattachement à une discipline principale, on note en fait que les projets mobilisant une équipe sont interdisciplinaires dans les faits, et cela de deux manières : soit dans la constitution de l'équipe projet, soit dans l'association entre chercheur(euse)s en SHS et un ingénieur de recherche (ou un ingénieur scientifique et technique) traitant d'aspects plus techniques. Plus précisément, voilà les trois modes de composition d'équipes projet observés :

- chercheuse ou chercheur portant un projet individuellement, notamment grâce au statut de chercheur(e) associé(e) (phase 2 : Sophie Gebeil, Alexander Delaporte ; phase 3 : Arthur de Forges)
- équipe projet avec uniquement des compétences SHS (phase 1 : projet Trielec Internet en campagne ; phase 2 : Web90, ASAP)
- équipe avec IR ou IST (phase 2 : Grande Guerre, Neonautes, Bodycapital, Lifranum, web mémoire)

C'est cet axe de différenciation qui sera retenu pour la typologie. Il peut être corrélé avec le type de productions des projets : certains projets sont en mesure de produire des process ou des programmes, d'autres non. Il convient en effet de souligner, pour les candidats à venir, que les projets conduits avec le DLN peuvent avoir une composante technique forte sans que ce soit un passage obligé. La compréhension des AW comme objet numérique reste par contre un prérequis, y compris pour les projets individuels travaillant sans compétences informatiques. À ce titre, le « ruissèlement » des compétences acquises en projet DLN par des chercheur(euse)s peut s'activer : Sophie Gebeil participait au projet Web90, Arthur de Farges a pour directrice de thèse Valérie Beaudouin.

Nous n'avons pas retenu ici le cadre dans lequel les projets s'insèrent : les projets de la phase 2 se sont faits avec des financements (labEx : grande guerre, ANR : Web90, Lifranum). Il est possible que la

phase 3 et les projets à venir diversifient les montages des recherches faites avec les archives du web, dans la mesure où on voit une généralisation des approches. Les projets s'inscrivant dans le dispositif du DataLab (phase 3 : Buzz-F) devront alors être observés pour comprendre les modalités de circulation des compétences et l'évolution des formes d'accompagnement. En effet, derrière la composition des équipes projet et les montages financiers, se dessinent en miroir les besoins spécifiques vis-à-vis des équipes supports comme la BnF : les IR ont besoin d'un échange technique, allant jusqu'à nouer des liens étroits avec les équipes informatiques de la BnF ; les projets individuels privilégient la connaissance des collections et donc les liens avec les chargés de collection, ou les usages fonctionnels des interfaces AW. À ce stade, seul le projet Bodycapital a intégré une documentaliste. Ce projet se singularise aussi, car il a été accueilli et accompagné par la BNUS et peut donc servir de préfiguration au partage de l'accompagnement entre les équipes centrales BnF et les équipes en réseau. L'ouverture du BnF DataLab, l'accueil de projet par des partenaires du réseau des BDLI et la mise en place d'une capsule ResPaDon pourraient à terme modifier la répartition des compétences et le besoin d'accompagnement des projets, le but étant de ne pas générer des superpositions, mais bien des synergies (par exemple avec l'accueil des IR au DataLab).

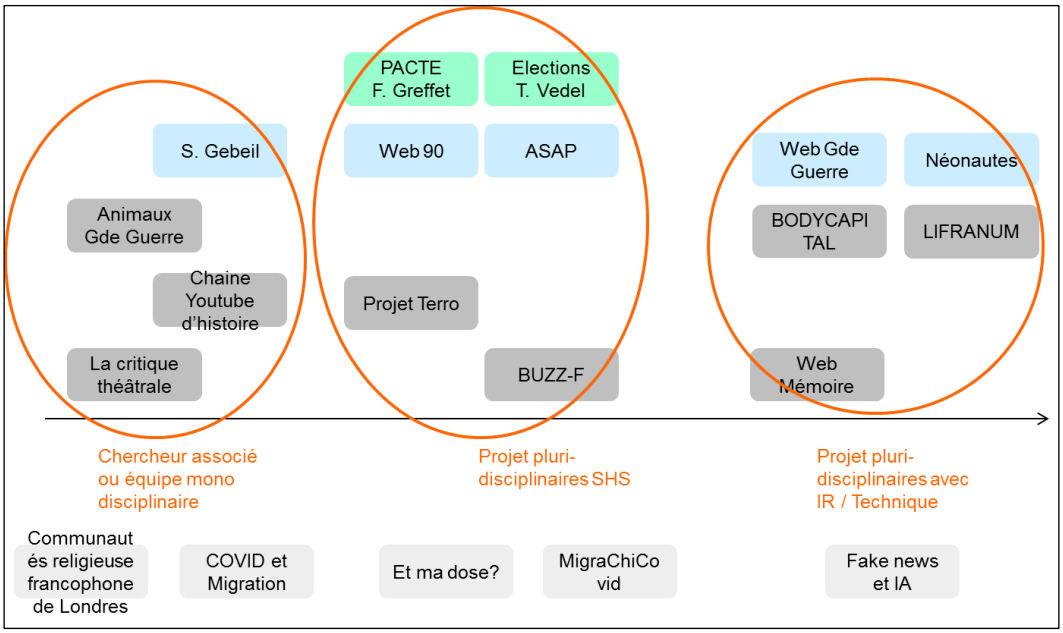


Illustration 2 : Répartition des 20 projets suivant l'axe de la composition du projet

Si les AW permettent donc à certains profils de renouveler leur recherche et de s'associer à des projets pluridisciplinaires, avec une compétence informatique ou non, on pourrait se demander si les chercheur(euse)s se lancent de la marge de leur discipline ou une fois qu'ils ont sécurisé leur trajectoire professionnelle : est-ce que les AW participent d'un braconnage des disciplines universitaires ou contribuent à une reconnaissance académique ? Ce point ne peut être traité sans des interviews plus poussés avec les acteurs, notamment pour comprendre à quel stade de leur carrière apparaissent les projets sur les archives du web, et ce que ces approches ont modifié ou infléchi dans leur trajectoire. Cette question liée au cadre académique est difficile à décrire, mais doit faire partie des connaissances à partager avec la BnF : les contextes dans lesquels se fait la recherche, que ce soit les thèses, les post-doctorats ou les projets, imposent des contraintes et des temporalités différentes et une responsabilité adaptée de l'établissement.

B.6. Les types des ressources utilisées

Pour revenir à la description des projets proprement dite, le deuxième axe de description concerne le type de ressources utilisées pour chaque recherche. Ce type peut s'appréhender de trois manières : la part des AW dans l'ensemble du corpus ; le mode de constitution des collections et enfin les données utilisées.

On peut d'abord considérer la place des archives du web dans le corpus mobilisé par le projet. En effet, les AW de la BnF peuvent être (1) l'unique ou la principale source documentaire utilisée par le projet ; (2) une source complémentaire par rapport à d'autres contenus : autres archives, web vivant, entretiens, ou autres sources primaires ; (3) une source indirectement mobilisée, par exemple dans le cas où l'équipe projet s'intéresse principalement au web vivant ou dans le cas où la recherche utilise les AW comme cadrage historiographique. L'investissement en temps ne peut pas être le même de la part des équipes projets en fonction de l'ampleur du projet et de la place qu'y prennent les AW.

Cette distinction est toutefois un élément constitutif du projet et n'impacte qu'indirectement le travail avec le DLN. Les deux éléments descriptifs suivants sont plus conséquents. Tout d'abord, malgré l'ampleur des collectes réalisées par la BnF, tout le web français n'est pas collecté à un rythme adéquat pour certaines recherches. Il y a donc des projets qui se servent des archives existantes comme un périmètre donné et daté, *a contrario* de projets qui nécessitent de compléter la collecte. Les corpus projet utilisés peuvent ainsi être distingués entre :

- Les corpus exclusivement issus des collectes BnF, pour lesquelles les sélections ont été faites par le réseau de correspondants internes et externes (phase 2 : ASAP, Grande Guerre) ; les collections AW sont alors le point de départ de la recherche ;
- Les corpus pour lesquels les sélections ont été ouvertes aux chercheurs et aux équipes projets pour que les collectes courantes puissent être complétées (phase 2 : Sophie Gebeil ; phase 3 : Cristina Tosseto) ; un des enjeux du projet est alors d'enrichir les AW et de pérenniser le matériau de recherche ;
- Les corpus pour lesquels le périmètre de la collecte et les sélections ont été définis par une équipe projet souvent avec un appui et une comparaison avec la base de travail BCweb (phase 2 : Bodycapital, Lifranum). Dans certains cas, l'existence d'archives anciennes permettant de retracer l'évolution d'un site est un élément essentiel du choix documentaire (phase 2 : Bodydcapital) ; les AW sont alors la base de travail, mais pas le point de départ ;
- Les corpus qui enrichissent les contenus avec les collectes faites par l'équipe projet partenaire, doublant souvent les collectes BnF dont elles peuvent reprendre les sélections et ce, afin de cibler un élément précis des contenus web comme le texte pour un projet de TAL ; dans ce cas, l'équipe projet demande ou cherche à produire un inventaire de contenus en mobilisant des métadonnées issues des bases archives du web : CDX, listing BCweb (phase 2 : Alexander Delaporte ; phase 3 : projet Terro) ; les AW s'articulent à un matériau plus large.

Ce dernier cas ouvre sur le descripteur du type de données utilisées : certains projets utilisent les contenus web des AW, d'autres l'indexation des contenus du web et d'autres encore, les métadonnées de la structure du web (liens, etc.). Cette distinction est importante, car le cadre juridique sur les types de matériau n'est pas le même : les jeux de métadonnées sont libres de droit, alors que les données en elles-mêmes non, leur exploitation doit être faite sur des infrastructures mises en place par la BnF ou, le cas échéant, par un de ses partenaires du réseau des BDLI.

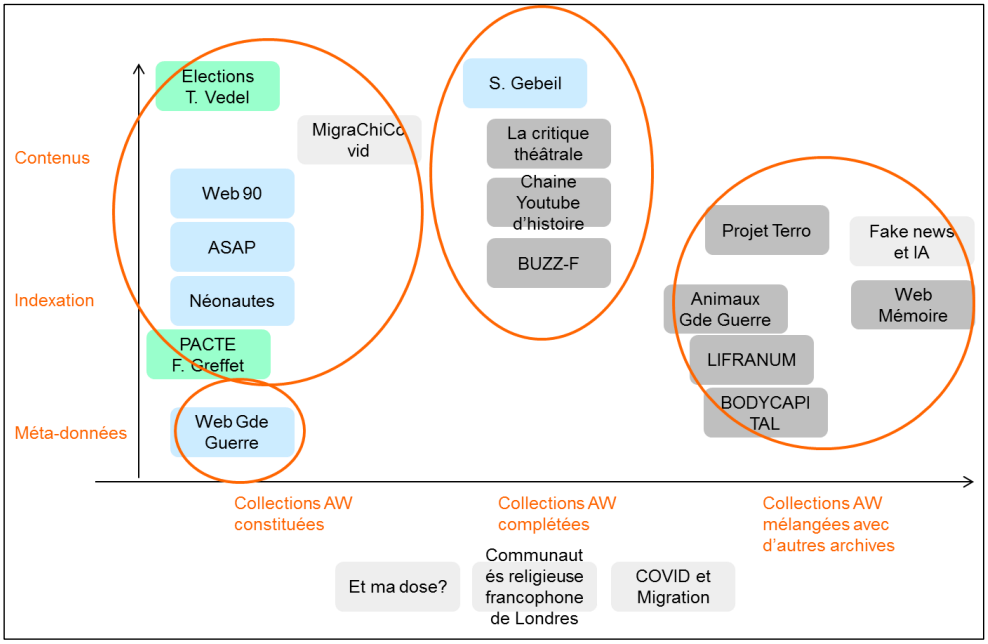


Illustration 3 : Répartition des 20 projets suivant l'axe des collections utilisées et des contenus travaillés

Cette description des projets du DLN montre qu'il y a une certaine continuité entre des projets qui étudient les contenus en utilisant l'indexation et les fichiers WARC, mais qu'une réelle marche semble à franchir pour les projets utilisant des métadonnées de structure du web. On note de plus que les projets de la phase 3 ont plus recours à des corpus mélangés avec d'autres sources, telles que celles de l'INA ou même des Archives nationales (projet Terro) : l'interopérabilité des collections de la BnF avec d'autres sources paraît alors un point essentiel pour garantir l'utilisabilité des AW, et ce aussi bien pour les projets individuels que pour les projets collectifs. L'idée que les collections des AW puissent être un corpus de référence autonome du web vivant ou d'autres sources semble ne pas correspondre aux réalités des recherches en cours.

B.7. Les méthodes mises en place et les outils utilisés

Intéressons-nous maintenant à ce qui est « fait » à ces collections. Dans les phases 1 et 2 décrites précédemment, une relative opposition pouvait apparaître entre les approches reposant sur la fouille de données et les approches historiques critiques.

La fouille de données permet de mener des analyses quantitatives sur les archives web. Un process de traitements est élaboré, testé préalablement sur un petit jeu de données, avant d'être mis en production. Le projet revêt un aspect technique fort et l'hétérogénéité des collections et des données (les archives du web ne sont en rien une base de données structurée) est une des principales difficultés rencontrées dans ce type de projet. Du fait des limites juridiques propres aux archives du web, un ingénieur de recherche attaché au projet peut être amené à devoir travailler dans les espaces de la BnF (phase 2 : Grande Guerre, Neonautes ; phase 3 : Lifranum) pour mettre en place le process. Celui-ci peut aboutir à la production de données dérivées, libres de droit et structurées, qui sont ensuite exportées pour être exploitées par l'équipe projet sur une infrastructure et des outils externes.

L'analyse archivistique repose elle sur un travail critique de la source web étudiée. Le chercheur réalise des enquêtes auprès des producteurs et/ou établir des comparaisons avec les médias TV, films (phase 2 : Sophie Gebeil ; phase 3 : Bodycapital, Arthur De Forges). Le travail de collecte mené par la BnF est essentiel à ce type de travaux, qui peuvent porter sur des collectes courantes, projets ou bien des collectes produites en collaboration étroite avec les chercheurs.

On voit ainsi se dessiner les projets plutôt inspirés d'une démarche de *close reading* des collections (phase 2 : Sophie Gebeil ; phase 3 : Arthur de Farges, Buzz-F), qui participent aux collectes à venir, par rapport à des démarches de *distant reading* des collections (phase 2 : Grande Guerre ; phase 3 : Bodycapital), même si la consultation précise des contenus et la posture critique reste une nécessité (par exemple, des entretiens ont été menés dans le projet Grande Guerre). En effet, cette opposition méthodologique semble s'estomper dans les projets de recherche les plus récents. Ceux-ci étudient des solutions pour mieux articuler les analyses quantitatives et qualitatives (Bodycapital, Lifranum, Buzz-F). L'annotation du corpus, la construction d'un sous-corpus représentatif ou des méthodes d'échantillonnage apportent déjà de premiers résultats encourageants. Ce critère de description semble encore trop spécifique à chaque projet et en évolution pour qu'un axe de distinction puisse se dégager de manière pérenne.

Un autre aspect lié aux démarches projets concerne le pilotage. Si certains projets ont finalement des besoins relativement circonscrits (recherche documentaire avancée, livraison d'un jeu de données, collecte de sauvegarde des contenus étudiés), d'autres nécessitent la mise en place d'une ingénierie du côté de la BnF pour instruire la faisabilité du projet, valider les objectifs communs et la méthodologie, exploiter les jeux de données et valoriser les résultats. Cette ingénierie se traduit très concrètement par la production d'un ensemble de documents de planification et de suivi. En exemple, la collecte menée dans le cadre du projet Bodycapital a dû être phasée dans le planning annuel des collectes du DLN, elle a donné lieu à un compte rendu technique. Par la suite, le projet a bénéficié d'une aide à la fouille de données. Pour répondre à ce besoin, une SolrWayback a été installée sur une infrastructure mise à disposition par la BNUS. Les premiers résultats de recherche, l'usage et les fonctions de l'outil ont donné lieu à plusieurs séances de travail. Au final, pour le suivi du projet, 9 points d'avancement ont été rédigés en 14 mois auxquels s'ajoutent les réunions préparatoires à la signature de la convention. L'appréhension d'un projet en première intention peut donc se résumer à cette question simple : l'équipe de recherche exprime-t-elle un besoin relativement circonscrit ou bien sera-t-il nécessaire de mettre en place une ingénierie de projet et des actions dans la durée (soutenir la montée en compétence des équipes, instruire avec elle la méthodologie, déployer une infrastructure et des outils, participer à la valorisation des résultats) ? Et dans ce deuxième cas, qui prend en charge ce travail et comment le capitaliser ?

B.8. Les « produits » de la recherche

Terminons avec la description des éléments produits par les projets de recherche. L'évaluation de l'apport de connaissances est bien sûr irréductible à un descripteur unique, on pourrait même considérer cette question comme problématique. Les apports sont en effet multiples, et d'ailleurs indirects. En regardant à partir des différents acteurs, on peut lister les « produits » suivants allant d'un horizon de temps du court au moyen terme :

- Pour les chercheurs : montée en compétences sur l'utilisation de matériau numérique (phase 2 : Sophie Gebel ; phase 3 : Arthur de Farges) ; collectes ponctuelles et préservation des données de recherche ; participation à des réseaux de pairs (ex : Valérie Schaefer) ;
- Pour le DLN et la BnF : production de parcours guidés dans les AW participant à la médiation et l'accessibilité des ressources ; enrichissement et qualification des collections pour une archive web thématique conçue comme un ensemble scientifique cohérent et documenté (phase 2 : Sophie Gebel ; phase 3 : Bodycapital, Lifranum) ; développements induits par le DSI et évolutions fonctionnelles ;
- Pour les communautés de recherche : publications académiques et articles de recherches ; formation des étudiants par les chercheur(euse)s participant aux projets ; production d'un jeu de données dérivées à partir des archives du web exploitables sur une autre infrastructure (Grande Guerre, Neonautes, Lifranum).

La consolidation de ces apports et productions mérite d'être observée dans le temps et avec une forme plutôt circulaire, au sens où les productions s'alimentent les unes les autres.

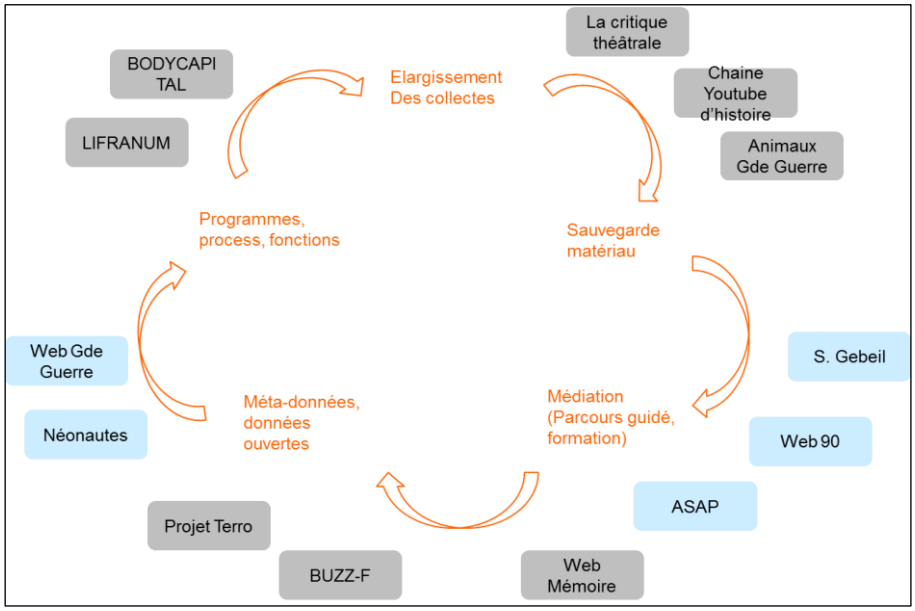


Illustration 4 : Répartition des 20 projets suivant le cercle de production

Cette représentation vertueuse des apports des projets les uns aux autres se fait aujourd'hui au niveau du DLN, par la centralisation des expertises et les apports réciproques. Si une organisation plus décentralisée se mettait en œuvre avec l'augmentation du nombre de projets et l'élargissement des accès, il conviendrait de formaliser le partage d'expérience et les acteurs participant de la circulation des retours. Autant le partage d'expérience se faisait jusque-là par une proximité des IR, des chercheur(euse)s et des équipes DLN lors des projets, autant la porosité des univers académiques et documentaires doit être pensée en intégrant les acteurs locaux. Si des communautés de pratique se créent, seront-elles transverses ou par métier ? Est-ce à la BnF d'assurer l'animation de cette communauté ?

C. Propositions de typologie

C.1. Propositions du Dépôt Légal Numérique de la BnF : une typologie pensée à partir de la notion de besoin

La typologie que nous avons décidé d'esquisser est centrée sur la notion de besoin et de service. Elle permet une approche plus synthétique en soulignant les organisations mises en œuvre pour parvenir aux résultats souhaités en fonction des collections et moyens utilisés.

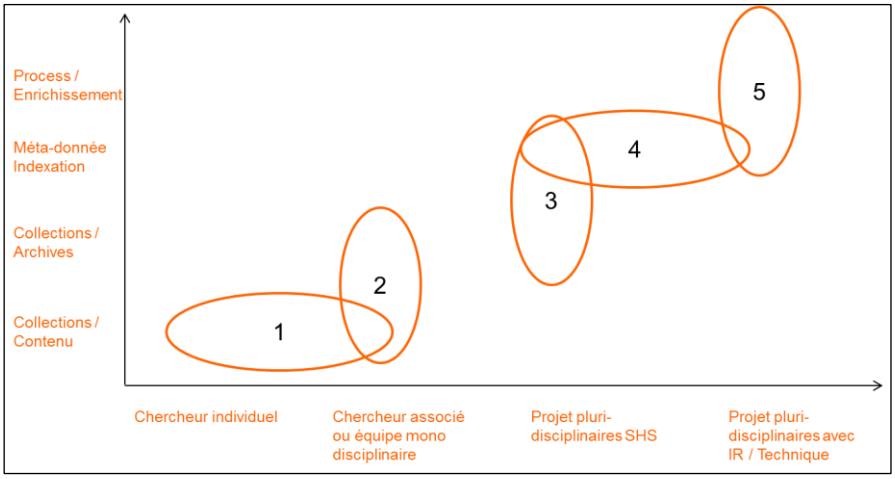


Illustration 5 : 5 types de projet ou besoin DLN

Il faut néanmoins considérer ces besoins de façon dynamique, puisqu'un projet complexe peut avoir plusieurs besoins et qu'une équipe projet peut également souhaiter explorer de nouvelles pistes méthodologiques lors du déroulement d'un projet. Ce découpage permet aussi de mieux saisir la dynamique des projets qui peuvent connaître des phases et des moments tournants, qu'il s'agit néanmoins de prévoir et de phaser au mieux dans l'intérêt du projet et pour sa réussite.

Nous avons également décidé de faire apparaître un besoin de recherche documentaire ponctuel dans cette typologie, bien qu'il s'agisse plutôt d'un service standard. En effet, la particularité des archives du web nécessite de prévoir – même pour une simple consultation éclairée - des temps de présentation des archives du web, des outils de consultation (en particulier des modes de recherche avancée), des différents moyens de rentrer dans les collections (I-Labs, listes API, parcours guidés).

C.2.Recherche ponctuelle

Exemple : Mathilde Henky, doctorante à Paris-Dauphine (politique de dématérialisation de l'État). Recherche documentaire experte dans les archives du web.

Type de service : aide à la recherche documentaire, hors projet

Public visé : chercheur, doctorant, éventuellement M1-M2

Accès au service : libre sur rendez-vous

Objectifs :

- montée en compétence du chercheur
    - brève présentation des archives du web et des outils d'accès
    - fonctions de recherche (URL, labs)
- identification d'archives web pour répondre à la demande
    - recherche pas à pas ou accompagné
    - vérification de la qualité de l'archive
- au besoin, élargissement de la recherche documentaire (réorientation vers d'autres bases ou sources d'information)

Durée de l'accompagnement : ponctuel et court

Mise en œuvre du service : accès à l'application archives du web, montée en compétence des IST sur les fonctionnalités de recherche et la connaissance des collections, production de mémo de recherche.

Évaluation : évaluer le degré d'autonomie acquis, satisfaction par rapport aux résultats trouvés

Amélioration du service : récupération des résultats en autonomie par le chercheur, citabilité des sources

C.3. Travail de recherche archivistique et d'enrichissement des collections

Exemple : Cristina Tosseto, post- doctorante (La critique théâtrale en ligne)

Type de service : aide à la recherche documentaire, constitution de corpus

Public visé : chercheur, doctorant

Accès au service : sur sélection

Objectif :

- montée en compétence du chercheur
    - présentation des archives du web et des outils d'accès
    - fonctions de recherche (URL, labs)
    - présentation du fonctionnement des collectes (process et notion de configuration technique)
    - formation avancée sur la navigation
- contrôle des archives web disponibles
    - par DNS : qualité et historique des archives,
    - fourniture de rapport
- au besoin, intégration du projet dans la politique documentaire des collectes courantes
    - ajout de sélection dans les collectes courantes
    - éventuellement, formation et ouverture des droits sur les outils de sélections
    - contrôle qualité
- valorisation du travail de recherche
    - production d'un parcours guidé
    - communication et/ou publication scientifique

Durée de l'accompagnement : long à accompagnement ponctuel (formation, besoin de collecte, valorisation du travail) – le dispositif « chercheur associé » est adapté à ce type de profil et propose un accompagnement d'un an renouvelable trois fois.

Mise en œuvre du service : accès à l'application archives du web, montée en compétence des IST sur les fonctionnalités de recherche et la connaissance des collections, production de mémo de recherche. Mise en place d'un dispositif contractuel d'accompagnement adapté à la durée du projet.

Évaluation : évaluer le degré d'autonomie acquis, point d'avancement régulier, enrichissement des archives (nouvelles sélections, description et signalement), mesure de la production scientifique (à court et moyen terme)

Amélioration du service : récupération des résultats en autonomie par le chercheur, citabilité des sources, faciliter la vérification de la qualité des archives et la fourniture de rapport depuis l'extérieur

C.4.Fouille et exploitation des jeux de métadonnées

Exemple : Buzz-F, listing des urls répondant à une requête précise (chaîne de caractère « lip-dub » présente dans l'URL, collectes 2011).

Type de service : aide à la recherche documentaire, extraction et fourniture de données ou métadonnées

Public visé : chercheur, doctorant, laboratoire de recherche

Accès au service : sur sélection après évaluation des objectifs, formulation et test de la requête

Objectifs :

- si besoin, montée en compétence du chercheur
    - présentation des archives du web et des outils d'accès
    - fonctions de recherche (URL, labs)
    - brève présentation des formats et jeux de données et métadonnées (liste API, fichiers CDX, fichiers WAT)
- définition de la demande et traduction technique (requête, faisabilité)
    - formulation du besoin et évaluation de la faisabilité technique
    - si besoin, test de la requête à partir de l'application des archives de l'Internet ou à partir d'un échantillon
- la réponse au besoin peut prendre plusieurs modalités selon que le chercheur puisse récupérer directement les métadonnées ou que celles-ci doivent être produites
    - si le chercheur est autonome (liste API)
    - la production de la liste passe par les équipes du DLN, voir d'Humanum et/ou du DSI (listing via Webkit cf urls d'un forum de jeux vidéo pour une capture donnée ; listing produit à partir du big index, des index solr…etc)
    - au besoin, aide pour le traitement des résultats (aide technique et documentaire)

Durée de l'accompagnement : nécessite plusieurs séances de travail, notamment pour bien définir et délimiter le besoin

Mise en œuvre du service : accès à l'application archives du web, montée en compétence des IST sur les fonctionnalités de recherche et la connaissance des collections, production de mémo sur les formats et les listings pouvant être produits, espace de récupération des métadonnées (local pour les jeux de données produits à partir d'une requête, en ligne pour les jeux de données standard cf API)

Évaluation : production d'une base de travail, satisfaction par rapport aux résultats trouvés, facilitation de la fouille de masse dans les collections

Amélioration du service : augmenter le nombre de jeux de données déjà prêt, améliorer la documentation des collectes et la présentation des listes API

C.5. Production et exploitation d'une collecte de référence

Exemple : Bodycapital

Type de service : aide à la recherche documentaire, constitution de corpus, aide à l'indexation

Public visé : laboratoire de recherche

Accès au service : sur sélection après évaluation des objectifs et des moyens mis en œuvre

Objectif :

- montée en compétence de l'équipe projet
    - présentation des archives du web et des outils d'accès
    - fonctions de recherche (URL, labs)
    - présentation du fonctionnement des collectes (process et notion de configuration technique)
    - formation avancée sur la navigation
- préparation de la collecte Corpus de recherche
    - définition des objectifs documentaires en association avec le correspondant concerné (BnF-DCO, BDLI)
    - phasage, préparation et lancement de la collecte
    - production d'un rapport de collecte,
- choix des outils et exploitation de la collecte,
    - présentation des outils disponibles (bibliothèque d'outils du BnF DataLab)
    - mise en place d'une infrastructure de travail et déploiement de l'outil
    - accompagnement et aide à l'exploitation des données
    - production du corpus utile et des résultats de recherche
- valorisation du travail de recherche
    - description et signalement de la collecte
    - amélioration de la bibliothèque d'outils
    - production d'un parcours guidé
    - communication et/ou publication scientifique

Durée de l'accompagnement : long

Mise en œuvre du service : accès à l'application archives du web, montée en compétence des IST sur les fonctionnalités de recherche et la connaissance des collections, production de mémo de recherche. Mise en place d'un dispositif contractuel d'accompagnement adapté à la durée du projet et précis sur les objectifs et engagements des partenaires.

Évaluation : évaluer le degré d'autonomie acquis, point d'avancement régulier, enrichissement des archives (nouvelles sélections, description et signalement), mesure de la production scientifique (à court et moyen terme)

Amélioration du service : accès aux listes API plus aisé, mises à jour régulières de la bibliothèque d'outil, visibilité de l'offre de service.

C.6. Mise en place d'un process de production et d'exploitation d'un corpus

Exemple : Neonaute

Type de service : aide à la recherche documentaire, aide à l'indexation, aide à l'annotation, mise en place de traitement automatique

Public visé : laboratoire de recherche

Accès au service : sur sélection après évaluation des objectifs et des moyens mis en œuvre

Objectif :

- montée en compétence de l'équipe projet
    - présentation des archives du web et des outils d'accès
    - fonctions de recherche (URL, labs)
    - présentation du fonctionnement des collectes (process et notion de configuration technique)
    - présentation des jeux de métadonnées et données (format, jeux tests)
    - éventuellement, présentation des outils de collecte et d'accès
- choix des outils
    - présentation des outils disponibles (bibliothèque d'outils du BnF DataLab)
- mise en place du process
    - définition des objectifs de traitement,
    - mise en place d'une infrastructure de travail
    - accès aux métadonnées et aux données
    - accompagnement et aide à l'exploitation des données
    - éventuellement, validation des traitements sur échantillon
    - production du corpus utile et des résultats de recherche
    - export et récupération des résultats (données dérivées)
- valorisation du travail de recherche
    - amélioration de la bibliothèque d'outils
    - documentation du code
    - communication et/ou publication scientifique

Durée de l'accompagnement : long

Mise en œuvre du service : accès à l'application archives du web, montée en compétence des IST sur les fonctionnalités de recherche et la connaissance des collections, production de mémo de recherche. Mise en place d'un dispositif contractuel d'accompagnement adapté à la durée du projet et précis sur les objectifs et engagements des partenaires.

Évaluation : évaluer le degré d'autonomie acquis, point d'avancement régulier, enrichissement des archives (nouvelles sélections, description et signalement), mesure de la production scientifique (à court et moyen terme)

Amélioration du service : accès aux listes API plus aisé, mises à jour régulières de la bibliothèque d'outil, visibilité de l'offre de service.

D. Préconisations issues des travaux autour des besoins des chercheurs

D.1. Ce que le réseau pourrait apporter (BnF)

L'approche précédente centrée sur les besoins des projets de recherche accompagnés par le DLN permet de mieux retranscrire la trajectoire des projets de recherche entrepris. Au cours de ce panorama, on observe à la fois des caractéristiques de projets qui sont spécifiques à cette phase de démarrage des recherches sur ce matériau et permettent « d'amorcer la pompe » et d'autres caractéristiques qui paraissent pérennes ou à pérenniser pour la généralisation des recherches. Dans les points d'attention permettant de sécuriser les usages à venir, on voudrait en conclusion en souligner deux.

Le premier point revient à la constitution des collections et à leur exploitation. S'il est admis de longue date que l'exhaustivité ambitionnée dans le dépôt légal ne peut s'appliquer au web, la question de la complémentarité des collections AW BnF avec d'autres matériaux devient centrale. Il s'agit à la fois d'expliciter les collectes et la constitution des collections, ce qui peut se faire en s'appuyant sur les compétences documentaires d'un réseau de professionnel(le)s des bibliothèques, et de prévoir les modalités d'enrichissement ou d'export des collections de manière à ce que les « compléments » soient possibles.

Le second point d'attention concerne la nécessité de penser le réseau de support aux projets AW en intégrant les différents métiers et compétences, que ce soit les compétences techniques de la BnF, les compétences d'accompagnement des interlocuteurs des BU, les compétences d'analyse des chercheurs. Dans cet inventaire, il manque l'explicitation de compétences sur la médiation et la valorisation des travaux (autant du côté des espaces documentaires que des espaces de recherche) et se pose la question de l'animation de communautés de pratique : est-ce que les collectifs de chercheurs jouent ce rôle ou parvient-on à maintenir une communauté mixte chercheur(euse) / professionnel(le) de bibliothèques ? Dans le réseau que souhaitent mettre en place les partenaires du projet ResPaDon, malgré les limites posées par le cadre juridique et technique, il serait dommage de limiter l'expérimentation au seul projet ayant un besoin circonscrit. Ce n'est pas forcément l'établissement local, mais plutôt le réseau qui pourrait proposer un accompagnement complet à ces futurs projets de recherche. Les premiers retours d'expérience permettront de mieux définir la place de l'opérateur national auprès des établissements partenaires.

Cette analyse doit bien sûr être couplée avec une mise en perspective des recherches conduites sur les archives du web hors des projets accompagnés par le DLN, ce qui est proposé par la recherche du laboratoire GERiiCO.

D.2. Les besoins exprimés par les chercheurs

Les chercheurs interrogés ont exprimé à la fois :

- des difficultés méthodologiques. En effet, avec ces matériaux de recherche, les chercheuses et chercheurs expérimentent de nouvelles méthodes et des problématiques associées à l'usage des archives du web. Qu'est-ce qu'il faut archiver ? Est-ce qu'il est possible de faire des échantillons qui sont représentatifs ? Quelle est la granularité proposée ? En effet, pour avoir de la profondeur, il faut des capacités importantes et donc une infrastructure performante, ce qui a un coût. Autant de questions méthodologiques sur lesquelles les personnes interrogées souhaitent avoir de l'appui.
- Des difficultés aussi techniques : Comment collecter et analyser des contenus dynamiques (par exemple ceux qui sont générés par des scripts JS ou du Flash ou Silverlight) ? Comment

accéder aux contenus d'une époque en utilisant les mêmes technologies qu'à l'époque ? En effet, l'ambition de ces chercheurs est de contribuer à rendre compte d'une époque, des sujets importants et de la manière dont cela est traité.

- Des besoins de réseautage, d'informations et de formation. Concernant le réseautage, avant même d'instituer une communauté de pratiques regroupant professionnels des bibliothèques et chercheurs, un annuaire des chercheurs et de leurs sujets serait très utile, ou bien une cartographie de l'expertise à la manière des organismes de transfert de technologie qui cherchent à mettre en relation demandeurs et « offreurs » de technologies. Les besoins d'information portent sur l'offre à la fois documentaire et de services dont peuvent bénéficier les chercheurs : bibliothèques de leur institution, BnF, infrastructures de recherche (HumaNum), grands équipements documentaires, etc. Les besoins de formation vont de la maîtrise d'outils pour capturer des données (des vidéos) ou traiter des corpus une fois constitués (outils de visualisation, analyse lexicométrique néanmoins très vulgarisée aujourd'hui) à la mise en œuvre de savoirs sur l'archivage (afin de constituer des corpus réutilisables). De même, nous pouvons souligner qu'une ambiguïté forte persiste entre archive du web et archive sur le web. Mais plus encore certains chercheurs ont exprimé le besoin d'avoir une formation sur Internet et le fonctionnement du Web, au-delà de l'exploitation immédiate qu'ils en ont pour leur sujet de recherche.

Enfin, ces personnes ont aussi des attentes vis-à-vis d'acteurs comme la BnF sur les points suivants :

- Mise à disposition d'outils pour analyser les textes ou les codes sources sur un volume conséquent de données.
- Mise à disposition d'archives dans plusieurs langues. Dans la plupart des cas, les projets de recherche ne sont pas uniquement francophones, il est alors important pour eux d'avoir accès à des archives dans d'autres langues.

L'ensemble des personnes entretenues soulignent la nécessité d'une volonté forte d'institutions, comme la BnF, pour mettre en place un service qui répond aux besoins susmentionnés. Cela nécessite aussi d'investir à la fois dans les infrastructures, les compétences informatiques et méthodologiques, mais aussi sur des aspects légaux comme l'accès dédites archives hors de la BnF ou des BDLI.

D.3. Conclusion

À l'aune des travaux d'enquête et d'analyse conduits, plusieurs constats ressortent.

Ainsi, le soutien que pourrait apporter une offre de services professionnels porterait à la fois sur la recherche d'information (problème de construction de corpus intégrant des sources diverses, dont des AW) et sur les outils (connaissance de l'offre d'outils pour répondre aux divers besoins des chercheurs depuis la collecte et le téléchargement jusqu'au traitement des ressources collectées). Un des problèmes majeurs auxquels sont confrontés les chercheurs est celui de l'obsolescence des formats numériques et des outils de lecture et de collecte. Les recherches, spécifiquement en sciences humaines et sociales, se font souvent sur le temps long et sur la collecte récurrente de documents à des époques différentes. Une illustration de ce problème d'obsolescence est celui de la collecte de vidéos : les outils de fabrication des vidéos par ceux qui les mettent en ligne comme les formats pour les lire et les traiter éventuellement ensuite évoluent rapidement. Or les chercheurs qui travaillent sur la vidéo ne sont pas nécessairement des spécialistes de l'image sur le plan technique.

III. LES PRINCIPAUX RESULTATS ET LEUR VALORISATION

E. Principaux résultats

Le groupe de travail, constitué de professionnels de l'information et des bibliothèques, d'une sociologue et d'enseignants-chercheurs, avait pour objet l'analyse des usages des archives du Web issues du dépôt légal numérique de la Bibliothèque Nationale de France (BnF) à l'occasion de la mise en place d'une capsule d'accès à distance à ces archives depuis les bibliothèques de l'Université de Lille. Le travail avait ainsi plusieurs intérêts :

- Situer les usages de ces archives pour la recherche aujourd'hui alors qu'une étude d'usage avait été réalisée par la BnF dix ans avant le début de ce projet ResPaDon
- Faire connaître ces fonds, en particulier au sein de l'Université de Lille, alors que se mettait en place, au sein du projet ResPaDon, la capsule d'accès à distance
- Tester la capsule dès qu'elle a été opérationnelle

L'intérêt de ce groupe de travail a d'abord été de pouvoir croiser l'expérience de la BnF concernant les projets de recherche dans lesquels elle a été sollicitée depuis 2011 avec des chercheurs ne connaissant pas nécessairement les services de la BnF et dont le récit a été recueilli lors d'une enquête par entretien. Le rapport présente les grandes lignes de l'analyse qui en est issue.

La particularité de l'utilisation des archives du Web par les chercheurs est qu'elle ne repose que très rarement sur une recherche ponctuelle telle que la ferait un journaliste ou un généalogiste : elle est destinée le plus souvent à alimenter des corpus c'est-à-dire un ensemble de sources rassemblées au cours d'un temps plus ou moins long autour d'un thème (ou plutôt une problématique). Ces corpus de sources doivent ensuite être archivés, ce qui suppose un savoir-faire que n'a pas la plupart des chercheurs. Reprendre, modifier, réutiliser des archives Web collectées à un moment donné pour une recherche représentent un ensemble de difficultés auquel tout chercheur est confronté. Les chercheurs peuvent avoir une double position : ils sont utilisateurs des services de la BnF pour compléter ou constituer leur corpus de recherche ; ils peuvent être aussi concepteurs, avec les professionnels des bibliothèques, de parcours guidés sur leur sujet de recherche, ouverts au public, dans le cadre d'un projet pensé comme tel dès l'origine. Mais nombre de corpus relèvent d'initiatives échappant à ce cadre.

La réalisation de corpus Web pour faire une recherche en sciences humaines et sociales ou en histoire des sciences et techniques est généralisée. Il est intéressant de constater l'amplitude des domaines étudiés tels qu'ils dépassent largement celui de l'histoire du Web qui reste un sujet majeur et fondateur comme en témoignent les travaux de N.Brugger et de V.Schaeffer notamment. Dans la mesure où il devient à peu près impossible d'étudier un phénomène social sans articuler l'activité réelle et l'activité numérique qu'il représente, la réalisation de corpus Web est incontournable.
Pourtant elle reste difficile, qu'il s'agisse de réaliser des corpus à partir du Web vivant et/ou à partir du Web archivé (du Web qui n'est plus accessible. L'accès suppose d'abord un déplacement à la BnF ou dans les bibliothèques de dépôt légal imprimeur (Lille dispose d'une telle bibliothèque). Les documents ne sont pas téléchargeables par l'utilisateur et l'existence de ces fonds est peu connue alors qu'Internet Archive (la Wayback machine) l'est par tous les chercheurs. Le droit reste un obstacle majeur et les fournisseurs d'archives autres que les bibliothèques se multiplient. Les éditeurs de bases de données (de presse par exemple) incluent maintenant des archives issues de l'écosystème du Web et des réseaux sociaux, les réseaux sociaux eux-mêmes vendent leurs archives. Des acteurs du « web-listening » travaillent avec de grandes équipes de recherche (en sciences politiques par exemple) et fabriquent des « candidats » aux archives de demain. On pourra se féliciter qu'il existe une diversité

d'acteurs dans ce domaine mais la politique de constitution de ces fonds n'a pas pour tous des buts scientifiques ou patrimoniaux et ne peut remplacer ceux d'une coopération entre bibliothèques et chercheurs. C'est la problématique du « contre-archivage » (Anat Ben-David, Francis Clavert). Ce « contre-archivage » conditionne le type de recherche qu'il est et sera possible de faire aujourd'hui et demain :
« We live in an era of data colonialism in which tech giants appropriate data as territory and act as self-appointed archons in the recording of history. To combat this, we should learn from post-colonial theory and counter-archiving as resistance *» (*Thorsen 2020 citant Ben David dans « Counter-Archiving: Combating Data Colonialism », *Medium* 16/11/2020).
La politique d'acquisition de ces fonds est d'autant plus déterminante que le Web et les réseaux sociaux prennent une part importante dans nos sociétés. Enfin, la difficulté est aussi épistémologique et technique. Elle est épistémologique car il faut pouvoir évaluer l'intérêt et la représentativité des archives trouvées et utiliser des sources complémentaires notamment issues du Web vivant. Elle nécessite aussi une maîtrise technique pour collecter, analyser (outils de text mining et de visualisation), ou archiver et permettre la possible réutilisation de ces matériaux. Cet ensemble de difficultés ne rend pas les chercheurs d'emblée autonomes. En revanche, ils ont conscience de leurs limites pour établir des corpus scientifiquement exploitables.

Outre l'enquête, le groupe de travail a pu faire connaître ces fonds aux collègues et étudiants grâce à la mise en place de la capsule d'accès à distance au sein de l'Université de Lille. En effet, non seulement l'ensemble des chercheurs lillois intéressés a été sollicité pour tester son fonctionnement et alimenter ainsi le groupe de travail sur la capsule mais aussi, certains chercheurs (Laurence Favier et Joana Casenave) du groupe de travail « Usages » ont pu réaliser des études de cas pour la recherche et l'enseignement à partir de la capsule lilloise. L'expérience pédagogique est relatée dans le rapport. Deux études de cas ont été présentées en 2023 au séminaire du laboratoire Gériico le 02/03 puis au colloque ResPaDon du 3 au 5 avril 2023. Elles ont permis de montrer ce que la collecte du Web électoral des élections de 2002 pouvait apporter à l'étude du vote électronique alors qu'il était expérimenté pour la première fois dans des élections nationales. Il a également pu mettre en évidence grâce à des sources issues de sites de presse (collecte « Actualités » des archives du Web) dans les années 2010 à 2012 comment le terme de « féminicide » entrait dans le débat public pour désigner des réalités françaises (et non seulement extérieures à la France). Le projet a ainsi permis aux enseignants du groupe de travail d'associer, tant leurs collègues extérieurs au projet que les étudiants de master, à l'usage de ces matériaux au-delà du seul test technique.

Le test technique de la capsule a également été très utile non seulement pour imaginer quelles fonctionnalités seraient souhaitables pour les chercheurs mais aussi pour révéler toutes les difficultés de l'accès à distance. Marie Cros qui faisait partie des groupes de travail sur les usages et sur la capsule a pu recueillir nos remarques à ce sujet. Nous avons pu aussi montrer aux étudiants que les ambitions de ce type d'accès à l'Université sont un peu différentes de celles de la bibliothèque municipale Jean Lévy de Lille que nous leur avons fait découvrir à l'occasion de ce projet.

F. Valorisation des résultats

L'enquête auprès des chercheurs, le retour d'expérience de la BnF, les expériences pédagogiques, les études de cas à partir de la capsule d'accès à distance lilloise dès qu'elle a été opérationnelle, le test de la capsule sont les principaux livrables ou participation à des livrables (dans le cas des tests techniques) que le groupe a produit. Ceux-ci ont été pensés et programmés lors de réunions régulières (7) à distance durant lesquelles les participants ont échangé régulièrement en dehors des évènements de valorisation des travaux.

L'étude des usages des archives du web s'est également construite par la contribution du groupe de travail sur les usages à la conception de temps d'échanges avec les communautés de recherche. Le

groupe du travail sur les usages a notamment participé de manière étroite à l'élaboration de deux événements :

- La journée de lancement du 17 mai 2021. Ce premier événement du projet ResPaDon avait vocation à dresser un premier état des lieux des usages des archives du web par les communautés de recherche. Les sessions plénières et les ateliers ont été l'occasion d'identifier une première typologie d'usages et d'ouvrir ensemble le débat sur ces usages.

- Le colloque Le web, source et archive du 3 au 5 avril 2023 : placé parmi les événements finaux du projet ResPaDon, ce colloque a eu pour objectif d'interroger la place de la source web dans les pratiques de recherche. Les axes thématiques retenus montrent l'ouverture de la réflexion épistémologique à la source web dans son ensemble. Les communications présentées viennent enrichir et compléter l'identification des usages menée au cours du projet.

  Par ailleurs, un numéro de la revue *Les Cahiers du numérique* destiné à valoriser certaines communications du colloque et à publier d'autres travaux qui n'ont pu être proposés à ce colloque a été diffusé le 13/11/2023. Lien vers l'appel à publication : https://lcn.revuesonline.com/revues/23/LCN_le_web_source_et_archive.pdf

Enfin, le groupe de travail sur les usages a contribué aux communications présentées lors :

- Du colloque annuel de la Maison Européenne des Sciences de l'Homme, DHNord, du 20 au 22 juin 2023 (https://www.meshs.fr/page/dhnord2022)

- Le colloque RESAW (Research Infrastructure for the Study of Archived Web Materials) « Exploring the archived web during a highly transformative age » (5 et 6 juin 2023, Mucem, Marseille) : « Session 8.B : A network to develop the use of web archives : three outcomes of the ResPaDon projects »

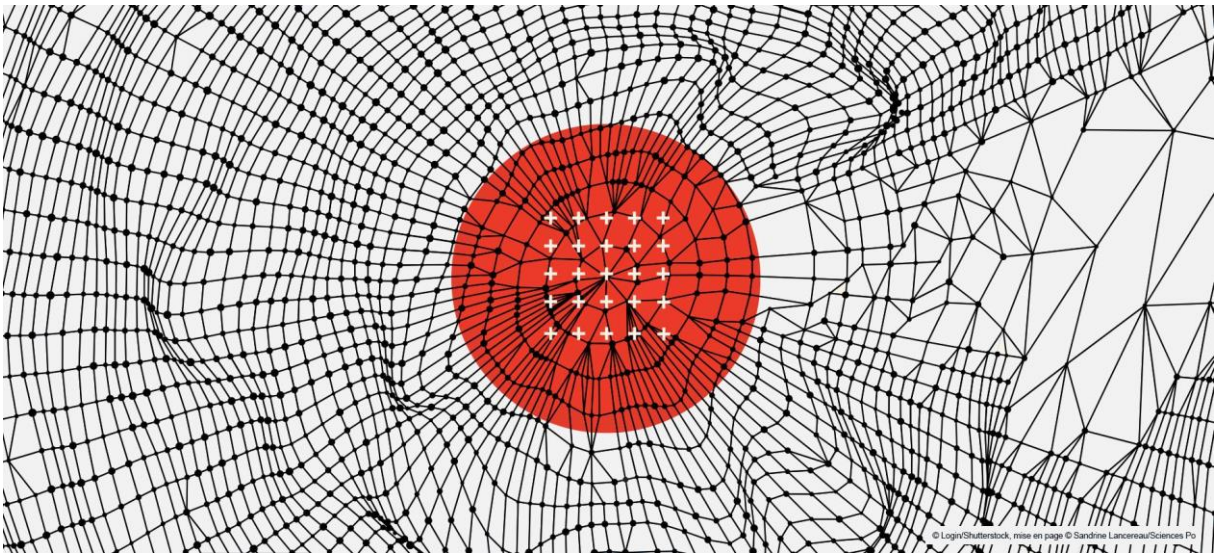
Annexe 3 :

L’expérimentation capsule au cœur du projet ResPaDon :

bilan et recommandations

RES
PA
DON

L’expérimentation capsule au coeur du projet ResPaDon : bilan et recommandations



Contenu

Introduction

Le projet ResPaDon avait pour objectif de développer de nouvelles formes d'exploitation des collections numériques de la BnF, et en particulier des archives du web, dans les établissements de l'enseignement supérieur. Le partenariat entre la BnF, Sciences Po, le Campus Condorcet et l'Université de Lille a permis de développer des services innovants sur ces collections.

Le présent document dresse le bilan de l'une des expérimentations conduites au titre du projet, appelée "expérimentation capsule". Elle consistait à proposer dans les murs de l'Université de Lille un accès distant aux collections de dépôt légal du web conservées à la BnF et une offre de services destinée à faciliter l'usage de ces collections.

Le déploiement de la capsule a consisté plus précisément en :

- la définition d'un cadre juridique pour l'expérimentation,
- la mise en place d'un accès distant sécurisé aux collections de dépôt légal du web de la BnF d'un point de vue technique et logistique,
- le développement et le déploiement d'outils et d'applications permettant la consultation, l'exploration, la fouille de texte et de données,
- la conception et la mise en place de services de médiation destinés à faciliter l'appropriation de ces collections, incluant la formation de "médiateurs" accueillant les chercheurs,
- l'organisation des tests du dispositif par les chercheurs de l'Université de Lille.

Contexte et objectifs de l'expérimentation

Le projet ResPaDon est né du constat que les archives du web sont sous-exploitées par les chercheurs. Un très large spectre de questions de recherche pourraient en effet faire appel aux archives du web, à titre de source principale ou complémentaire à d'autres. Le postulat qui a inspiré l'expérimentation capsule est que cette situation s'explique par un "coût d'entrée" dans les archives du web trop élevé pour les chercheurs qui commencent à s'y intéresser. Le déploiement de la capsule à l'Université de Lille a eu pour objectif d'encourager des usages académiques plus larges des archives du web en réduisant ce coût, considéré dans ses différents aspects :

- **coût d'entrée juridique** : il est dû au fait que les archives du web, autrement dit les contenus collectés au titre du dépôt légal du web par la BnF et ses partenaires, sont des contenus sous droit et soumis à des restrictions d'accès. Le Code du patrimoine stipule que les archives du web sont uniquement consultables dans les salles recherche de la BnF, ainsi que dans un réseau de bibliothèques de dépôt légal imprimeur (BDLI) précisément listées dans un arrêté de 2014. Ces bibliothèques sont, à l'exception notable de la Bibliothèque interuniversitaire de Strasbourg, des bibliothèques de lecture publique. Pour consulter les archives du web, il est donc nécessaire de se déplacer dans l'un de ces lieux, alors que les pratiques de recherche actuelles ont tendance à privilégier un accès nomade, immédiat et distant, aux données. La quasi-absence de point d'accès à ces collections en Université a ainsi été identifiée comme un réel obstacle à leur exploitation plus large en contexte académique. Elle explique en grande partie leur faible visibilité et une relative méconnaissance de l'existence de

ce matériau par les chercheurs de l'ESR. L'expérimentation capsule a constitué une première réponse à ce problème, en offrant un point d'accès aux archives du web dans l'emprise d'une grande Université pluridisciplinaire, au plus près des équipes de recherche. Il est important de noter qu'il s'agissait avec ce dispositif expérimental de proposer, en accord avec la transposition récente en droit français des exceptions au droit d'auteur, un environnement sécurisé permettant la fouille de texte et de données sur des collections sous droit.

- **coût d'entrée méthodologique :** les archives du web sont une source relativement complexe à exploiter. Les mobiliser dans un travail scientifique requiert d'en comprendre la fabrique, c'est-à-dire les modalités, documentaires et techniques, de constitution de ces collections. D'un point de vue technique, ces archives sont collectées de manière semi-automatisée par des robots logiciels de collecte ou *crawlers* qui copient les pages web et les éléments qui les composent, puis accessibles dans une application qui reconstitue un contexte de navigation similaire au contexte orignel. D'un point de vue documentaire, les archives du web sont le fruit d'un modèle mixte, qui combine l'agrégation de listes de domaines français fournies par les gestionnaires de noms de domaine (collecte dite "large", conduite une fois par an), et une sélection de contenus à archiver plus fréquemment et plus en profondeur par un large réseau d'experts à la BnF et dans des établissements partenaires, parmi lesquels des laboratoires de recherche. La collecte poursuit ainsi une logique d'échantillonnage raisonné, le web, en constante évolution, étant par nature impossible à archiver dans sa totalité. L'archive web fige un flux, offre une image ou une trace de ce qu'était le web à un moment donné. La singularité de ce matériau et sa nature d'artefact numérique a des conséquences sur les méthodes et les outils qui permettent de l'étudier, qui sont à de nombreux égards spécifiques. La capsule ResPaDon a eu pour objectif de réduire ce temps d'acculturation à l'archive, de faciliter la découverte et la compréhension d'un nouveau matériau de recherche et des méthodes qui permettent de l'étudier, en proposant dans l'enceinte de l'Université des services d'accompagnement à la prise en main des collections. C'est dans cet esprit qu'ont été conçus et mis en œuvre pour les besoins de l'expérimentation un accueil par les personnels du SCD de l'Université de Lille et des services de médiation.

- **coût d'entrée technique :** celui-ci tient en premier lieu à la spécificité et à la complexité des formats de stockage des archives du web, le format WARC, décrit par la norme ISO 28500, qui compile l'ensemble des données hétérogènes collectées (code source des pages en html, les feuilles de style, les images, etc.) ainsi que des métadonnées dans des fichiers valises. Il s'explique également par la volumétrie des données concernées. Croiser une lecture qualitative des documents et des analyses quantitatives requiert ainsi un outillage et des compétences numériques spécifiques, un prérequis commun aux archives du web et aux autres matériaux mobilisés par les Humanités numériques. L'hétérogénéité des types de données et de formats trouvés sur le web ainsi que le caractère universaliste des collections a d'autre part pour conséquence que les méthodologies et les outils de constitution de corpus sont spécifiques aux différentes disciplines, voire nécessitent d'être adaptés en fonction des questions de recherche. Enfin, favoriser la découvrabilité des archives du web conservées à la BnF, qui ne sont pas intégralement indexées en plein texte, mais dans leur majorité accessibles

uniquement par une recherche par URL, est un défi de taille. La capsule déployée à Lille a eu pour objectif de pallier ce coût d'entrée technique, en proposant à titre expérimental des outils d'aide à l'exploration enrichie des données hétérogènes qui composent les archives du web (texte, image, vidéo, etc.), ainsi que des outils d'aide à la fouille de texte et de données.

La capsule déployée à Lille est ainsi une réponse aux nombreux défis à relever pour permettre une utilisation plus intensive des archives du web par les chercheurs. L'objectif est d'encourager une très large palette d'usages, de l'utilisation ponctuelle des archives en complément d'autres matériaux de recherche à son utilisation comme source principale, de l'analyse qualitative d'un nombre restreint de captures web aux approches quantitatives de type fouille de texte et de données. En ce sens, la capsule prolonge et étend la démarche de facilitation et de promotion des usages innovants autour des collections numériques conservées à la BnF entreprise avec le BnF DataLab, ouvert en octobre 2021, et peut être considérée comme un DataLab hors les murs. Les deux dispositifs s'inspirent et se nourrissent l'une l'autre.

Un dispositif expérimental évalué au fil de l'eau et testé en conditions réelles pendant un an

L'expérimentation capsule a consisté à mettre en place et à tester en conditions réelles un prototype d'offre de services et d'outils conçu pour encourager les usages de collections d'intérêt national. L'objectif final de l'expérimentation était de formuler des préconisations pour améliorer le dispositif et penser sa reproductibilité dans d'autres universités.

L'évaluation du dispositif s'est faite au fil de l'eau, par une attention portée aux difficultés rencontrées et par la documentation des circuits et processus de mise en œuvre. Les applications et services proposés dans la capsule ont de plus été testés en conditions réelles par 49 chercheurs et étudiants de l'Université de Lille entre juin 2022 et septembre 2023.

Les retours de tests, recueillis par les médiateurs sous forme d'entretiens informels, fournissent un éclairage intéressant sur les outils et services proposés et les améliorations à leur apporter. Ils viennent ainsi utilement compléter les observations faites au fil de l'eau par les différents acteurs impliqués dans le déploiement de la capsule.

Le présent bilan s'appuie ainsi à la fois sur les leçons tirées de la mise en œuvre et sur l'analyse des retours de tests. Une attention particulière est portée aux moyens matériels et humains impliqués ainsi qu'aux compétences requises dans la mise en œuvre de la capsule. Le bilan dégage également dans cette perspective des recommandations pour améliorer, pérenniser le dispositif de capsule et garantir sa reproductibilité.

I) Une emprise BnF au sein de l'Université de Lille

Une convention d'application BnF / U-Lille

En complément de la convention cadre et multipartite ResPaDon associant l'ensemble des partenaires du projet, une "convention d'application" signée par l'Université de Lille et la BnF a fixé le cadre juridique du déploiement de la capsule.

Cette convention :

- permet le déploiement, à titre expérimental, d'un point d'accès distant aux collections de dépôt légal du web conservées à la BnF dans l'Université de Lille, qui ne fait pas partie des établissements habilités à fournir un accès aux archives du web listés dans le code du Patrimoine (la liste des bibliothèques de dépôt légal imprimeur (BDLI) concernées figure dans un arrêté de 2014). La convention d'application crée à cette fin une emprise BnF à l'Université de Lille qui met à disposition de la BnF pour occupation, à titre grâcieux et temporaire, deux espaces physiques (deux salles) où se déroulent les tests ;
- encadre l'utilisation des collections de dépôt légal du web et des applications permettant de les explorer, les analyser et les fouiller. Les principes retenus sont les mêmes que ceux en vigueur dans les salles de recherche de la BnF et le BnF DataLab et interdisent le téléchargement massif de données. La fouille de texte et de données est permise sur un corpus restreint et au sein d'un environnement sécurisé coupé du web ;
- encadre le déroulement des tests, le recueil et l'exploitation du résultat des entretiens conduits par les médiateurs. Aux termes de la convention, l'Université de Lille est responsable de la supervision des tests réalisés par les chercheurs qu'elle emploie ou héberge. La convention encadre le recueil et l'usage des données personnelles collectées à cette fin ;
- propose une Charte à signer (sous forme imprimée) par laquelle les testeurs acceptent les règles d'utilisation des données et s'engagent à fournir des retours sur leurs tests.

L'élaboration de cette convention d'application et de la Charte a mobilisé les services juridiques des deux établissements, et quatre membres du projet. Des rendez-vous spécifiques avec le Data Protection Officer de l'Université ont permis de préciser l'utilisation des données personnelles recueillies lors des tests et a été suivi par une déclaration de traitement dans le registre de l'Université. Le travail effectué constitue une excellente base pour proposer à l'avenir une convention type applicable à d'autres universités.

Recommandations

Recommandation n°1 : Faire évoluer le cadre législatif et réglementaire pour permettre l'implantation de points d'accès aux collections de dépôt légal du web dans les emprises des services communs de documentation des établissements de l'ESR.

Cette évolution est nécessaire pour pérenniser la capsule mise en place à Lille et permettre le déploiement de capsules dans d'autres établissements. Ce point rejoint les préconisations n°7 et n°8 du projet ResPaDon, "Faciliter l'accès et la réutilisation des archives du web en faisant évoluer les conditions réglementaires actuelles" et "Déployer et pérenniser des capsules d'accès aux archives du web dans des établissements de l'enseignement supérieur et de la recherche"

<u>Recommandation n°2</u> : Rédiger une convention type associant les établissements accueillant des capsules et la BnF et précisant les obligations des deux parties.

Cette convention peut consister en une adaptation de la "Convention d'application" pour la rendre plus générique.

Afin de tirer parti des enseignements de l'expérimentation, les améliorations suivantes pourraient être intégrées à cette convention type et/ou au cadre législatif :

- **étendre la définition du public habilité à consulter et fouiller les collections de dépôt légal du web** : la définition des destinataires du dispositif doit être étendue de façon à prévoir explicitement les usages pédagogiques des archives du web dans la Convention, ceux-ci étant indissociables des usages recherche. L'ensemble de la communauté universitaire doit être habilitée à consulter et explorer les collections de dépôt légal du web. Il sera à cette fin utile de prévoir une procédure d'accréditation des chercheurs déléguée aux établissements accueillant des capsules ;
- **simplifier les conditions de recueil et de transfert des données personnelles des usagers de la capsule** : les tests conduits dans le cadre de l'expérimentation nécessitaient un recueil de données personnelles en vue d'exploiter les résultats. Dans la perspective de points d'accès pérennes, il n'y aurait pas lieu de prévoir le recueil et les traitements de données personnelles des usagers de la capsule. Les traitements effectués étant plus génériques, ils pourraient entrer dans le cadre déjà défini pour l'utilisation et la connexion aux équipements et réseaux informatiques de l'Université accueillant une capsule. Les éventuelles données personnelles recueillies par la BnF lors de l'accès à ses applications restent à préciser en fonction du système d'accès distant qui sera retenu, encore en cours de définition et spécification au terme du projet ;
- **dématérialiser la charte d'utilisation des données et services numériques de la BnF** : l'acceptation des conditions d'utilisation des données et applications en ligne peut se faire au moment de la connexion au système d'accès distant, sur le modèle de ce qui existe dans les bibliothèques de dépôt légal imprimeur.

<u>Recommandation n°3</u> : Implanter des points de consultation dans les espaces fréquentés par les chercheurs

<u>Recommandation n°4</u> : Décrire, préciser et faciliter les usages qui peuvent être faits des différents types de données relatifs au dépôt légal du web mis à disposition dans les capsules.

La mise en oeuvre de cette recommandation peut se traduire par la rédaction d'un guide à l'usage des chercheurs sur les conditions de consultation, de traitement et d'exploitation des données disponibles dans la capsule dans un cadre de recherche, et/ou d'une Foire aux questions.

Les entretiens montrent que les incertitudes concernant les conditions de citation, d'export et de réutilisation des données et résultats d'analyse dans les publications scientifiques est un frein aux usages académiques des archives du web, dès la toute première exploration. Les nombreuses questions posées par les chercheurs pendant les tests montrent qu'il convient de mieux spécifier les régimes encadrant l'usage des différents types de données (données collectées, métadonnées et données dérivées techniques ou documentaires, données transformées ou enrichies) et de distinguer les différents usages et types de contextes (accès, analyse et exploitation dont TDM, copie privée, publication et diffusion). La clarification des usages permis rejoint la préconisation n°4 du projet ResPaDon,

“Inscrire les sources web, archives et web vivant, dans l’évolution des pratiques de recherche et dans l’ouverture des processus et résultats de la recherche”.

Un guide juridique ou une Foire aux questions à destination des chercheurs seraient de nature à réduire ces incertitudes et permettrait d’illustrer par des exemples concrets les différents cas de figure. Ce guide pourrait préciser les points suivants :

Accès aux données et conditions de consultation, (en d’autres termes, ce qui peut être sorti de la capsule ou non) :

- décrire les régimes de propriété intellectuelle régissant les différents types de données : distinguer notamment les données dérivées et métadonnées d’ordre documentaire ou technique, libres de droit, diffusées ou diffusables sous licence EtaLab, qui peuvent être consultées en dehors de la capsule et réutilisées librement d’une part, et les données collectées au titre du dépôt légal du web proprement dites d’autre part, qui sont soumises au droit d’auteur, dont la consultation se peut se faire que dans la capsule ;

- en ce qui concerne les données collectées au titre du dépôt légal elles-mêmes (données soumises au droit d’auteur), distinguer l’export massif des données hors de la capsule, interdit, d’une part, des copiés-collés ou copies d’écran ponctuels des contenus consultés que peuvent réaliser les chercheurs pour pouvoir les analyser, et qui peuvent être exportés hors de la capsule. Cette distinction n’était pas assez claire pour les testeurs et de nombreuses questions ont porté sur ce point.

Analyse, exploitation et traitement des données, incluant l'exploration et la fouille de texte et de données : la question de savoir comment permettre l’exercice de la fouille de texte et de données autorisée par l’exception TDM sur des corpus sous droits en accès restreint tels que les archives du web est complexe. La capsule constitue l’une des réponses à cette question en ce sens qu’elle propose un environnement coupé du web contenant des outils et applications d’exploration et d’analyse, et d’aide à la fouille de texte et de données permettant d’effectuer des traitements sur les données. Les traitements sont réalisés sur une machine virtuelle sécurisée et coupée du web. Des questions demeurent néanmoins sur le statut des données ainsi produites, et de leurs conditions de réexploitation dans le cadre de publications académiques.

Réutilisation, diffusion et publications des résultats de la recherche, qu’il s’agisse des données transformées et enrichies, de corpus intermédiaires ou des résultats eux-mêmes : la convention précise que les testeurs peuvent demander l’export des données issues de la recherche, c’est-à-dire “ produites avec les outils d’analyse, de requêtage, de programmation ou de visualisation livrés dans l’environnement sécurisé”, stipule que “les données ainsi enrichies et transformées constituent le résultat original de la recherche” et que “les chercheurs pourront demander à l’issue de l’expérimentation l’export sécurisé de leurs résultats” en envoyant un mail. Toutefois cela ne prend pas en compte tous les cas de figure, les types de corpus intermédiaires ou données agrégées ou transformés sont divers, et une large variété de cas de figure existe que le guide juridique pourrait exposer et illustrer :

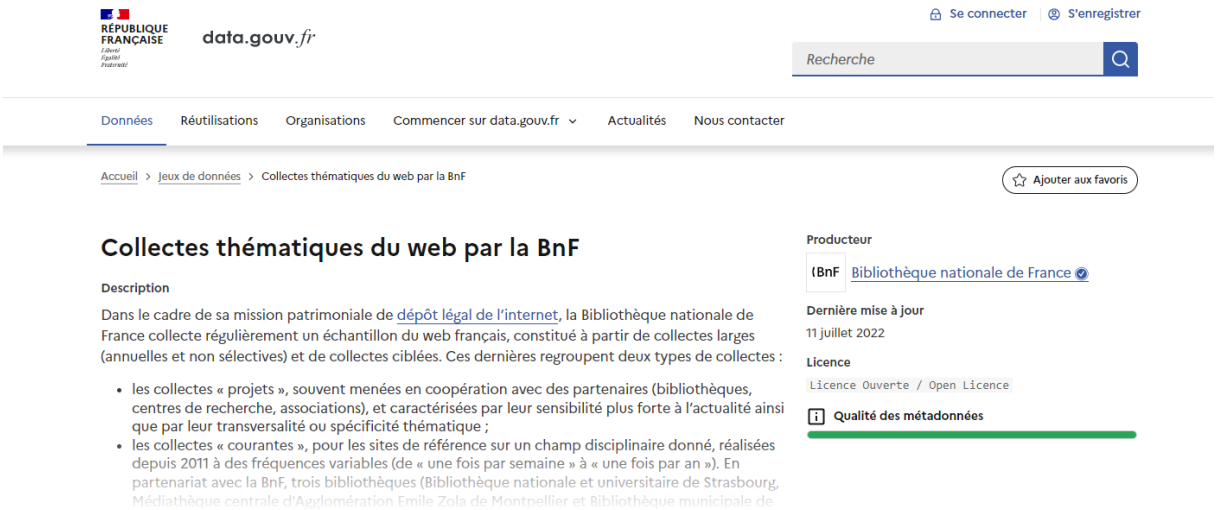
- certaines de ces données transformées ou enrichies sont entièrement la propriété du chercheur qui les a produites, on peut penser notamment au graphe de liens

catégorisant différents types d'acteurs produit à partir d'un export des liens hypertextes présents dans les pages ou encore aux résultats d'analyse d'occurrences des termes dans un corpus produits avec le logiciel Iramuteq (Gephi et Iramuteq étaient proposés au sein de la capsule) : dans ce cas, les données intermédiaires, le fichier contenant l'ensemble des liens présents dans les pages, est une donnée libre de droit ; le résultat produit (graphe de liens) est, au même titre que son analyse, le résultat original de la recherche et il appartient au chercheur qui l'a produit de fixer les conditions de sa ré-exploitation ;

- en revanche, dans le cas d'une analyse sémantique effectuée sur le plein texte des pages, le plein texte des pages web collectées enrichi des entités nommées ou d'autres traitements sémantiques est à la fois une donnée transformée et un corpus intermédiaire qui incorporent le travail d'analyse du chercheur, en d'autres termes une donnée de la recherche, et une donnée soumise au droit d'auteur détenue par les producteurs de contenus web.

Diffusion, notamment dans les publications académiques :

- le guide pourrait proposer de bonnes pratiques de citation des archives du web par l'utilisation des URL pérennes ;
- les questions ont montré qu'il fallait préciser ce qu'il était possible de faire avec les captures d'écran ou les copiés-collés de contenus textuels exportés hors de la capsule : si celui-ci est autorisé pour usage privé, les règles de reproduction dans une publication sont plus complexes : la courte citation de contenus textuels d'une page web entre dans l'exception recherche, une ambiguïté demeure sur les captures d'écran des archives, y compris en basse définition ;
- la façon d'assurer la reproductibilité de la recherche au-delà même du partage des corpus intermédiaires pourrait également faire l'objet de recommandations et d'exemples : par exemple, publication des scripts ayant servi à extraire les corpus ou liste des URL des captures utilisées, pour rapprocher l'usage des archives du web d'un usage FAIR.



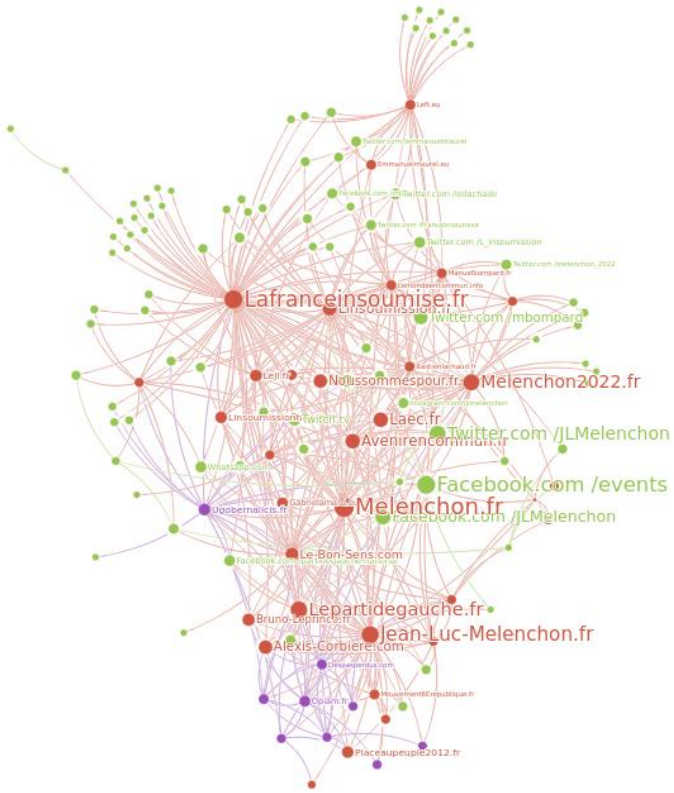
Exemple de métadonnées documentaires diffusées sous licence EtaLab : liste des sélections faites dans le cadre des collectes ciblées par la BnF

solrwayback_linkgraph_2023-06-15_16-43-18.csv

```
jbr2002.com,mdbafrance.com
fr.fm,free.fr,netscape.com
verts87.org,eu.org,les-verts.org,verts-limousin.org
aschieri.net,microsoft.com,netscape.com
corinne-lepage.com,esseclive.com,gay.com,lapolitique.com,professionpolitique.net
franceelections2002.com,alainmadelin.com,elysee.fr,presidentielles.org,vumetrix.com
olivierbesancenot.org,lcr-rouge.org,uzine.net
lipietz2002.net,assemblee-nationale.fr,elections-legislatives.fr,electionsverts.org,eludefrance.net,perline.org,uzine.net
grandmanitou.net,uzine.net,weborama.com,weborama.fr,xiti.com
jeanclaudethomas.com,legispack2002.com
les-verts.org,etatsgeneraux.org,les-verts-europe.org,ouvaton.org,sgdg.org,souris-verte.net,voila.fr
francisdemay.org,legispack2002.com
ouvaton.org,amisdelaterre.org,conso.net,ecoloparade.org,greenpeace.fr,les-verts.org,nedstatbasic.net
verts-noisy.org,apache.org,gnu.org,mysql.com,ouvaton.net,php.net,postnuke.com,verts-sylvieduffrene.org,xiti.com
amaurynardone.net,legispack2002.com
2002enseignement-recherche.net,adobe.fr

```

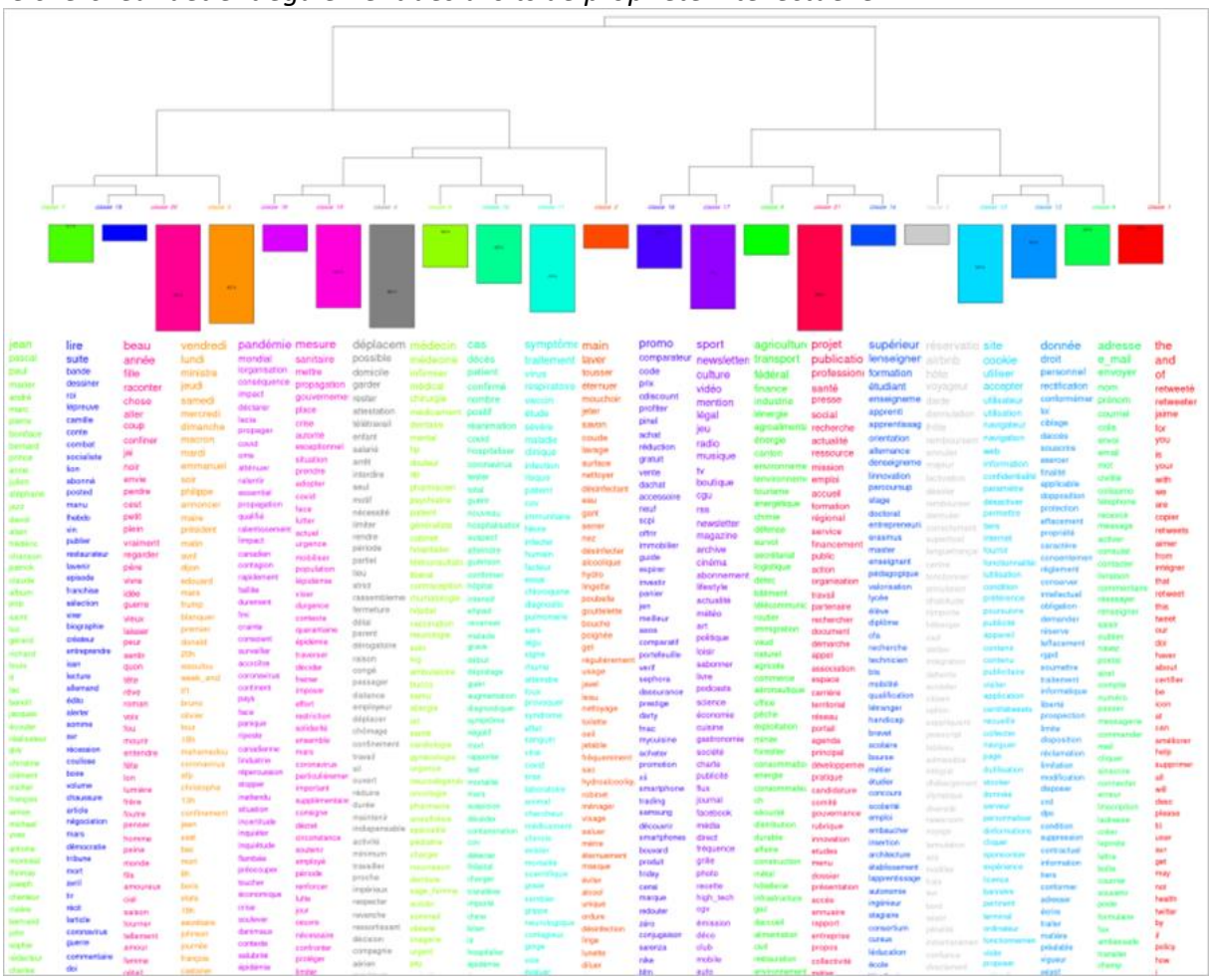
Fichier texte contenant les liens entrants et sortants de chaque site pouvant servir à produire un graphe de liens dans un logiciel dédié, par exemple Gephi, disponible dans la capsule. Donnée de la recherche / corpus intermédiaire, libre de droit, reproductible dans une publication



Graphe produit par les analyses (catégorisation des types d'acteur) : résultat original de la recherche, dont le chercheur détient les droits de propriété intellectuelle


```
"tokens":"Archives de la catégorie : ' Article 17 CEDH ' Le rédacteur en chef de deux journaux publiés en Azerbaïdjan , Eynulla Fatu
"lemmas":"archives de le catégorie : ' article 17 cedh ' le rédacteur en chef de deux journal publier en azerbaïdjan , eynulla fatul
"lemmas_tags":"archives/NOUN__Gender=Fem|Number=Plur de/ADP___ le/DET__Definite=Def|Gender=Fem|Number=Sing|PronType=Art catégorie/NC
```

Plein texte enrichi par des analyses en TAL : les producteurs de contenus détiennent les droits sur le plein texte des pages, autorisation nécessaire pour le reproduire ; le travail d'analyse sémantique fait qu'il s'agit d'une donnée enrichie, donnée dérivée de la recherche, sur laquelle le chercheur détient également des droits de propriété intellectuelle



Analyse d'occurrence des termes produite avec Iramuteq : le corpus intermédiaire ayant servi de bases aux analyses est une donnée sous droit, le résultat est exportable et librement diffusable

<u>Recommandation 4 (suite)</u> : Transcrire dans l'architecture des outils cette variété d'usages : en effet, les tests ont montré que la confusion entre ce qu'il est techniquement possible de faire et ce qu'on a juridiquement le droit de faire entrave les usages des archives du web. Il est ainsi souhaitable que les fonctionnalités embarquées dans les outils d'exploration intègrent la variété d'usages autorisés pour les différents types de données : outils de copiés-collés, mise en accès libre des métadonnées techniques et documentaires hors droit, possibilité d'exporter les données issues de la recherche au fil de l'eau.

II) Un service d'accès distant aux collections de dépôt légal du web

Deux solutions techniques

L'accès aux collections de dépôt légal du web depuis l'Université de Lille a nécessité la mise en place d'un environnement d'accès distant permettant la recherche, la consultation et la manipulation des collections tout en garantissant la sécurité des systèmes d'informations de la BnF et de l'Université de Lille ainsi que la sécurité des collections.

Deux solutions techniques ont été expérimentées : la solution inWebo et la solution WALLIX. Les deux solutions permettent d'authentifier les accès externes au SI de la BnF et la mise à disposition d'un environnement de travail sur les archives du web via un mécanisme de déport d'affichage : les utilisateurs manipulent depuis leur navigateur local des applications qui s'exécutent en réalité sur un serveur situé à la BnF.

La solution inWebo

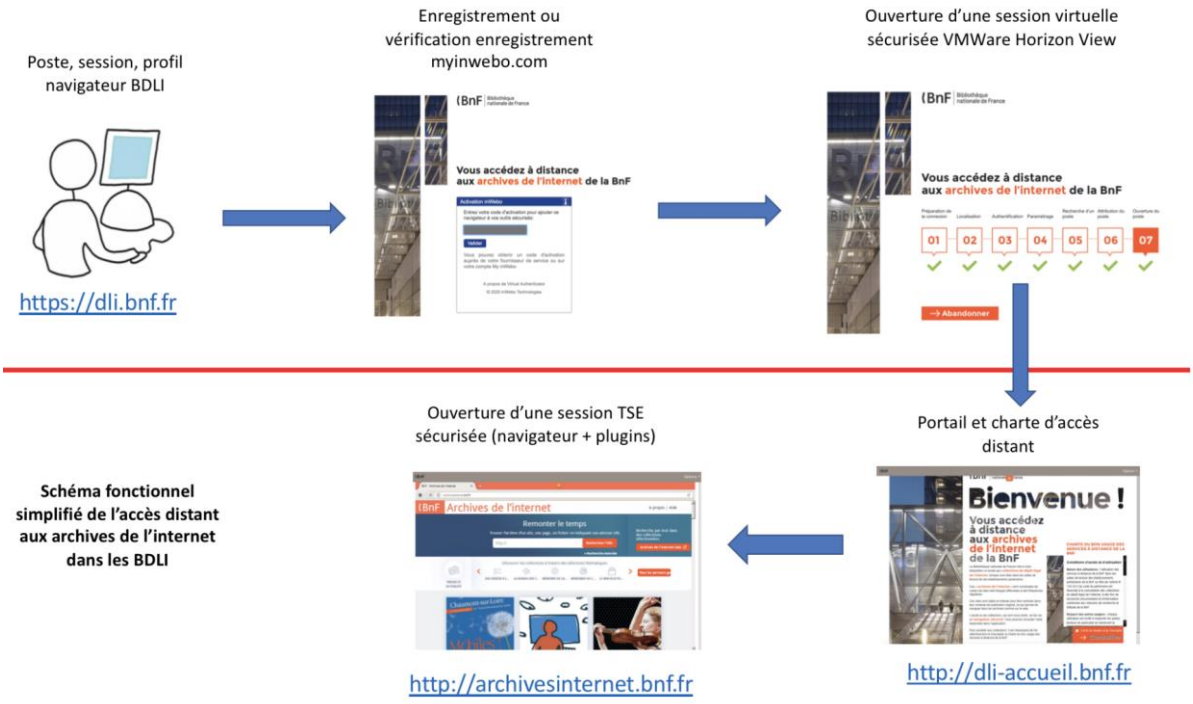


Schéma fonctionnel simplifié de l'accès distant aux archives de l'internet avec la solution inWebo

Du point de vue de l'Université de Lille, la mise en place de cette solution a nécessité la création d'une session utilisateur générique sur chacun des deux PC dédiés au projet ResPaDon, l'installation de l'extension Chrome inWebo Helium Backup et l'enrôlement initial du poste, de la session utilisateur et du profil du navigateur local par l'intermédiaire d'un jeton attribué par un administrateur DSI de la BnF via le service en ligne myinwebo.com de la société TrustBuilder.

Une fois l'enrôlement initial terminé, chaque utilisateur qui veut consulter les archives se connecte au service web https://dli2.bnf.fr et ouvre une première session virtuelle sécurisée à l'intérieur du navigateur local. S'il valide la "charte du bon usage des services à distance de la BnF", il peut ouvrir une seconde session virtuelle qui lui permet d'accéder à l'ensemble des collections avec l'application "Archives de l'internet" et aux quelques collections indexées en plein texte avec l'application "Archives de l'internet Labs". L'application s'ouvre dans un navigateur de couleur orange qui est réservé à la consultation des archives et coupé du web vivant.
En conformité avec la politique de sécurisation des accès aux données relevant du dépôt légal du web et soumises au droit d'auteur, il est impossible de télécharger des contenus archivés depuis cet environnement. Il est possible de récupérer des références et de courts extraits de texte via un système de presse papier.

Du point de vue de la BnF, cette solution repose sur une infrastructure technique complexe basée sur le service en Saas myinwebo.com, le logiciel VMWare Horizon View et un ensemble de serveurs et qui interviennent dans le processus d'ouverture des sessions distantes à l'intérieur du navigateur local.

Matérialisation des différentes étapes d'ouverture d'une session distante avec la solution inWebo

Deux comptes inWebo et deux machines virtuelles ont été créés pour l'Université de Lille.

Utilisé depuis 2014 pour l'accès distant à l'application Archives de l'internet dans les bibliothèques de dépôt légal imprimeur (BDLI), ce dispositif a eu plusieurs défaillances pendant la durée du projet (problèmes réseau, problèmes de licence, disparition du jeton d'enrôlement) qui ont entraîné des remontées d'incidents par les médiateurs et des interventions techniques des administrateurs DSI de la BnF.

La solution WALLIX

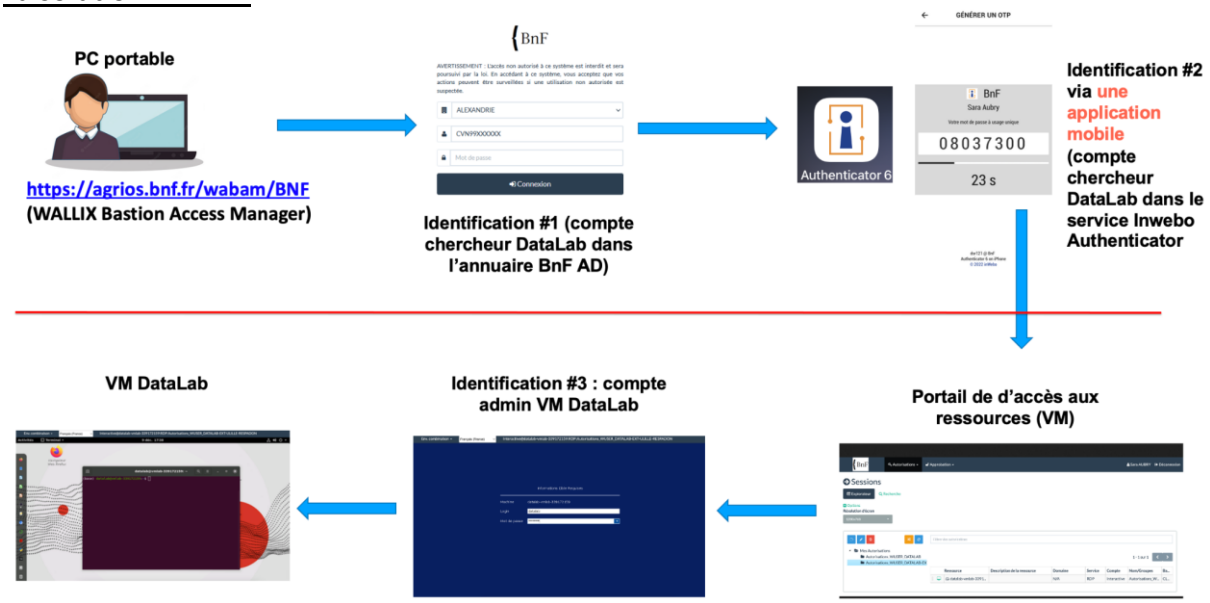
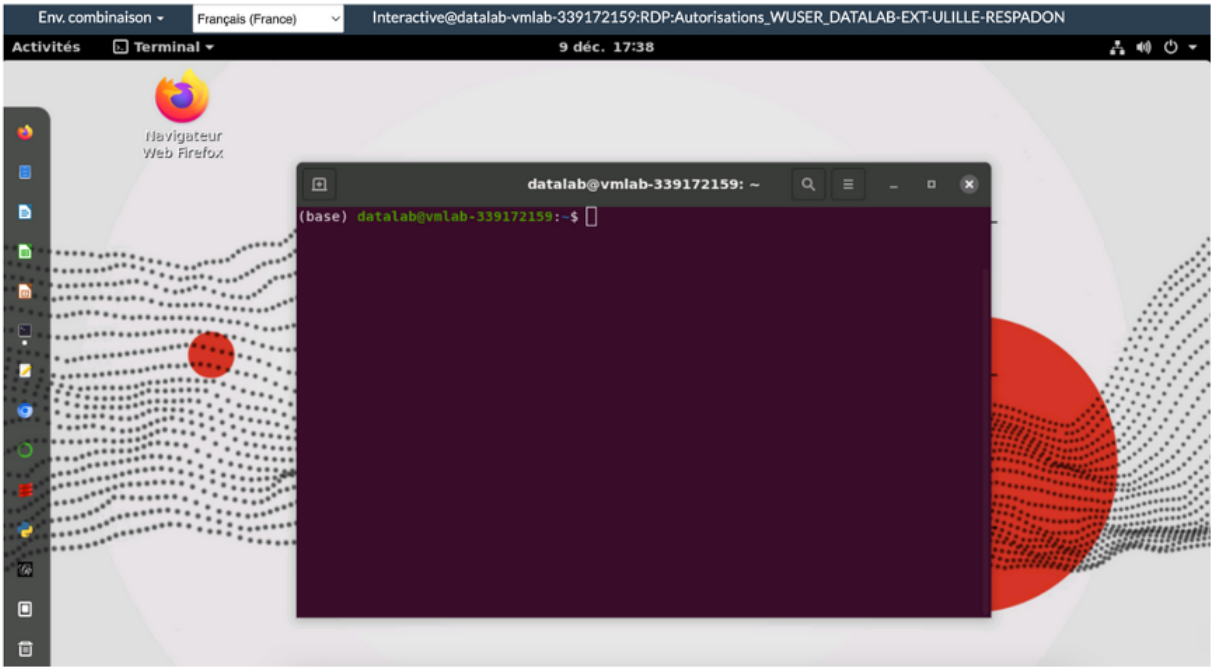


Schéma fonctionnel simplifié de l'accès distant à la capsule d'exploration des archives élections 2002 avec la solution WALLIX

La solution WALLIX a été choisie pour accéder aux applications et aux données relatives aux élections 2002, la collection retenue par les partenaires du projet ResPaDon pour expérimenter de nouvelles techniques et de nouveaux outils d'exploration et fouille de textes et de données (TDM). Cette solution repose sur l'utilisation d'une application sur un téléphone portable qui permet de générer un jeton à usage unique à chaque connexion au service web. Contrairement à la première solution, elle ne nécessite pas l'installation d'une extension de navigateur.

Du point de vue de l'Université de Lille, la mise en place de cette solution a impliqué l'enregistrement préalable de deux référents du SCD (assimilés à des chercheurs Datalab) dans l'annuaire de la BnF et l'installation et le paramétrage de l'application inWebo Authenticator sur deux téléphones portables.
A chaque fois qu'un utilisateur veut accéder aux applications et aux données relatives aux élections 2002, un référent du SCD doit au préalable s'identifier. L'identification identifiant/mot de passe est couplée avec la saisie d'un jeton à usage unique (OTP) généré par l'application inWebo Authenticator sur un téléphone portable. Le référent accède ensuite à un "portail de ressources" sur lequel il sélectionne un poste virtuel puis saisit d'autres identifiants (login/mot de passe) qui lui permettent d'ouvrir une session. L'utilisateur dispose d'un poste de travail complet embarquant, sur un disque d'un volume de 1To, un système Linux Ubuntu et un ensemble de logiciels (navigateur, logiciels bureautiques, logiciels de programmation et d'analyse de données) qui lui permettent de travailler sur et avec les applications et les données relatives aux élections 2002. En conformité avec la politique de sécurisation des accès aux données relevant du dépôt légal du web et soumises au droit d'auteur, ce poste de travail est totalement déconnecté du web vivant. Pour des raisons techniques, il est actuellement impossible pour l'utilisateur de récupérer des références, des extraits de texte ou le résultat d'un travail de recherche.



Poste de travail contenant les applications et les données relatives aux élections 2002

Du point de vue de la BnF, il a été décidé d'utiliser une infrastructure actuellement utilisée par les administrateurs du DSI lorsqu'ils sont en télétravail. Cette infrastructure est complètement différente et séparée de l'infrastructure sur laquelle repose la solution inWebo.
L'accès distant sécurisé pour des utilisateurs non-agents de la BnF avait été étudié lors de la mise en place de l'infrastructure technique du BnF DataLab. Ces travaux ont été repris pour permettre de sécuriser et de cloisonner les accès externes des référents du SCD au seul poste de travail contenant les applications et données relatives aux élections 2002. Pour garantir la sécurité des systèmes, des applications et des collections, ainsi que la stabilité et la traçabilité des usages, cette infrastructure n'a pas vocation à être utilisée au-delà de l'expérimentation menée dans le cadre de ResPaDon.

Recommandations

<u>Recommandation n°5</u> : Mettre en place à la BnF une infrastructure informatique permettant le passage à l'échelle du dispositif expérimental

<u>Recommandation n°6</u> : Améliorer la solution d'accès sécurisée aux collections de dépôt légal du web pour la rendre plus robuste et conforme à la politique de sécurité des établissements de l'ESR :

- **Conformité à la politique de sécurité** : la solution inWebo fonctionne uniquement avec une session utilisateur générique, ce qui est contraire à la politique de sécurité des postes informatiques de l'Université de Lille qui impose que les utilisateurs soient personnellement identifiés. Le support informatique de l'Université de Lille a exceptionnellement accepté que cette identification des usagers soit faite par le biais de la connexion au réseau wifi pendant l'expérimentation. La solution inWebo ne peut pas être exploitée telle quelle dans le cadre d'un service régulier. La solution cible doit prendre en compte les politiques de sécurité des systèmes d'information des Universités. Une formalisation de ces exigences validée par des instances telle que la conférence des DSI permettrait de concevoir d'emblée un dispositif intégrant les

exigences des deux parties en matière de sécurité et d'éviter les adaptations au cas par cas coûteuses en temps et en ressources.

- **Conformité aux procédures de maintenance et de gestion des postes** : la solution inWebo repose sur une extension qui doit être réinstallée lors de chaque mise à jour du navigateur ou processus de nettoyage/réinitialisation du poste informatique. La solution cible doit être compatible avec les procédures de mise à jour et gestion courante des postes informatiques.

- **Simplification de l'authentification** : les deux solutions techniques reposent sur deux systèmes d'inscriptions et d'authentification des usagers externes différents : inWebo repose sur l'enregistrement du poste et de la session informatiques, WALLIX sur la création d'un compte personnel dans l'annuaire des chercheurs du DataLab. La connexion au SI de la BnF via inWebo est transparente pour l'usager, celle via WALLIX implique la saisie d'un identifiant, d'un jeton à usage unique généré via un smartphone, d'un mot passe puis de nouveaux identifiant/mot de passe. Deux URL opaques et distinctes servent de point d'entrée aux deux dispositifs. La solution cible doit permettre une fluidification et une simplification de l'authentification et une meilleure lisibilité des points d'entrée.

- **Homogénéisation** : avoir un dispositif composé de deux solutions techniques différentes (inWebo et WALLIX), l'une donnant accès à l'application Archives de l'internet, l'autre embarquant de nouveaux outils d'exploration et fouille de textes et de données (TDM) sur une collection particulière est complexe à installer, à appréhender et à utiliser. Les deux environnements sont hermétiques, sans possibilité de passer de l'un à l'autre et ne permettent pas une utilisation en complémentarité (par exemple : identifier un article sur les élections 2002 et retrouver les sites en lien avec cet article dans les Archives de l'internet). Il est indispensable de proposer un seul environnement proposant l'ensemble des applications et des données utiles aux projets de recherche. Cela doit permettre de simplifier les modalités de connexion et l'utilisation complémentaire des applications et collections mises à disposition.

- **Robustesse de la solution** : la solution inWebo est peu robuste car dépendante d'une extension du navigateur Chrome et reposant sur une infrastructure qui doit être renouvelée. La solution WALLIX a été conçue pour un usage interne aux administrateurs du DSI et non pour un usage public, elle n'est donc pas supervisée comme les autres services proposés aux usagers de la BnF. La BnF doit concevoir un nouveau schéma d'architecture et installer une nouvelle infrastructure dans la perspective d'ouverture d'un service de capsule.

- **Support technique** : dans le cadre de l'expérimentation, plusieurs incidents techniques (problèmes réseau, problèmes de licence, problèmes de disparition du jeton d'enrôlement) ont été remontés par les médiateurs : ces incidents doivent pouvoir être signalés à un service support chargé du suivi de leur résolution. Les procédures d'exploitation doivent être renforcées et partagées entre plusieurs administrateurs sur le modèle des procédures existantes.

<u>Recommandation n°7</u> : Améliorer l'ergonomie de la solution d'accès distant sécurisé pour la rendre conforme aux usages de recherche, notamment en facilitant l'export et le copié collé des données.
Pour permettre des usages recherche intensifs, il est nécessaire d'améliorer l'ergonomie de cet environnement de travail distant et d'introduire :

- la possibilité d'exporter ponctuellement des données à des fins d'analyse ou de travail : facilitation du copier-coller d'extraits de texte, et de permaliens, utilitaire de capture d'écran ;
- une semi-automatisation de la procédure d'export des données issues de la recherche, c'est-à-dire des données "produites avec les outils d'analyse, de requêtage, de programmation ou de visualisation livrés dans l'environnement sécurisé, données qui ainsi enrichies et transformées constituent le résultat original de la recherche" : il serait souhaitable de permettre au fil de l'eau cet export et non seulement par une demande par mail à l'issue du travail de recherche.

III) Outils pour la découverte, l'exploration et la fouille des collections

La capsule ResPaDon propose un ensemble de données et d'applications permettant l'exploration, le traitement, l'analyse et la fouille des données collectées au titre du dépôt légal du web.

Un premier ensemble d'applications étaient déjà proposées en bibliothèque de dépôt légal imprimeur (BDLI) ainsi que sur les postes de lecture des espaces recherche de la BnF. Déployé à l'Université de Lille lors d'une première phase de tests sous le nom de "capsule découverte", à compter de début octobre 2021, ce premier ensemble comprenait :

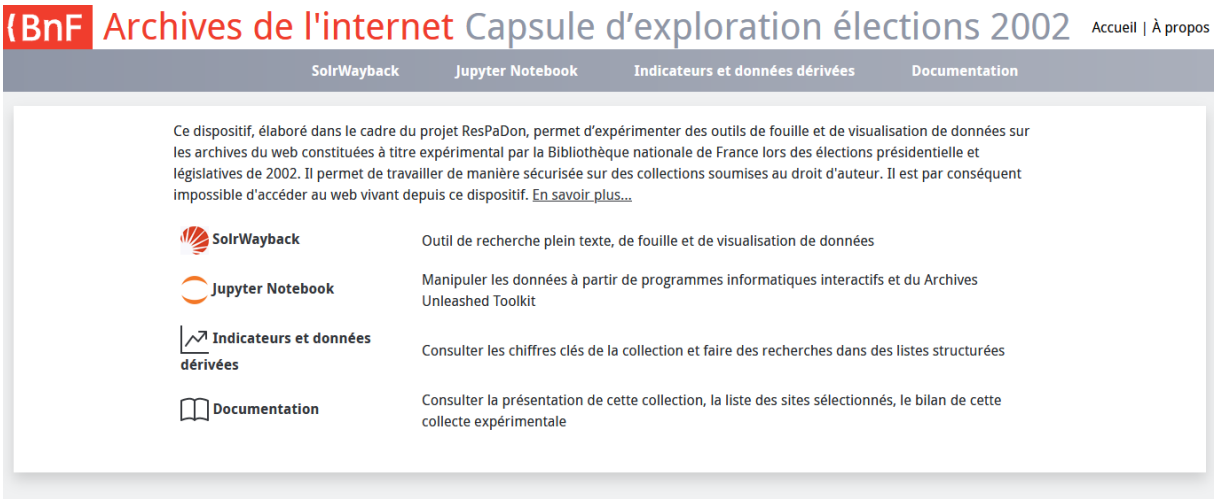
- l'application Archives de l'internet permettant de consulter l'ensemble des collections de dépôt légal du web (1996 à aujourd'hui, représentant 1,8 Po de données) à partir d'une recherche par URL, ou au travers de parcours guidés, tels que « Cliquer, voter : l'internet électoral » portant sur les élections 2002 à 2007 et « Le web électoral de 2010 à 2015 »,
- l'application Archives de l'internet Labs proposant une recherche plein texte sur quatre sous-ensembles documentaires[1] : la collection "Presse et actualité" constituée par la collecte quotidienne d'une centaine de titres de presse et de sites d'actualité depuis 2010, la collection relative aux attentats de Paris de 2015, la collection relative à la première vague de l'épidémie de Covid-19, et la collection "Incunables du Web" constituée par l'acquisition rétrospective auprès d'Internet Archive des premiers sites du domaine français collectés entre 1996 et 2000. Cette application embarque également une recherche ngram permettant de visualiser l'évolution de la fréquence d'un ou plusieurs termes ou expressions dans les contenus collectés au fil du temps.

Ces deux applications embarquent une aide en ligne et sont des outils éprouvés, développés et maintenus par la BnF depuis de longues années. Dans le cadre de la capsule, un accompagnement personnalisé à la prise en main et découverte de ces collections et une documentation a été conçue, décrit plus bas.

Un second ensemble d'applications et de données baptisé "capsule élections 2002" est venu compléter ce premier ensemble à compter de mai 2022. Il était destiné à encourager la fouille de texte et de données sur un corpus de petite taille, isolable du reste des collections de dépôt légal du web, et présentant un réel intérêt scientifique, la collection élections 2002, constituée par la collecte du web relatif aux élections présidentielle et législatives françaises de 2002. Les fichiers composant cette collection (26 M de fichiers web, convertis et conservés dans 1654 fichiers au format ARC compressés, soit au total 167 Go de données), et des applications permettant de l'explorer et de les fouiller ont été déployées à titre expérimental pour les besoins du projet et installés sur une machine virtuelle Linux conçue comme un environnement de travail à part entière coupé du web vivant. Conçue pour permettre l'exploration et la fouille, cette capsule avait également pour objectif de donner un aperçu aussi large que possible des outils, données et méthodes mobilisables pour travailler sur les archives du web, notamment pour croiser lecture distante et lecture rapprochée, analyses qualitatives et méthodes quantitatives. C'est dans cet esprit qu'ont été proposés dans cette

[1] A l'issue de l'expérimentation capsule en septembre 2023. De nouveaux corpus sont régulièrement indexés en plein texte.

“capsule Elections 2002” des applications d’aide à la fouille de texte et de données, différents types de métadonnées et de données dérivées, des indicateurs statistiques, ainsi que des logiciels permettant de conduire des analyses et des traitements sur les données décrits ci-dessous.



Page d’accueil du portail de la capsule élections 2002

Les applications d’aide à la fouille de texte et de données

Trois applications ont été installées et adaptées pour permettre d’explorer et de fouiller la collection élections 2002 :

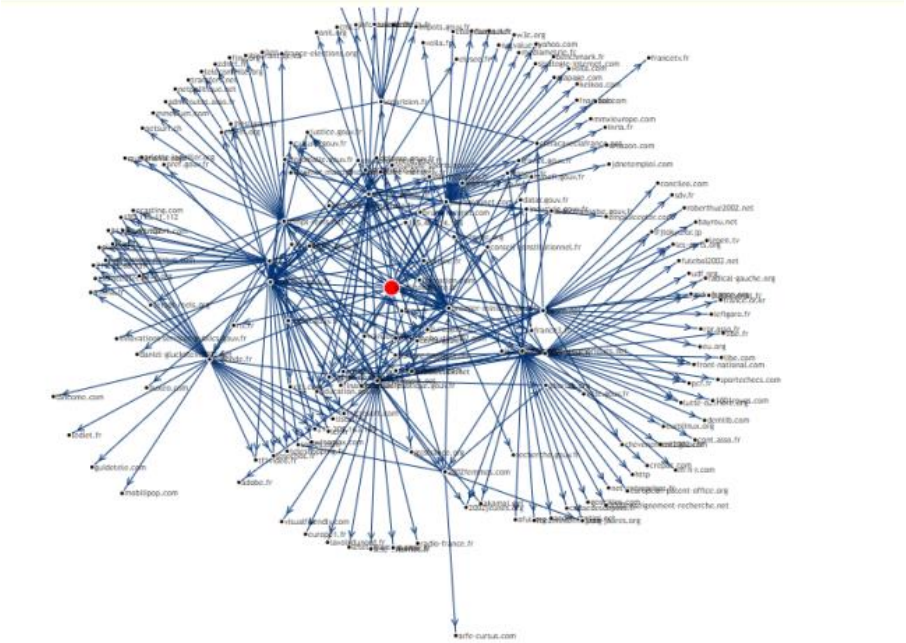
- SolrWayback (https://github.com/netarchivesuite/solrwayback) est un outil open source développé par la Bibliothèque Royale du Danemark et au développement duquel la BnF contribue depuis 2022 qui propose des fonctionnalités d’exploration : recherche par adresse URL, recherche plein texte, navigation et visualisation. Elle propose également des fonctionnalités de fouille permettant d'analyser les contenus à l’aide d’outils de data visualisation : nuage de mots, graphe de liens entrants et sortants interactif, export Gephi, statistiques par domaine, recherche n-gram. SolrWayback permet également de faire des recherches sur des images par mot clé ou par similitude en chargeant une image et des recherches sur les coordonnées GPS qui se trouvent dans les images.

La fonction "recherche image" de SolrWayback, ici avec le mot "nucléaire", permet de faire ressortir les images présentes dans la collection élections 2002 qui soit ont des métadonnées contenant le mot nucléaire, soit se trouvent dans une pages à proximité d'un texte contenant ce mot

La "boîte à outils" de SolrWayback permet de générer différentes formes de visualisations de données : ici un nuage de mots clés généré à partir des pages du site "front-national.com" qui montre les mots ayant le plus grand nombre d'occurrences

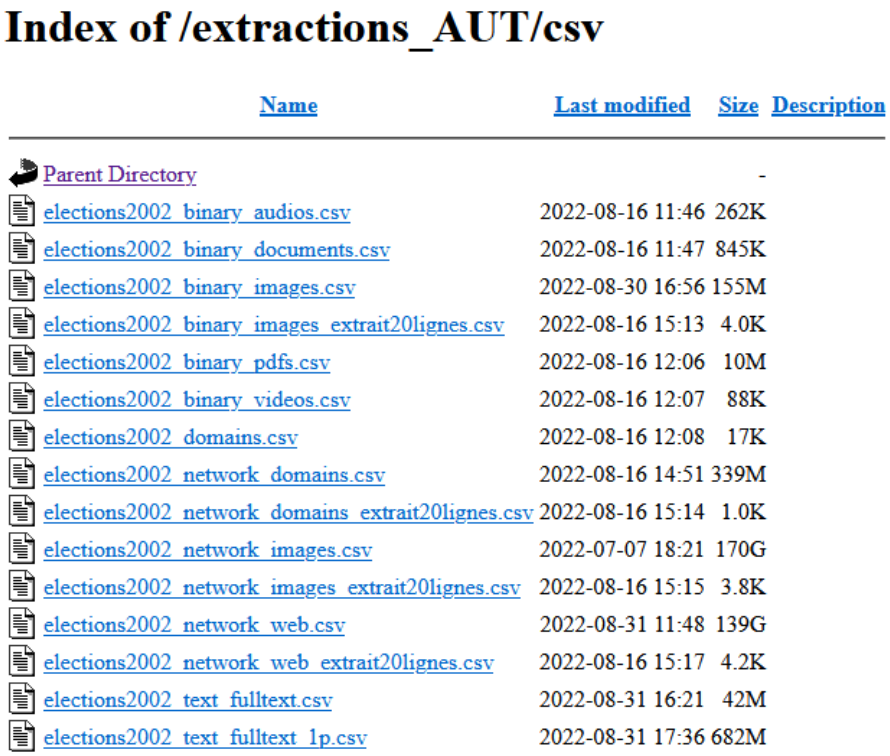
	2002-01	2002-02	2002-03	2002-04	2002-05	2002-06	2002-07	2002-08
Size in kilobytes	0	53,172	2,136,652	7,768,931	8,736,618	6,426,155	0	0
Pages	0	2,692	157,467	579,287	477,825	327,113	0	0
Ingoing links	1	6	43	60	58	72	7	0

La fonction "statistiques par domaine" permet de visualiser, pour un site web donné, ici "parti-socialiste.fr" l'évolution du nombre de pages collectées, du poids des pages collectées, du nombre de liens hypertextes pointant vers les pages. Les évolutions sont représentées par mois sur une période de 6 mois, une fonctionnalité développée par la BnF sur cet outil à l'occasion du projet



La "boîte à outils" de SolrWayback permet de produire des cartographies de sites web qui pointent vers un site particulier, ou, à l'inverse, les sites vers lesquels ce site pointe, une fonctionnalité qui peut aider à l'analyse des réseaux et communautés d'acteurs. Les listes de liens peuvent aussi être extraites afin d'utiliser le logiciel Gephi (installé dans l'environnement de travail) pour produire des visualisations plus complexes

- La boîte à outils Archives Unleashed (AUT, https://archivesunleashed.org/aut/) initiée par l'Université de Toronto est composée d'un ensemble de librairies conçu pour faciliter l'analyse des fichiers W/ARC composant une collection d'archives web et le travail sur les résultats. Elle embarque des fonctions qui peuvent être utilisées à travers des scripts en langage Python ou Scala pour faire des analyses au niveau de la collection, sur le texte des pages, sur les liens entre les pages, sur les images et les autres contenus binaires.



Index of /extractions_AUT/csv

Name	Last modified	Size	Description
Parent Directory		-	
elections2002_binary_audios.csv	2022-08-16 11:46	262K	
elections2002_binary_documents.csv	2022-08-16 11:47	845K	
elections2002_binary_images.csv	2022-08-30 16:56	155M	
elections2002_binary_images_extrait20lignes.csv	2022-08-16 15:13	4.0K	
elections2002_binary_pdfs.csv	2022-08-16 12:06	10M	
elections2002_binary_videos.csv	2022-08-16 12:07	88K	
elections2002_domains.csv	2022-08-16 12:08	17K	
elections2002_network_domains.csv	2022-08-16 14:51	339M	
elections2002_network_domains_extrait20lignes.csv	2022-08-16 15:14	1.0K	
elections2002_network_images.csv	2022-07-07 18:21	170G	
elections2002_network_images_extrait20lignes.csv	2022-08-16 15:15	3.8K	
elections2002_network_web.csv	2022-08-31 11:48	139G	
elections2002_network_web_extrait20lignes.csv	2022-08-16 15:17	4.2K	
elections2002_text_fulltext.csv	2022-08-31 16:21	42M	
elections2002_text_fulltext_1p.csv	2022-08-31 17:36	682M	

Extractions pré-générées avec la boîte à outils Archives Unleashed Toolkit pour permettre l'identification de certains types de données (par exemple les enregistrements audios ou vidéos) et l'exécution de scripts (par exemple pour analyser le texte des pages d'un site web particulier)

- Des carnets Jupyter ou *Jupyter Notebooks* (https://github.com/archivesunleashed/notebooks/ et https://glam-workbench.net/web-archives/) permettent de faciliter l'utilisation de la boîte à outils Archive Unleashed Toolkit (AUT). Ils associent du code, des graphiques, des visualisations et du texte dans des carnets interactifs qui s'exécutent directement dans un navigateur web. Le développement et l'utilisation de carnets Jupyter est en plein essor dans le domaine de l'enseignement supérieur et de la recherche. Ils sont conçus pour des utilisateurs qui ne maîtrisent pas le code informatique et fournissent des exemples de requêtes qu'il est possible d'adapter pour obtenir des résultats ciblés sur d'autres recherches.Trois des

carnets conçus par Archives Unleashed ont été adaptés pour permettre d'explorer la collection élections 2002 et ses données dérivées.

Entrée [59]:
```
import tldextract

domain_frequency["tld"] = domain_frequency.apply(
    lambda row: tldextract.extract(row.apply(str).domaine).suffix, axis=1
)
domain_frequency
```

Out[59]:

	domaine	count	tld
0	parti-socialiste.fr	4935553	fr
1	lioneljospin.net	3231432	net
2	ladepeche.com	2008761	com
3	pcf.fr	1928150	fr
4	lapolitique.com	1492945	com
...	...	...	...
860	patrick-bonnet-prg.com	2	com
861	lepen.net	2	net
862	verts-76-10.org	1	org
863	google.com	1	com
864	google.org	1	org

865 rows × 3 columns

Utilisation d'un carnet Jupyter pour produire la liste et un graphe des noms de domaines les plus représentés dans la collection élections 2002

Ensuite, nous allons lancer le processeur de langage naturel de SpaCy, et ensuite afficher la sortie NER qui identifie les organisations, les personnes...

Entrée [30]:
```
ner = nlp(page)
displacy.render(ner, style="ent", jupyter=True)
```

DU 12 AVRIL 1999 A L'ESPACE MONCASSIN Philippe Cosnay PER : Bien. Si vous voulez bien rejoindre vos places. Nous allons commencer la séance. Au nom de la section Javel-Grenelle LOC , je vous remercie tous d'être présents, d'être venus si nombreux dans le 15ème arrondissement. Je te remercie bien sûr, François PER , et Pervenche LOC qui va nous rejoindre, respectivement tête de liste et numéro deux de la liste que nous allons défendre tout au long des deux mois, quasiment jour pour jour maintenant, qui nous reste dans cette campagne (...) des élections européennes. Je salue l'arrivée de Jean-Marie Le Guen PER , Premier Secrétaire fédéral PER de Paris LOC ; je salue Michel Charzat PER , sénateur-maire du 20ème arrondissement ; j' MISC aperçois certains candidats dans la salle : Jean Malot PER , d'autres candidats vont nous rejoindre au cours de la soirée. Je vois dans votre présence un appui à notre démarche, et je vous en remercie chaleureusement. Je tiens à remercier tout particulièrement nos camarades européens qui ont bien voulu participer à cette grande rencontre socialiste européenne : Françoise Auquier PER , candidate aux élections régionales en Belgique LOC , qui auront lieu également le 13 juin ; Alexis Rowell PER , qui représente le Parti Travailliste ORG à Paris LOC ; ainsi bien sûr que les sections membres de notre réseau, le REseau MISC de Sections Socialistes Européennes,

Exemple d'analyse produite sur un échantillon de texte avec un carnet Jupyter utilisant la bibliothèque Python Spacy spécialisée dans le traitement automatique des langues. Le carnet utilise la reconnaissance d'entités nommées, ou NER, qui permet d'identifier des "entités" du texte, à savoir des noms de personnes, de lieux ou d'organisations

Les métadonnées, indicateurs et données dérivées

Cette capsule embarque également un ensemble d'indicateurs, de données dérivées et de métadonnées sur la collection élections 2002. Les indicateurs correspondent à des indicateurs généraux (nombre d'URL, nombre et répartition des domaines, nombre et répartition des types MIME…) qui permettent de quantifier et qualifier la collection, et d'en faire une première lecture. Les données dérivées ont été créées avec la boîte à outils Archives

Unleashed et s'appuient sur les modèles définis par les équipes d'AUT et d'Internet Archive dans le cadre du service ARCH (Archive Research Compute Hub) : données dérivées sur la collection complète, sur les liens, le texte et tous les objets binaires composant la collection. Les métadonnées correspondent aux informations documentaires issues de la sélection pour les campagnes électorales et publiées sur http://api.bnf.fr.

INDICATEURS

Les fichiers ci-dessous regroupent un ensemble de chiffres clés permettant d'appréhender le contenu de la collection élections 2002. Ils sont également accessibles via un Terminal dans le répertoire ~/Documents/espace-BnF/data/elections2002/indicateurs.

Fichier	Poids	Description	
elections2002_indicateurs_generaux.txt	357 o	Indicateurs de niveau collection	Télécharger
elections2002_repartition_domaine.txt	17 Ko	Répartition des URL collectées par nom de domaine	Télécharger
elections2002_repartition_tld.txt	228 o	Répartition des URL collectées par TLD (extension : .com, .fr, etc.)	Télécharger
elections2002_repartition_typemime.txt	725 o	Répartition des URL collectées par type MIME (information sur les formats représentés sur internet)	Télécharger

DONNÉES DÉRIVÉES

Les données dérivées représentent un type de traitement ou d'agrégation réalisées à partir des fichiers ARC composant la collection élections 2002. Elles ont été générées à l'aide du Archives Unleashed Toolkit. Elles portent sur le niveau collection, le niveau réseau de liens, le contenu textuel des pages ou les fichiers binaires. Ces fichiers sont également accessibles via un Terminal dans le répertoire ~/Documents/espace-BnF/data/elections2002/donnees-derivees. Les URL présentes dans ces fichiers peuvent être consultées via SolrWayback en cochant "URL Search".

Fichier	Poids	Description	
elections2002.cdx	5,2 Go	Index CDX de l'ensemble des URL collectées	Disponible uniquement via le Terminal
elections2002_binary_audios.csv	262 Ko	Liste des fichiers audios (1 371 fichiers)	Télécharger
elections2002_binary_documents.csv	845 Ko	Liste des fichiers de type documents (4 261 fichiers)	Télécharger
elections2002_binary_images.csv	155 Mo	Liste et emplacement des images (785 450 fichiers)	Télécharger
elections2002_binary_images_extrait20lignes.csv	4 Ko	Échantillon de 20 images	Télécharger
elections2002_binary_pdfs.csv	10 Mo	Liste des fichiers PDF (51 359 fichiers)	Télécharger
elections2002_binary_videos.csv	88 Ko	Liste des fichiers de type vidéos (433 fichiers)	Télécharger

Les indicateurs et données dérivées générés à partir des logs des outils de collecte ou via la boîte à outils Archives Unleashed permettent d'analyser la constitution de la collection dans sa globalité et d'en faciliter la lecture distante. Les listes de fichiers bureautiques ou de fichiers audios, listes de liens, sont un exemple d'aide à la fouille des données qui la composent. Ces données peuvent ensuite être visualisées, requêtées et exploitées via les applications et logiciels disponibles dans la capsule

Toutes les applications (SolrWayback, boîte à outils Archies Unleashed, carnets Jupyter) et les données (indicateurs, données dérivées, métadonnées) mises à disposition dans la capsule peuvent être utilisées de manière autonome ou complémentaire et constituent autant de points de départ potentiels pour étudier la collection. Ainsi, il est possible d'étudier dans les données dérivées la liste des fichiers images présents dans la collection et de visualiser, via la recherche par URL disponible au sein de SolrWayback, les images en question et les sites qui les contiennent ; ou encore de se servir de la liste des URL de départ pour repérer les contenus humoristiques présents au sein de la collection et effectuer des requêtes plus ciblées, ou encore de faire une recherche plein texte, d'isoler via les facettes un domaine intéressant et de générer un graphe de lien à l'aide de la boîte à outils Archives Unleashed.

La documentation

La capsule élections 2002 embarque un ensemble de documents qui accompagnent la prise en main des différentes applications et de la collection. Elle contient :

- des pas à pas et exemples de recherche sur les différentes applications,
- la liste des sites sélectionnés et transmis au robot de collecte. Cette liste contient également des métadonnées descriptives conformément à la typologie documentaire adoptée : "Fréquence", "Profondeur", "Typologie", "Type d'élections", "Parti", "Candidat", "Autres mots clés", "Historique des URL".
- un bilan statistique rédigé en janvier 2003 qui fournit des indicateurs quantitatifs sur les données effectivement collectées par catégories de sites et par candidat.
- le bilan qui documente le processus expérimental de création de la collection élections 2002 dans ses aspects technique, organisationnel et documentaire. Il présente le modèle de production adopté, et explicite les logiques documentaires qui ont présidé à la sélection des contenus à archiver. Ce bilan fournit également de nombreux éléments d'analyse concernant les traits saillants de la campagne.
- un glossaire des termes techniques spécifiques aux archives du web et aux outils proposés dans la capsule (format ARC, index CDX, profondeur de collecte, etc.).

Recherche d'images par mots clés (en cochant "Images")

- tract
- le pen
- nucléaire

Recherche par adresse URL (en cochant "URL Search")

- http://www.conseil-constitutionnel.fr/dossier/presidentielles/2002/documents/liste/liste.htm
- http://www.gauchestory.com
- http://www.gauchestory.com/jeu.swf
- http://www.bayrou.net/forum/index.html
- http://www.front-national.com/discours/2002/21-04-2002.htm

Visualisation des pages archivées

La date de capture, le calendrier de capture, ainsi que des informations sur la collecte de chaque page sont disponibles en cliquant sur le bouton "Toolbar" situé en haut à gauche de chaque page consultée.

Dans la boîte à outil (Toolbox)

Wordcloud

- chiracaveclafrance.net
- olivierbesancenot.org
- christineboutin2002.com

Link graph

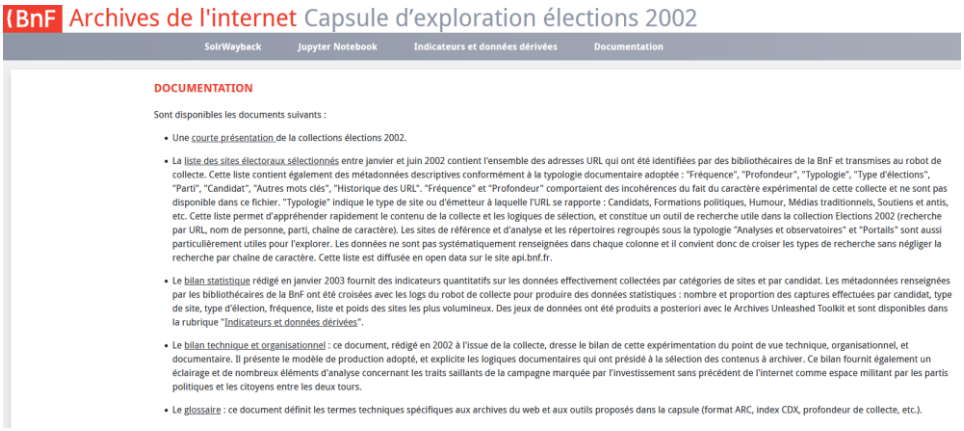
Max. node degree correspond au nombre maximum de liens entrants (ingoing) ou sortants (outgoing) d'un noeud

- pfc.fr (Max. node degree à 10, Ingoing, affiche les 10 sites qui comportent le plus de liens hypertextes pointant vers pcf.fr)
- front-national.com (Max. node degree à 10, Outgoing)
- lcr-rouge.org (Max. node degree à 10, Outgoing)

Link graph Gephi export

- crawl_date:[2002-04-20T00:00:00Z TO 2002-05-06T23:59:59Z] pour avoir une vue sur l'ensemble des sites collectés entre les deux scrutins

Exemples pour l'utilisation de l'application SorlWayback



Liste de la documentation accessible à l'intérieur de la capsule

Les logiciels complémentaires de traitement des données

En complément des applications et des données, la capsule embarque un ensemble de logiciels communément utilisés par les chercheurs en sciences sociales et dans le champ des humanités numériques :

- un ensemble de plugins, notamment Flash car la collection élections 2002 contient de nombreuses animations,
- IRaMuTeQ : un logiciel open source d'analyse de données textuelles ou de statistique textuelle,
- Anaconda : un système open source pour faciliter l'utilisation de langages de programmation Python et R appliqués au développement d'applications dédiées à la science des données et à l'apprentissage automatique,
- Gephi : un logiciel libre d'analyse et de visualisation de réseaux.

La collection élections 2002

La collecte du web électoral réalisée à l'occasion des élections présidentielle et législatives de 2002 est l'une des toutes premières collectes automatisées réalisées par la BnF. Il s'agissait d'une expérimentation destinée à préciser les contours techniques et documentaires du dépôt légal du web français tels qu'ils seront entérinés par la loi relative au droit d'auteur et aux droits voisins dans la société de l'information (DADVSI) de 2006.

La collecte a duré 19 semaines et a été construite autour des deux tours de l'élection présidentielle et des élections législatives. Le temps d'un événement, elle a permis d'expérimenter une infrastructure de collecte, un modèle de prospection et de sélection de contenus web, une typologie documentaire.

Volumétrie Au total, 1906 sites web, parties de sites web ou newsletter ont été sélectionnés par un groupe de bibliothécaires de la BnF et collectés par robot à intervalles réguliers entre janvier et juin 2002. La collection représente 26 M de fichiers web, convertis et conservés dans 1654 fichiers au format ARC, soit au total 167 Go de données.

Typologie documentaire Les sites sélectionnés se répartissent entre les catégories documentaires suivantes, déclinées suivant le type d'émetteur :

Candidats (sites de chaque candidat) : 483 sites Formations politiques (partis, syndicats, mouvements, personnalités) : 550 sites Médias traditionnels (presse, radio, télé) : 242 sites Soutiens et antis (soutiens aux candidats, anti-candidats ou contestataires, autres) : 157 sites Analyses et observatoires (sondages, marketing politique, communication, aspects juridiques) : 46 sites Enseignement et recherche (centres de recherche, enseignements supérieurs, écoles, étudiants) : 6 sites Humour (contenus satiriques, humoristiques, ludiques) : 71 sites Net-politique (forums, chats, débats... en ligne) : 12 sites Officiels et institutionnels / gouvernementaux, locaux : 17 sites Portails spécialisés présidentielles, répertoires, annuaires, autres : 26 sites Webzine (Cyber-Médias / webzine, pages spéciales) : 50 sites Divers (autres sites inclassables dans les rubriques précédentes, e-commerce) : 246 sites La liste complète des sites sélectionnés est disponible dans la capsule au format CSV (cf. Documentation).

Les sites ont été capturés à diverses fréquences et périodicités :

- plusieurs fois par semaine, pour les sites constituant le « noyau » de la collecte : sites de candidats, de partis et formations politiques,

- une fois par semaine, périodicité proposée par défaut pour les autres adresses, et la plus fréquemment utilisée pour des sites de médias,

- trois fois suivant les temps forts du scrutin : avant, entre les deux tours, après le deuxième tour, - une seule fois, notamment pour certains articles de presse isolés. Ces périodicités et fréquences de collecte ont toutefois évolué au cours des 19 semaines qu'a duré la collecte, tant pour rendre compte des changements survenus entre les deux tours que du fait du caractère expérimental des outils et de l'organisation.

La collecte visait à constituer un échantillon représentatif du web électoral et à refléter la diversité des utilisations de l'internet dans la campagne. Les sites des candidats et partis en campagne constituent ainsi le cœur de la collection, et permettent notamment de mesurer l'inégal investissement du web comme outil de propagande ou de mobilisation par les grandes familles politiques. Les sélections suivent également la présence sur le web de différents types d'acteurs de la société civile, associations, syndicats, communauté académique et de rendre compte des différents registres d'expression et d'action mobilisés pendant la campagne. La multiplication des contenus parodiques et humoristiques est ainsi apparue comme un trait saillant du web électoral de 2002. Une attention particulière a été portée à documenter l'émergence d'une « Net-politique » combinant outils d'analyse traditionnels et instruments de mesure spécifiques au web, ou encore à la constitution, par les principaux fournisseurs d'accès à Internet de l'époque, de portails web et répertoires dédiés aux élections. Ces contenus et sources d'analyse à chaud par les acteurs de l'époque ont également permis de nourrir les stratégies de prospection documentaire, au même titre que les « webrings » ou listes de sites amis présents sur les blogs.

Le 21 avril 2002 et la capture sur le vif des réactions numériques à un événement

Le séisme qu'a constitué la qualification de Jean-Marie Le Pen au second tour de l'élection présidentielle s'est traduit par un investissement sans précédent du web comme outil de

mobilisation politique, et par l'émergence de nouvelles formes d'appropriation du web comme espace de débat et d'expression citoyenne. La capture régulière des sites associatifs et de forums, de sites pétitionnaires ou encore la collecte des réactions de la presse étrangère aux résultats de l'élection sont autant de matériaux pour analyser ces réactions dans la temporalité de l'événement, ainsi que ses limites, dans la mesure où les observateurs et sélectionneurs ont d'emblée noté que les prises de positions de certains acteurs dans l'arène politique se répercutaient inégalement sur leurs sites web. Enfin la collecte de contenus web en lien avec les élections législatives a permis d'esquisser de nouvelles méthodes de prospection, sur une base géographique cette fois. La collection d'archives web élections 2002 constitue ainsi un matériau unique en son genre, qui a d'ores et déjà donné lieu à des travaux de recherche très riches sur le rôle de l'internet dans la communication politique, son inégal investissement par les formations politiques, le renouvellement du militantisme avec l'arrivée d'internet, le marketing politique. Les pistes d'exploration scientifique, 20 ans après l'événement, restent nombreuses, et les outils d'exploration et de fouille déployés dans cette capsule à titre expérimental entendent accompagner susciter de nouveaux usages. Cette collecte inaugure une série de 20 collectes électorales qui se sont succédé sur le même modèle entre 2002 et 2022 au rythme de l'agenda électoral national. La collection élections 2002 est à ce titre un échantillon emblématique d'un ensemble documentaire plus vaste, les collectes du web électoral, qui sont consultables via une recherche par URL dans la capsule ResPaDon.

Recommandations

Recommandation n°8 : Consolider et enrichir la palette d'outils d'exploration et d'aide à la fouille de texte et de données proposées dans la capsule, afin d'améliorer la découvrabilité des collections.

Recommandation n°9 : Travailler sur le design et l'ergonomie de l'interface usager de la capsule en s'appuyant sur les retours de tests pour concevoir un parcours unifié.

Les retours de tests fournissent un éclairage intéressant sur les améliorations à apporter aux outils proposés :

- **Articulation entre les différents outils** : pour faciliter la mise en œuvre de la capsule, les outils d'exploration enrichie étaient proposés dans un dispositif technique autonome et distinct du dispositif BDLI existant. La construction d'un seul et unique environnement permettrait d'offrir une facilité et une fluidité dans l'utilisation de la capsule ainsi qu'une meilleure articulation et le renforcement des différentes fonctionnalités.
- **Expérience usager** : la consolidation et l'amélioration du parcours usager et du design applicatif apparaissent comme un enjeu crucial pour faciliter l'appropriation des archives du web.
- **Outils de capture d'écran et copier-coller** : les fonctions étaient impossibles dans le dispositif WALLIX et jugées trop fastidieuses ou insuffisantes dans le dispositif BDLI. Elles doivent être ouvertes et plus intuitives.
- Améliorations à apporter aux **outils ou fonctionnalités complémentaires** :
    - Application "Archives de l'internet" : l'ergonomie de l'interface et la qualité de la navigation dans les archives ont été appréciées. L'amélioration de la découvrabilité ou recherchabilité du contenu font partie des demandes d'amélioration les plus fréquemment exprimées, par l'intermédiaire d'une

recherche plein texte sur l'ensemble des collections. Moins coûteuse, la mise en place d'outils d'aide à la recherche sur les adresses URL, notamment celles des sites sélectionnées dans le cadre des collectes ciblées, serait très utile aux usages pédagogiques et de recherche. Les nombreuses métadonnées existantes et dont une partie est déjà disponible sur api.bnf.fr et data.gouv.fr pourraient être davantage valorisées à cette fin.

- SolrWayback : la recherche par image et les fonctionnalités de visualisation ont été appréciées des testeurs.
  - De très nombreuses questions ont concerné le fonctionnement de l'algorithme de recherche et le classement des résultats, pour la recherche image ou la recherche par mot. Une aide contextuelle explicitant ce fonctionnement, ainsi qu'une interface en français répondraient sans doute à ce besoin. L'internationalisation de l'interface fait partie de la feuille de route de développement de SolrWayback et la BnF appuie et contribue à cette évolution.
  - L'autre demande d'amélioration exprimée lors des tests était de pouvoir paramétrer de manière plus fine les visualisations, par exemple en représentant des évolutions par semaine ou mois et pas seulement par année. De nombreux outils de la "boîte à outils" ne permettaient de visualiser les évolutions qu'à l'échelle d'une année, ce qui n'était pas pertinent pour la collection élections 2002. Cette fonctionnalité a pu être implémentée par la BnF au cours du projet.

- Archives Unleashed Toolkit et Jupyter Notebooks :
  - Certains testeurs ont demandé à pouvoir utiliser une infrastructure plus performante pour permettre de faire tourner des requêtes sur un plus gros volume de données, notamment les requêtes concernant le plein texte ;
  - de façon générale, plusieurs testeurs se sont montrés curieux de découvrir ces notebooks, mais ont fait remarquer qu'ils manquaient de pistes pour mobiliser les résultats ou les requêtes dans une recherche, ou encore pour personnaliser les requêtes. L'élaboration de Jupyter Notebooks en lien avec des cas d'usage recherche réels et susceptibles d'inspirer la formulation de questionnements similaires nécessite un travail approfondi, à la frontière entre le travail scientifique et le travail de médiation sur la collection. L'enrichissement de cette palette de notebooks pourrait être l'un des apports d'un travail en réseau entre établissements de l'ESR où seraient déployés les futures capsules : des formations à l'utilisation de Jupyter notebooks pourraient être proposées en début de projet, et déboucher en cours de projet sur la réalisation de notebooks spécifiques à une approche disciplinaire s'appuyant sur les travaux pédagogiques ou les usages recherche des capsules dans les différents établissements.

IV) Implantation de la capsule à l'Université de Lille

Les lieux

Dans la convention d'application juridique signée entre la BnF et l'Université de Lille, deux salles du Service Commun de Documentation de l'Université sont identifiées comme emprises de la BnF, situées dans deux de ses implantations : l'une à la Bibliothèque Universitaire Droit-Gestion, et l'autre à Lilliad learning center. Les bibliothèques choisies sont celles à proximité des communautés disciplinaires qui étaient à priori les plus susceptibles d'être intéressées par les archives du web : sociologie, sciences de l'information et de la communication, sciences politiques, histoire…

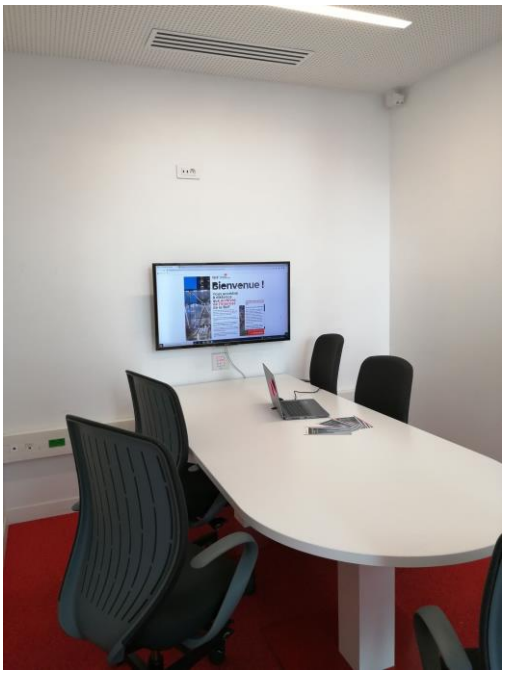
Les salles

Le choix des salles dans ces deux bibliothèques a été fait en concertation avec les équipes qui gèrent les équipements afin de concilier de la manière la plus souple possible les tests de la capsule et les demandes d'utilisation des espaces en provenance des usagers habituels des bibliothèques. D'autres critères ont orienté ces choix : choix d'espaces calmes, présence de tables et chaises de bureau dans la salle (mobilier haut), accessibilité du réseau wifi, salles intégrées ou intégrables dans le système centralisé de réservation d'espaces. Il a également été décidé de proposer la consultation des archives du web à proximité des postes de consultation multimédia de l'Institut national de l'audiovisuel.

Dans le cadre de l'expérimentation, le nombre restreint de tests et leur irrégularité, ainsi que des problématiques d'usages différents sur les deux bibliothèques d'implantation ont conduit à opter pour deux solutions différentes. A Lilliad, la forte tension sur les places disponibles a conduit au choix d'une salle de travail en groupe réservable à la fois par les usagers habituels et par l'équipe ResPaDon, l'équipe pouvant bloquer la réservation de la salle avant les autres usagers. A la BU Droit-Gestion, le poste de consultation des archives du web a été installé dans une salle dont l'usage était jusque-là peu défini. Le déplacement des postes de consultation multimédia de l'INA dans cette salle a permis de consolider une fonctionnalité "recherche" plus affirmée pour cet espace.

Une signalétique multiformat a été développée aux couleurs du projet ResPaDon pour permettre d'identifier la localisation du point d'accès et en assurer la communication : kakemono, affiches, flyers, écrans dynamiques…

BU Droit-gestion, salle recherche



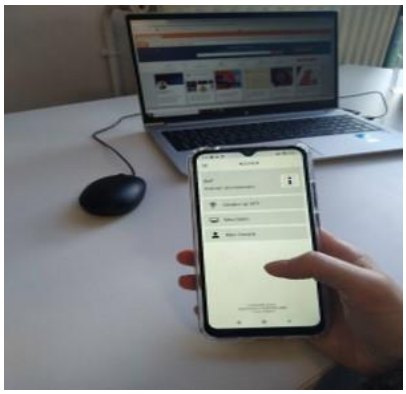
Lilliad Salle 1S01

L'équipement

Dans chaque point d'accès, le SCD a fourni le matériel nécessaire à la connexion et à la consultation des capsules, à savoir un ordinateur portable et un téléphone portable permettant la connexion sécurisée aux applications des archives du web. Les services informatiques du SCD ont commandé et installé le matériel ; leurs échanges avec les interlocuteurs de la BnF leur ont permis de configurer les accès distants au système d'information de la BnF et de trouver des solutions pour résoudre les difficultés apparues au fil du processus.

L'implémentation technique expérimentale des outils de fouille et de visualisation de données faisait que certaines fonctionnalités de datavisualisation ne pouvaient s'afficher que sur un tiers de l'écran de l'ordinateur portable ce qui en rendait la lecture particulièrement ardue. Aussi, il est rapidement apparu intéressant de connecter l'ordinateur portable à un écran de grande dimension permettant d'avoir une vue plus confortable. Les retours d'expérience des chercheurs ont aussi pointé la nécessité de disposer d'une souris pour faire des sélections plus précises et pouvoir sélectionner et naviguer entre les différentes fonctionnalités avec plus de facilité et de précision.

Lors de l'expérimentation, l'ajout d'une borne wifi a été nécessaire pour assurer une meilleure fluidité de la consultation. Les testeurs ont aussi pointé l'importance d'un confort lumineux et thermique, sur des temps de consultation pouvant atteindre plusieurs heures consécutives.



Recommandations

<u>Recommandations n°10</u> : Aménager et équiper des espaces d'accueil en impliquant les services informatiques de l'Université dès le début du projet.

- **Des points de consultation à proximité des usages de recherche** : pour favoriser l'attractivité, la visibilité et les usages du dispositif, il est souhaitable de proposer l'accès aux archives du web dans les espaces les plus proches des unités de recherche et chercheurs qui en seront les utilisateurs les plus probables / nombreux (SHS, sciences politiques notamment), ou dans un "espace chercheurs" dans lequel des habitudes de fréquentation existent déjà. Par ailleurs, la proximité du poste de consultation archives du web avec un poste de consultation multimédia (PCM) de l'INA, peut être intéressante pour favoriser l'attractivité des deux accès et encourager la complémentarité des approches.

- **Une salle permettant la cohabitation avec d'autres usages** : dans le cadre législatif actuel, la consultation des archives du web ne pouvait se faire que dans une salle désignée dans la convention d'application juridique. La recommandation est de choisir une salle permettant la cohabitation de plusieurs usages. La couverture wifi de la salle doit être de très bonne qualité. Le choix d'une salle calme, disposant d'un confort matériel, lumineux et thermique est recommandé, les séances de consultation pouvant se prolonger plusieurs heures consécutives.
  La gestion des créneaux de consultation nécessite que la salle soit identifiée dans un système centralisé de réservation, accessible aux membres de l'équipe qui gèrent l'accès, ou directement via les chercheurs, en fonction des modalités d'accès qui seront choisies.

- **Du matériel adapté** : pour chaque point de consultation il est nécessaire de prévoir un ordinateur (portable ou fixe en fonction des contraintes de gestion de salles). En fonction de la solution d'accès distant retenue par la BnF à l'issue de cette expérimentation, il pourra être nécessaire d'acquérir un smartphone avec accès wifi. L'expérimentation a de plus démontré la nécessité de périphériques associés au poste informatique : une souris pour faciliter la navigation notamment dans les outils de visualisation, un casque pour consulter les contenus audio/vidéo, ainsi qu'un écran de grande taille (avec, le cas échéant, la connectique permettant de relier le tout), indispensable pour permettre de visualiser correctement les résultats des

manipulations effectuées avec les outils de fouille et de visualisation. Un document recensant les matériels à acquérir par les établissements accueillant un point d'accès devra être fourni.

- **Mobilisation des personnels du SCD sur le choix des espaces** : la concertation avec les équipes en charge des espaces publics / des espaces chercheurs est indispensable afin que l'implantation soit pertinente mais aussi qu'elle s'articule au mieux avec d'éventuels autres usages des espaces choisis. La mobilisation de compétences et/ou services en conception de signalétique et en communication est utile à la fois pour matérialiser les espaces et communiquer sur son existence et son actualité de manière régulière, répétée et ciblée.

- **Implication des services supports informatiques dès le début du projet :** la mobilisation des services informatiques de l'établissement d'accueil est indispensable, à la fois dans la phase de commande de matériel, mais aussi pour la configuration des ordinateurs et téléphones, des paramètres réseaux et des accès aux applications des archives du web. L'expérimentation a fait la preuve que la mise en relation directe des services informatiques de l'établissement avec les interlocuteurs de la BnF permettaient de parvenir à des solutions techniques à partir de situations complexes. La recommandation serait de débuter les échanges entre les deux services de manière très précoce lors de l'installation d'une capsule, pour identifier au plus tôt les problématiques et se donner le temps de dénouer les problèmes.

V) Médiation et accompagnement à l'usage des collections

Le rôle de médiateur : formation et compétences

Les équipes de l'Université de Lille et de la BnF ont travaillé ensemble pour développer un service d'accompagnement visant à faciliter la compréhension du contexte de constitution de la collection et à abaisser le coût d'entrée.

L'objectif est d'accompagner les chercheurs, et de faciliter la découverte et l'appropriation des archives du web, avec l'objectif de les rendre plus rapidement accessibles et plus facilement appropriables.

Au service commun de documentation de l'université de Lille, la fonction de « médiateur des archives du web » a été créée, avec pour mission d'accompagner les tests des chercheurs et de faciliter leurs premiers pas dans la capsule. A la suite de la sollicitation des chefs de département et des présentations du projet faites en interne, six agents du SCD se sont portés volontaires pour ce rôle. Ces agents étaient rattachés au département des "services à la recherche et aux chercheurs", au département de la politique documentaire, au département "animation culturelle scientifique et technique", ou encore à l'équipe d'accueil des bibliothèques. Toutes les catégories (A,B,C) étaient représentées parmi les médiateurs.

Deux membres de la BnF se sont déplacés à Lille à deux reprises pour former les médiateurs, à raison de deux journées pour la "capsule découverte", et de deux journées pour la "capsule élections 2002".

Lors des formations, les médiateurs ont bénéficié d'apports théoriques, de temps pratiques d'exploration des archives du web sur chacune des deux capsules, et d'échanges avec les équipes de la BnF. Ils se sont approprié les contenus appris et en ont restitué les points essentiels dans un document qui est devenu le support de présentation du dispositif aux chercheurs lors de leurs tests.

Plusieurs éléments ont permis de consolider et d'entretenir les connaissances et compétences des médiateurs, ainsi que leur confiance pour assurer ce rôle particulier auprès des chercheurs :

- la création de documents « support » permettant avant un test de se remettre en tête les informations utiles : récapitulatif des procédures de connexion, guide de toutes les étapes d'accompagnement d'un test, synthèse des points de la Charte juridique, tutoriels des outils mis à disposition dans les capsules ;
- un échange de 2h en visio avec les collègues de la BnF pour faire des rappels sur le fonctionnement des capsules, et compléter les connaissances sur les outils les plus complexes proposés dans les capsules ;
- des entraînements à l'utilisation des fonctionnalités des capsules archives du web à l'initiative des médiateurs qui en ressentaient le besoin ;

- la proposition de sessions de test à des collègues du SCD pour permettre aux médiateurs de s'entraîner à présenter les archives du web et à dérouler toutes les étapes de l'accompagnement ;
- la mise en place d'un document partagé entre les médiateurs permettant d'y faire figurer et de consulter les questions et problèmes techniques survenus, et la manière de les résoudre.

Au fil de l'expérimentation, au-delà des éléments dispensés lors des formations de la BnF, certaines compétences des médiateurs se sont révélées particulièrement utiles pour faciliter leur appropriation du média et du dispositif et pour assurer un accompagnement des chercheurs adapté : une bonne compréhension du processus de recherche (notamment en sciences humaines et sociales) et de la manière de mobiliser les outils d'exploration et de fouille, une aisance à la manipulation des outils informatique et des connaissances sur les notions juridiques de propriété intellectuelle.

Recommandations

Recommandation n°11 : Mettre en place un réseau d'acteurs locaux et nationaux pour conduire les actions de sensibilisation, formation et communication sur les archives du web

- Étendre la logique de travail en réseau qui a fait ses preuves dans le projet ResPaDon en créant de noeuds locaux de professionnels de l'IST et de chercheurs qui pourraient porter la démarche de sensibilisation et de communication sur cette source (interventions, séminaires, offre de formation, formations de formateurs, etc.). L'animation du réseau des médiateurs et de l'ensemble des acteurs concernés serait une pièce centrale du dispositif pour permettre la formation continue des professionnels de l'IST sur les archives du web.

Recommandation n°12 : Nommer deux médiateurs par point d'accès, sur la base d'une quotité de travail de 30% leur permettant de se former, d'accompagner les chercheurs et de participer à la vie du réseau.

Formaliser le rôle de médiateur :

- Identifier des médiateurs coutumiers du travail avec les chercheurs et avec les manipulations informatiques. Une familiarisation avec le processus de recherche et le fonctionnement des recherches et des outils de traitement et d'analyse permet notamment des échanges plus constructifs avec le chercheur.
- Une des pistes à envisager serait de se limiter à deux médiateurs par point d'accès, afin qu'ils puissent consacrer un temps conséquent à la formation continue et aux échanges avec d'autres médiateurs dans d'autres universités.
- Faire figurer le rôle de "médiateur des archives du web" sur la fiche de poste des agents concernés sur la base d'une quotité de 30% afin qu'ils puissent consacrer du temps à la formation, à l'accompagnement des usagers.

Recommandation n°13 : Organiser la formation initiale et le maintien des compétences des médiateurs en s'appuyant notamment sur le réseau et des organismes de formation :

- Prévoir un temps initial de formation de deux jours incluant apprentissages théoriques et temps de démonstration et de pratique.

- Prévoir deux jours de formation supplémentaires aux outils d'exploration enrichie des contenus et d'aide à la fouille de texte et de données pour permettre un meilleur accompagnement des chercheurs sur ces outils.
- Prévoir un système d'échanges (document collaboratif, canal de discussion…) entre médiateurs pour le partage d'expériences et la résolution de problèmes, et des temps d'échanges.
- Organiser des sessions d'entraînement à intervalles réguliers en s'appuyant sur la documentation mutualisée.
- Proposer avec l'aide des URFIST ou les CRFCB des formations générales au traitement et à la fouille des données et d'initiation aux outils de visualisation, en complément des formations ciblées sur les archives du web.

L'organisation de l'accueil des chercheurs

Lors de l'expérimentation les créneaux de test étaient accessibles sur rendez-vous. Pour pouvoir bénéficier d'un créneau, les chercheurs prenaient rendez-vous en écrivant à une adresse mail générique centralisant toutes les demandes. Un créneau de test leur était alors proposé en fonction de leurs disponibilités, de celles des médiateurs, et de celle du lieu de test souhaité (BU Droit-Gestion ou Lilliad).

La salle était réservée généralement pour une durée de 3h30, incluant le temps d'installation du matériel avant l'arrivée du chercheur. L'expérimentation a montré qu'il était judicieux, dans la mesure du possible, de réserver la salle pour une durée plus étendue que le souhait initial du chercheur. Il est en effet souvent arrivé que le chercheur, immergé dans ses explorations, ait du mal à s'arrêter d'explorer les archives du web et prolonge sa séance. Les étudiants en master qui ont testé le dispositif avec une recherche précise à effectuer ont été accueillis sur des durées plus courtes.

Les parcours de tests

Deux parcours de tests étaient proposés aux utilisateurs :

- Un parcours découverte débutant par une présentation et des échanges consacrés au cadre législatif, documentaire et technique de la collecte du web et permettant l'appropriation des principaux outils proposés, puis permettant un temps d'exploration libre des collections d'archives du web via les applications Archives de l'internet et Archives de l'internet Labs ;

- Un parcours approfondissement, centré sur les outils d'exploration enrichie de la capsule élections 2002. En plus des explications liées à la capsule découverte, les utilisateurs bénéficiaient d'une présentation des collectes électorales, des outils avancés et des données proposées permettant l'analyse et la fouille de corpus web. Ils avaient ensuite la possibilité d'utiliser librement le dispositif.

Il était possible de bénéficier d'un rendez-vous avec des experts de la BnF pour un accompagnement avancé ; cette possibilité n'a pas été utilisée pendant l'expérimentation.

L'accompagnement des chercheurs par les médiateurs

Lors de chaque session de test des archives du web, le médiateur accueillait et installait le chercheur dans la salle, et suivait toutes les étapes nécessaires pour connecter l'ordinateur aux serveurs de la BnF. Il assurait ensuite une présentation incluant :

- une description succincte du projet ResPaDon dans lequel l'expérimentation se tenait,
- une explication du cadre juridique de l'expérimentation et de ses implications, complétée par la signature de la Charte juridique par le chercheur,
- une courte introduction aux archives du web sur la base du support produit lors de la formation des médiateurs,
- des explications sur le fonctionnement des capsules à disposition et les fonctionnalités présentes,
- des explications complémentaires portant sur les outils de recherche, de fouille et de visualisation disponibles (pour le parcours approfondissement).

Cette présentation durait 20 minutes au minimum pour les parcours découverte, mais pouvait être beaucoup plus longue pour les explications liées aux deux capsules, et également en fonction des questions et des échanges avec le chercheur.

Le chercheur disposait ensuite d'un temps autonome de navigation dans les archives du web, pendant lequel il pouvait recontacter le médiateur (téléphone ou mail) pour un support spécifique, une réponse à une question, ou la résolution d'un problème technique. Dans cette phase d'expérimentation, il est souvent arrivé à l'équipe des médiateurs de contacter les collègues de la BnF pour un dépannage technique, des réponses à des questions posées par les utilisateurs, ou un besoin de documentation complémentaire. Ces éléments ont été précieux pour améliorer le dispositif et la documentation au fil de l'eau. Le temps d'exploration était au minimum d'une heure pour une navigation simple dans la capsule découverte. Dès lors que le chercheur s'intéressait également aux outils d'exploration enrichie, ou à un objet de recherche précis, le temps de la session pouvait s'étendre sur plusieurs heures, et le chercheur pouvait décider de revenir plusieurs fois.

Recommandations

Recommandation n°14 : créer des supports de médiation et d'accompagnement mutualisés et personnalisables et organiser des modalités d'échanges de pratiques et de retours d'expérience inter-établissement

- **Créer un support de médiation mutualisé et personnalisable** permettant de présenter les archives du web et les fonctionnalités. Le support réalisé pendant l'expérimentation à l'université de Lille peut être décliné et enrichi à cette fin.
- **Partager des pratiques et des questions/réponses entre établissements** accueillant des capsules et enrichir au fil de l'eau une documentation détaillée et mutualisée sur les archives du web et les outils mais aussi sur les aspects techniques, juridiques, les procédures de connexion et de résolution de problèmes. Cela rejoint la préconisation n°12 du projet ResPaDon qui prévoit de "Mettre en œuvre une co-animation du réseau

par les nœuds et les acteurs nationaux : documentation mutualisée, rencontres régulières, échanges de pratiques…".

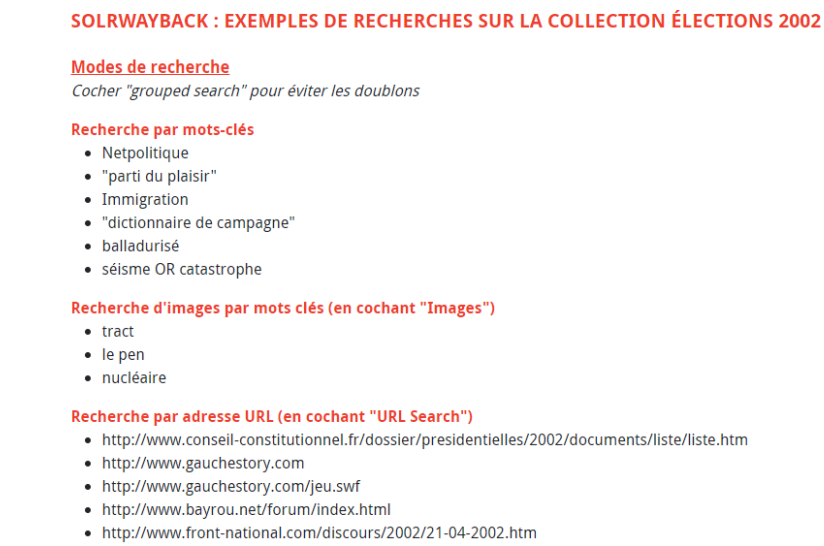
- **Créer un document qui récapitule toutes les étapes de l'accompagnement** pour que les médiateurs puissent le consulter si besoin avant chaque nouvelle séance (en particulier si les sessions ne sont pas régulières dans le temps). Le document créé à l'université de Lille peut être repris et adapté.
- **Réserver une plage horaire confortable** pour la séance de consultation incluant le temps d'installation du matériel avant l'accueil du chercheur, l'accueil et la présentation, et le temps d'exploration autonome du chercheur.

La documentation complémentaire

Documents mis à disposition sous forme imprimée

La documentation a été identifiée comme un enjeu important pour permettre la découverte et l'appropriation des collections et des outils proposés dans la capsule. La documentation des applications disponibles dans les deux capsules a été enrichie au fil de l'eau en fonction des demandes des médiateurs et des usagers.
Elle concerne aujourd'hui :

- les modalités de connexion (pas à pas),
- l'archivage du web et ses techniques (glossaire, bibliographies),
- les collections et leurs modalités de constitution, avec un focus sur les collections web électoral en général (bilans techniques et documentaires des collectes, explicitation de la politique documentaire), et sur la collection élections 2002 en particulier,
- des cas d'usages recherche des collections d'archive web : un document décrit la méthode adoptée par plusieurs projets de recherche ayant porté sur les archives web conservées à la BnF ces dernières années et s'appuie sur le travail d'analyse conduit dans le WP usages,
- les outils et applications proposés dans la capsule : à la documentation intégrée aux applications s'est ajoutée une description des outils d'exploration enrichie proposés dans la capsule élections 2002. Des exemples de recherche adaptés ont également été fournis pour chacun des outils, afin d'aider à démarrer.



SOLRWAYBACK : EXEMPLES DE RECHERCHES SUR LA COLLECTION ÉLECTIONS 2002

Modes de recherche

Cocher "grouped search" pour éviter les doublons

Recherche par mots-clés

- Netpolitique
- "parti du plaisir"
- Immigration
- "dictionnaire de campagne"
- balladurisé
- séisme OR catastrophe

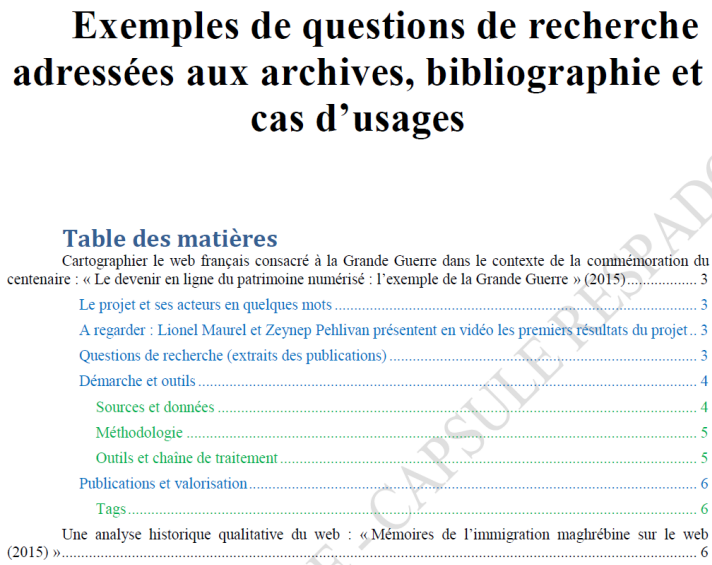
Recherche d'images par mots clés (en cochant "Images")

- tract
- le pen
- nucléaire

Recherche par adresse URL (en cochant "URL Search")

- http://www.conseil-constitutionnel.fr/dossier/presidentielles/2002/documents/liste/liste.htm
- http://www.gauchestory.com
- http://www.gauchestory.com/jeu.swf
- http://www.bayrou.net/forum/index.html
- http://www.front-national.com/discours/2002/21-04-2002.htm

Exemples de recherches pour guider les usagers dans l'utilisation de SolrWayback et favoriser la découverte de la collection élections 2002



Exemples de questions de recherche adressées aux archives, bibliographie et cas d'usages

Table des matières

Document rassemblant des cas d'usage des archives du web dans le cadre de travaux de recherche

Recommandations

Il s'avère qu'il est nécessaire de rendre disponible la documentation hors de la capsule : une partie importante de cette documentation était intégrée aux outils, dans les rubriques aide ou les aides contextuelles. Les retours de tests soulignent le besoin de disposer d'une documentation accessible en dehors de la capsule en amont et en aval de la visite ; à leur demande, celle-ci a été imprimée et aussi envoyée par mail à de nombreux testeurs. Une plateforme en accès libre centralisant la documentation ainsi que, plus généralement,

l'ensemble des données dérivées et métadonnées libres de droit et permettant de préparer sa visite constituerait une amélioration importante du dispositif.

Recommandations n°15 : **Créer un bac à sable à usage pédagogique et de recherche en accès libre**
Ce bac à sable permettrait de regrouper la documentation, les cas d'usages, des outils de recherche dans les métadonnées et d'exploration des collections, la présentation de résultats de recherche, une version open source des applications avec des corpus libres de droit ou dont les droits ont été négociés ; la conception de ce bac à sable s'appuiera sur l'ensemble des résultats des tests de la capsule déployée à Lille pour consolider l'ergonomie, le design et l'interactivité du parcours usager. Il doit être pensé en complémentarité et articulation étroite avec l'offre d'outils enrichis accessibles uniquement au sein des capsules et permettre d'offrir une diffusion plus large aux méthodes et outils disponibles, ainsi que de préparer ou de prolonger sa visite. Enrichi par le réseau d'établissements, Il faciliterait le travail de prise en main des capsules par les établissements suivants et l'autonomie des chercheurs dans l'appropriation de ces collections pour des usages pédagogiques ou recherche.

- **Approfondir les cas d'usages** : l'un des objectifs de la capsule était de donner à voir différents usages des archives du web, d'offrir une large palette d'outils pour permettre aux chercheurs de se projeter dans une démarche de recherche et un cheminement méthodologique. La capsule a été une première réponse à cet enjeu. La présentation initiale faite par les médiateurs permet de lever les premiers obstacles à l'usage des archives et d'accélérer leur prise en main, et, du point de vue de la documentation, la présentation des projets de recherche accueillis à la BnF et de leurs démarches et méthodologies a rapidement été identifiée comme importante en complément de l'offre d'outils. Les tests confirment toutefois le besoin d'un accompagnement plus poussé sur la façon de mobiliser les outils, les données, les indicateurs fournis, et, *in fine*, de transcrire une question de recherche en une démarche d'exploration des archives du web articulant ces différents outils. Cette facilitation reste un vrai défi et le travail conduit pendant l'expérimentation gagnera à être poursuivi. La démarche de documentation des outils et des collections, l'illustration des méthodologies au travers de cas d'usage approfondis et des collections doit donc être approfondie, ainsi que la démarche de médiation. Ce constat corrobore les résultats de travaux conduits à l'étranger sur les archives du web, et sous-tend par exemple la démarche d'animation de communautés et de construction d'outils conduite par le projet Archives Unleashed, ou encore l'élaboration de la section "Web archives" du GLAM Lab[2]. La *sandbox* en cours de développement par le consortium international pour la préservation de l'Internet s'inscrit dans cette même perspective.

[2] Nick Ruest, Jimmy Lin, Ian Milligan, Samantha Fritz. The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives. Proceedings of the 2020 IEEE/ACM Joint Conference on Digital Libraries (JCDL 2020), Wuhan, China.
Tim Sherratt. (2021). GLAM Workbench (version v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.5603060

Les actions de valorisation et de communication

Une stratégie de communication multiforme a été menée pour faire connaître l'expérimentation et inciter les chercheurs à venir tester les archives du web.

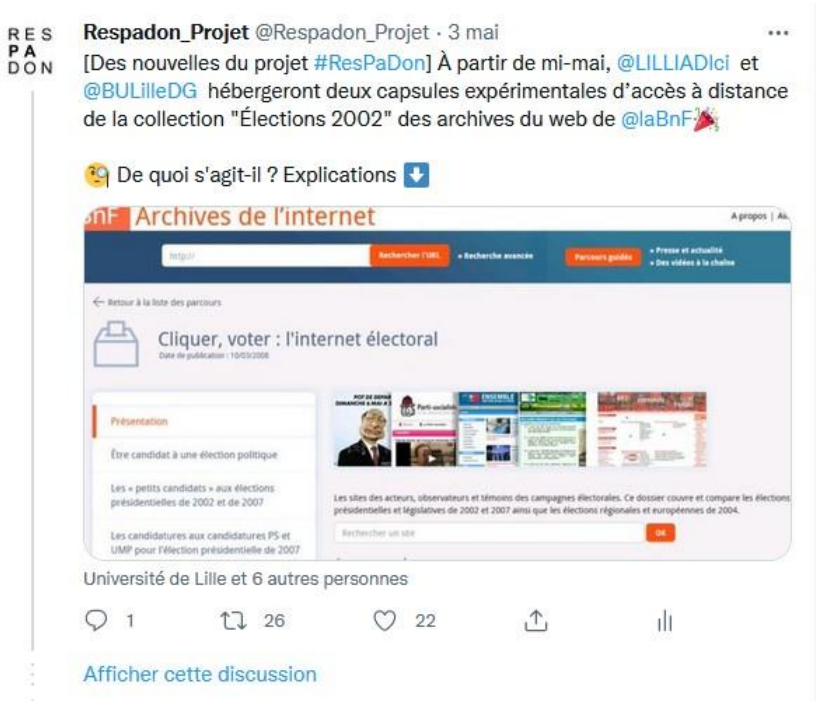
Toute la communication visuelle s'est appuyée sur la charte graphique du projet ResPaDon ce qui a apporté une cohérence aux affiches, kakemono, vignettes et visuels développés.

Des actions ont été développées à l'échelle de l'université : articles dans la lettre recherche de l'université, campagne de posts sur les réseaux sociaux, affichage, déploiement de kakemono, webinaire animé par des personnels de la BnF sur la fabrique des archives du web avec des retours d'expérience et présentation de travaux de chercheurs ayant travaillé sur les archives du web.

Les unités de recherche ont par ailleurs été visées par plusieurs types d'actions : présentation lors de l'assemblée générale du laboratoire de sciences politiques, campagnes de mailing et prises de contact avec des directeurs d'unité.

Des actions ont aussi été conduites directement au niveau des chercheurs : ciblage et invitation personnalisée par mail de chercheurs potentiellement intéressés par les archives du web à partir de l'étude de leurs profils (sujet de recherche, sources et méthodologies habituellement employées…), prises de contact personnelles via des rencontres lors d'événements.

Globalement, et malgré la somme des efforts déployés en ce sens, il s'est avéré finalement assez compliqué pendant l'expérimentation d'attirer et de convaincre les chercheurs de venir tester les archives du web. Les actions qui ont été le plus efficaces pour amener les chercheurs à tester les archives du web sont les contacts inter-personnels. Le manque de temps et l'absence de familiarité avec la source sont les raisons principales évoquées par les chercheurs qui ne souhaitaient pas tester le dispositif car ils n'y voyaient pas un intérêt immédiat pour leur recherche.



Communication sur le fil Twitter du projet ResPaDon à l'occasion de l'ouverture de la capsule

Recommandations

Recommandation n°16 : développer un kit de communication mutualisé et construire des actions de sensibilisation en s'appuyant sur le réseau des établissements partenaires et en impliquant les chercheurs

- **utiliser une charte graphique identifiable** pour développer des supports de communication multiformes (kakemono, réseaux sociaux, lettres d'actualité, mailing…), proposer aux établissements du réseau un kit de communication personnalisable et adaptable,
- **diversifier et cumuler les canaux de communication envers les chercheurs** : actions au niveau de l'établissement, des unités de recherche, des chercheurs individuels, des écoles doctorales, des masters,
- **co-construire des actions de sensibilisation (séminaires, interventions…) avec les unités de recherche** en s'appuyant sur le réseau des établissements partenaires,
- **intégrer des retours d'expérience de chercheurs dans les actions de sensibilisation** à destination des unités de recherche, en s'appuyant notamment sur l'expérience de chercheurs de l'unité de recherche.

VI) Les testeurs et usages de la capsule

Au total, 49 personnes ont testé la capsule entre mai 2022 et juin 2023, avec des profils variés : 11 personnels IST, 10 chercheurs et doctorants et 28 étudiants en master. Les personnels IST ont été accueillis à la fois pour permettre aux médiateurs de s'entraîner dans leur présentation de la source et des outils et aussi par curiosité professionnelle pour ce nouveau service. Les chercheurs ont été accueillis dans une démarche d'exploration avec ou sans question de recherche particulière. Quant aux étudiants, leur accueil s'est déroulé dans le cadre d'un cours organisé par des enseignants.

Les usages recherche

Les profils des chercheurs

- Chercheur en sciences de l'information et de la communication. Thème : histoire et épistémologie de la philologie, édition critique de textes médiévaux.
- Doctorant en histoire. Thème : historiographie médiévale
- Chercheur en sciences de l'information et de la communication. Thème : web électoral
- Chercheur en science politique. Thème : sociologie du numérique
- Chercheurs en sociologie. Thème : réseaux sociaux
- Doctorant en sociologie. Thème : sociologie des migrations
- Doctorant en sociologie. Thème : sociologie du travail
- Chercheur en sciences de l'information et de la communication. Thème : le numérique, la transformation des organisations
- Chercheur en sciences de l'information et de la communication. Thème : web électoral, vote électronique
- Chercheur en science politique. Thème : visibilité médiatique

Les tests ont permis de recueillir des retours d'usage approfondis concernant les interfaces, outils et fonctionnalités proposés dans la capsule et sont décrits plus haut en partie III.

Les courts entretiens menés à l'issue des tests ont en outre permis de retracer certains parcours d'exploration des archives du web décrits ci-dessous. Les attentes étaient variées : plusieurs testeurs sont venus pour découvrir une nouvelle source, des outils et des méthodes, d'autres venaient avec une question de recherche précise voire des listes d'adresses de sites. Les étudiants en master quant à eux avaient des questions de recherche à investiguer en temps limité.

Des exemples de démarches d'exploration

Doctorante en sociologie

Une doctorante vient soumettre son sujet de recherche aux archives du web. Elle arrive en ayant préparé une liste d'adresses URL d'entreprises liées à son sujet de thèse, avec la perspective d'examiner les visuels et les contenus mis en avant sur leurs sites web au fil du temps et plus globalement de mesurer l'évolution de la présence numérique et des discours des entreprises au fil du temps.

Ses premières impressions sont très positives, notamment sur l'ergonomie de l'application Archives de l'internet qui favorise la recherche de données ; elle trouve la présentation et l'indexation des sources claires et apprécie la facilité de navigation sur l'application. Sa connaissance antérieure de la Wayback Machine d'Internet Archive lui permet de comparer les deux systèmes, et elle évalue la navigation plus fluide et constate que les archives de l'internet de la BnF sont plus complètes que les archives disponibles via la Wayback Machine. Elle apprécie les conditions matérielles de l'accès qui lui permettent de naviguer sans être gênée par des temps de latence dûs au chargement des pages.

La consultation des différents sites archivés au fil du temps lui a permis de trouver des informations et listings dont elle n'avait jusqu'alors pas trouvé de trace lors de ses recherches dans d'autres sources ; elle a également pu étudier l'évolution du type d'image mis en avant par les entreprises pour les représenter, et répertorier la teneur des messages délivrés au fil du temps et leur évolution.

Même si la collection élections 2002 n'était pas réellement adapté à sa recherche, la doctorante a tout de même exploré les modules de fouille et d'analyse mis à disposition dans la SolrWayback, et a pu en tirer quelques enseignements, en réussissant à se projeter dans l'utilisation qu'elle pourrait en faire sur son corpus idéal. Elle a observé que le Word Cloud permettait d'approcher les sites par les plus fortes occurrences de mots utilisés et d'avoir une familiarisation visuelle d'un contenu à la base complexe. La visualisation des résultats par domaine lui a permis de mesurer l'évolution de la fréquence d'utilisation de certains de ses mots-clés. A partir de l'URL d'un site web, elle a aussi utilisé les graphes de liens qui présentent l'ensemble des sites reliés au site source, et en en découvrant le potentiel elle regrettait que lors de cette phase expérimentale le corpus ne soit pas plus adapté à sa recherche, car elle imaginait tout à fait utiliser cet outil pour visualiser le réseau de chacune de ses entreprises, et l'identification d'acteurs liés peu visibles par ailleurs.

Chercheuse en sciences de l'information et de la communication

La chercheuse explore les archives du web pour étudier de quelle manière certains sites ont évolué au fil du temps, en repérant la modification de la structure de site, l'ajout d'outils… Elle avait préparé une liste d'URL de sites à consulter. Elle a pu repérer d'autres sites à partir d'une recherche avec l'outil n-gram et également en utilisant la recherche classique et en triant les résultats grâce aux facettes disponibles. Elle compte utiliser les résultats de ses recherches dans les archives du web en complémentarité avec d'autres sources et notamment en explorant le web vivant. *« C'est un moyen aussi de faire de la recherche à un degré un peu plus général, avec une vision d'évolution historique et aussi l'occasion de faire un peu de quantitatif par rapport à une recherche classique. » « On voit qu'il y a plein de possibilités et ça donne envie, mais peut être que c'est le genre de recherche qui demande un peu de temps pour se lancer dans le sujet et le contexte, il faut avoir un peu de temps devant soi. La difficulté, c'est quand même de pouvoir venir assez longtemps. »*

Chercheur en sciences de l'information et de la communication

Une chercheuse qui pratique déjà la Wayback Machine d'Internet Archive, et qui vient explorer les archives du web avec un thème précis à investiguer. Elle a au préalable fait des recherches dans la presse de l'époque qu'elle étudie, pour pouvoir établir des comparaisons avec ce qu'elle espère trouver dans les archives du web. Elle est venue plusieurs fois utiliser la capsule et a poussé plus loin ses investigations à chaque fois, en testant les nombreuses fonctionnalités disponibles pour trouver des résultats et en comprendre la pertinence. Elle

indique le bénéfice de revenir plusieurs fois en étant plus familière des outils proposés. Elle regrette l'absence de lien technique entre les deux capsules, ayant trouvé d'une part un parcours guidé sur sa thématique, et en parallèle dans la capsule Elections 2002 des outils d'exploration plus performants.

Elle a pu faire une cartographie de sites web grâce aux outils de la boîte à outils. *« Grâce au corpus Web électoral, on trouve des sites qui n'existent plus et qui sont intéressants, rares »*. La consultation des Archives du web dans l'autre capsule lui a permis d'aller consulter ces sites et d'identifier combien de temps ont duré ces sites.

Son expérience de consultation lui fait dire *« ça me fait raisonner autrement, ça alimente une version différente du sujet . »* Grâce aux outils de fouille, elle a notamment pu repérer une série d'acteurs en lien avec sa thématique dont elle ne connaissait pas l'existence.

Au terme de ses séances de consultation, elle indique que plusieurs niveaux d'accompagnement des chercheurs seraient souhaitables : des explications techniques pour faciliter la navigation dans les capsules ; un accompagnement plus important sur les modes d'interrogation et de formulation des requêtes afin de préciser la pertinence des recherches effectuées ; et enfin un accompagnement plus général autour de la problématique de recherche et de la manière de l'explorer dans les archives du web avec l'ensemble des fonctionnalités disponibles. Elle pointe la nécessité de pouvoir discuter avec un ingénieur d'études pour expliquer la recherche et obtenir des conseils. Elle insiste par ailleurs sur la nécessité d'avoir une assistance pour construire une collecte spécifique dans des cas particuliers.

Doctorant en histoire

« J'ai déjà utilisé les archives, notamment Gallica de la BnF, les Annales, mais pas les archives d'INA. J'ai commencé à utiliser Internet Archive pendant la période COVID. J'ai trouvé des ressources pour mes recherches ainsi que les cours. L'utilisation est simple, les parcours variés, cela permet de découvrir beaucoup de choses, cela rend presque nostalgique. J'ai beaucoup aimé la recherche n-gram, cela donne envie de faire de la recherche ! La « neutralité » de ces outils nous permet de nous positionner par rapport à l'opinion publique, de sonder les opinions et des sujets d'une époque. Cette expérience m'a permis de développer quelques idées de recherche. J'aurais aimé trouvé une liste de sites en rapport avec ma thématique. »

Ces retours, collectés par les médiateurs lors de brefs entretiens avec les chercheurs à l'issue des tests, soulignent :

- l'intérêt de la médiation notamment par rapport aux questionnements et étonnements récurrents sur la nature de la source (multiplicité des captures, lacunes, profondeur de l'archive, compréhension des modalités de constitution, de fonctionnement des outils) ;
- l'intérêt et enthousiasme pour la richesse des contenus explorés : plus-value des archives en complémentarité avec d'autres sources habituellement utilisées, notamment pour l'analyse diachronique des discours d'acteurs variés sur une même thématique (associations, acteurs institutionnels, etc.), ou encore des outils d'exploration (graphes de liens, logique de la navigation), pour favoriser l'identification d'informations, d'acteurs, de thématiques jusqu'alors inconnus dans leur recherches, que ce soit par sérendipité ou grâce aux sélections proposées dans les parcours guidés ;

- mais aussi le besoin d'un accompagnement méthodologique approfondi pour traduire une question de recherche sur les archives en une démarche d'exploration des archives web (source longue à prendre en main, multiplicité des outils, etc.) ;
- et certaines frustrations liées à la volonté que les outils d'exploration avancés soient déployés sur leur domaine d'expertise / question de recherche et non seulement sur la collection élections 2002.

Recommandations

Accompagner les projets de recherche dans les établissements de l'ESR

Recommandation n° 17 : Encourager la participation des chercheurs aux collectes de corpus web dans leur domaine d'expertise, afin de favoriser une connaissance plus approfondie des collections proposées.

Les sélections faites dans ce cadre viendraient enrichir les collections de dépôt légal du web comme c'est déjà le cas sur certaines collectes et pourraient également faire l'objet d'une indexation spécifique.

Recommandation n° 18 : Dans les capsules, déployer des outils d'exploration enrichie sur ces corpus co-produits ou définis et extraits avec les équipes de recherche.

La collection élections 2002 servirait de démonstrateur des outils et méthodes avancées d'analyse de corpus web en première phase du projet, pour permettre la découverte de la source, et dans une seconde phase de la vie de la capsule, ces mêmes outils seraient déployés pour permettre l'exploration et la fouille des corpus intéressant les équipes locales. Le déploiement d'une capsule pourrait ainsi s'appuyer sur des projets de recherche en lien avec des pôles d'expertise locaux et permettre dans les années suivantes des projets pédagogiques autour de ces corpus avec les étudiants.

Recommandation n°19 : Accompagner au mieux l'émergence de projets de recherche sur les archives du web en formant un ingénieur de recherche dédié aux réseaux des établissements accueillant des capsules et mutualisé entre ces établissements

Accompagner au mieux l'émergence de projets de recherche sur le archives du web en formant un ingénieur de recherche dédié aux réseaux des établissements accueillant des capsules : cet ingénieur pourrait fournir une aide méthodologique dans la construction d'une démarche de recherche sur les archives du web, et jouer un rôle de passeur des compétences et d'acquis méthodologiques pour ces différents projets, capable d'orienter plus rapidement les chercheurs vers tel outil ou telle démarche spécifique, en complément de l'accompagnement de premier niveau fourni par les médiateurs.

Les usages pédagogiques

De façon générale, et c'est l'une des leçons du dispositif, les usages pédagogiques de la capsule ont été plus importants qu'attendus, plusieurs enseignants-chercheurs ayant demandé à pouvoir envoyer leurs étudiants utiliser la capsule pour leur faire découvrir une nouvelle source et ses caractéristiques à au travers de cas d'usages concrets.

Une session pédagogique a été organisée pour une quinzaine d'étudiants de DEUST dans le cadre d'un cours plus global sur les archives. Ils ont pu découvrir les contenus et fonctionnalités à partir d'exercices proposés par leur enseignant.

Près d'une trentaine d'étudiants de master ont exploité les archives du web à partir de sujets variés proposés par leurs enseignants ou choisis par eux-mêmes. Dans le cadre de leur cours, certains avaient déjà consulté des archives traditionnelles, utilisé Internet Archive, Gallica, ou encore exploré les collections de l'INA.

<u>Exemples de sujets investigués :</u>

- La peine de mort en France depuis le XVIe et jusque son abolition
- L'avènement du MMA (mixed martial arts)
- L'invisibilisation des femmes artistes
- L'évolution de l'Astrologie
- Les bibliothèques comme tiers lieux
- Le bilan écologique et humain de la coupe du Monde de football 2022
- La dictature en Haïti
- L'histoire de la télévision
- L'histoire du féminisme en France depuis 17ème siècle
- Les féminicides
- L'évolution de la représentation visuelle des dragons dans le temps
- L'exploration spatiale

Les étudiants de master ont eu accès à la capsule "découverte". Ils ont effectué des recherches assez différentes les unes des autres et n'ont pas tous utilisé les mêmes fonctionnalités. Quelques exemples d'utilisation :

- consultation du site internet d'un journal à partir de son URL pour retrouver des articles et pages web dont ils connaissaient l'existence mais qui n'étaient pas accessibles sur le site internet actuel du journal, ou antérieurs aux articles accessibles via d'autres biais comme europress
- utilisation des outils disponibles dans l'application "Archives de l'internet Labs" et en particulier du n-gram pour comparer l'évolution de la fréquence d'apparition de certains mots ou concepts dans la collection Actualités,
- utilisation des parcours guidés et des listes API comme clé d'entrée pour identifier des sites internet pouvant avoir un lien avec une question de recherche, puis permettant de rebondir de site en site ;
- recherche par mot clé dans la collection Actualités pour repérer les acteurs impliqués (ou ayant un discours) sur un sujet de société au fil du temps, et évolution de ces acteurs ;
- recherche d'images pour étudier l'évolution de la représentation visuelle d'un objet dans le temps.

Les entretiens menés auprès des étudiants à l'issue de leur consultation ont permis de recueillir des éléments relevant d'un "rapport d'étonnement" :

- au premier abord pour certains, une sorte de scepticisme à devoir utiliser cette source en plus des autres sources déjà mobilisées dans le cadre de leur cours, mais au final

une source jugée intéressante, complémentaire aux autres sources consultées, et des étudiants heureux de l'avoir découverte ("une mine d'or" pour un des étudiants) ;
- la nécessité d'avoir une connaissance préalable du sujet pour pouvoir faire des premières recherches dans les archives ;
- passé l'étonnement du mode particulier de recherche par URL ou de certaines fonctionnalités, une utilisation jugée généralement assez simple, fluide et intuitive pour beaucoup, même si cela n'a pas été le cas pour tous, certains trouvant la manipulation des listes API en particulier compliquée ;
- une difficulté à ne pas connaître et à retrouver des URL de sites disparus en lien avec leur thématique de recherche ;
- une difficulté à identifier dans les collections des sites en lien avec des thématiques marginales ou plutôt émergentes, n'ayant pas de site dédié, ou une présence faible dans la collection actualités ;
- le constat que la préparation des URL à consulter en amont de la consultation représente un gain de temps certain ;
- une satisfaction à trouver des contenus qui n'auraient pas pu être trouvés par un autre moyen, ou encore un étonnement à découvrir l'ancienneté d'utilisation d'un terme ou d'une polémique sociétale

Les tests confirment l'intérêt des archives du web comme matériau pédagogique, mobilisable dans des cadres variés : formation à la recherche d'information (fonctionnement d'un moteur de recherche), étude critique des sources (spécificités de la source par rapport à d'autres ou du média web, sources nativement numériques), ou en sciences sociales (étude du traitement médiatique des événements, étude des représentations, etc.). Ils soulignent l'intérêt de disposer d'un outillage plus poussé pour repérer des URL ou d'une base de supports pédagogiques consolidés.

Recommandations

Recommandation n°20 : Développer les usages pédagogiques des archives du web et en faire une des dimensions importantes de la stratégie de communication et de sensibilisation à cette source :
Favoriser les séances de sensibilisation à vocation pédagogique et **cibler spécifiquement les doctorants en début de thèse** dont les méthodologies ne sont pas figées. Les usages pédagogiques des archives du web et la possibilité pour les enseignants-chercheurs de monter en partenariat avec le SCD des cours de découverte des archives du web ont rencontré un réel succès et ces usages pédagogiques pourraient avoir une place centrale dans le dispositif de communication et de sensibilisation. Outre la sensibilisation des futurs chercheurs et l'inscription de la formation aux archives du web dans le développement de la littératie numérique, l'expérience montre que cette offre de formation permet de plus d'amener les enseignants-chercheurs à utiliser la source dans leurs propres recherches.

Recommandation n° 21 : Permettre l'enrichissement progressif du bac à sable pédagogique et de recherche en accès libre par le réseau.
Ce bac à sable décrit plus haut regrouperait et centraliserait la documentation, les cas d'usages, des outils de recherche dans les métadonnées et d'exploration des collections, la présentation de résultats de recherche, une version open source des applications avec des corpus libres de droit ou dont les droits ont été négociés.

Cet enrichissement pourrait reposer sur la formalisation de travaux de groupes d'étudiants sur une thématique précise, ou sur la mise en récit des méthodologies des chercheurs utilisateurs de la capsule ou sur des entretiens oraux. Il faciliterait le travail de prise en main des capsules par les établissements suivants et l'autonomie des chercheurs dans l'appropriation de ces collections pour des usages pédagogiques ou recherche.

- Proposer dans ce bac à sable la description de sessions pédagogiques de découverte et d'exploration des archives du web accessibles à des étudiants de master, voire d'autres niveaux.
- Mettre à disposition une liste de sujets pouvant être explorés grâce aux archives du web pour faciliter la phase de découverte
- Documenter des parcours-exemple de navigation dans les archives du web pour faciliter une appropriation plus simple de l'environnement, une meilleure compréhension des différentes fonctionnalités et de leur complémentarité
- Permettre des usages exploratoires.

VII) Les perspectives ouvertes par l'expérimentation

A l'issue de cette expérimentation, il est possible de préciser les contours d'un dispositif de capsule réplicable et soutenable dans un réseau d'établissements de l'ESR s'appuyant sur un ensemble de recommandations.

Autoriser le déploiement d'un accès distant aux collections de dépôt légal du web dans les établissements de l'enseignement supérieur et de la recherche.

- **Recommandation n°1 : Faire évoluer le cadre législatif et réglementaire pour permettre l'implantation de points d'accès aux collections de dépôt légal du web dans les emprises des services de documentation des établissements de l'ESR.**
  Cette évolution est nécessaire pour envisager le déploiement de capsules dans un réseau élargi d'établissements et leur pérennisation, au-delà de l'expérimentation conduite durant le projet ResPaDon. Le cadre réglementaire et législatif actuel limite en effet l'accès aux archives du web à un réseau d'établissements partenaires, principalement des bibliothèques publiques, listées dans un arrêté.

- **Recommandation n°2 : Rédiger une convention type associant chaque établissement hébergeant une capsule et la BnF et précisant les obligations des deux parties.** Cette convention type pourrait servir de base aux échanges juridiques entre la BnF et l'établissement se préparant à accueillir une capsule**.**

- **Recommandation n°3 : Implanter des points de consultation dans les espaces fréquentés par les chercheurs, à raison de un à deux points par établissement, dans les locaux des services de documentation qui proposent des ressources documentaires complémentaires et une aide à leur utilisation.**

- **Recommandation n°10 : Aménager et équiper des espaces d'accueil en impliquant les services informatiques de l'université dès le début du projet.**
  Le choix de salles identifiées dans un système centralisé de réservation, proches des unités de recherche (campus SHS et sciences politiques notamment), accueillant le cas échéant le poste de consultation multimédia de l'INA, et permettant la cohabitation avec d'autres usages, est souhaitable pour la visibilité et l'attractivité du dispositif. Les échanges avec les services informatiques de l'université pourraient s'appuyer sur un pas à pas et une liste de prérequis mutualisés, ou une Foire aux questions enrichie au fur et à mesure des déploiements.

Développer un réseau national de médiateurs pour l'accompagnement et la communication autour des archives du web.

- **Recommandation n°11 : Mettre en place un réseau d'acteurs locaux et nationaux pour conduire les actions de sensibilisation, formation et communication sur les archives du web.**
  Il s'agit d'étendre la logique de travail en réseau qui a fait ses preuves dans le projet ResPaDon Le caractère coûteux des actions de communication au niveau local plaide pour un changement d'échelle. La communication et la sensibilisation seraient portées

au sein d'un réseau national d'acteurs professionnels de l'information et équipes de recherche : interventions, séminaires, offre de formation, formations de formateurs, etc.

- **Recommandation n°12 : Nommer deux médiateurs par point d'accès, sur la base d'une quotité de travail de 30% leur permettant de se former, d'accompagner les chercheurs et de participer à la vie du réseau.** L'expérimentation a montré que le caractère essentiel du rôle de médiateur, chargé d'accompagner la découverte des archives, de répondre aux questions des chercheurs et de contribuer à l'organisation de sessions de formation. Il est souhaitable que ces médiateurs soient coutumiers du travail avec les chercheurs et avec les manipulations informatiques. Une familiarisation avec le processus de recherche et le fonctionnement des outils de traitement et d'analyse permet notamment des échanges plus constructifs avec le chercheur.

- **Recommandation n°13 : Organiser la formation initiale et le maintien des compétences des médiateurs en s'appuyant notamment sur le réseau et des organismes de formation.**
  Un temps de formation initiale de 6 jours, contre 4 jours dans l'expérimentation lilloise, apparaît nécessaire pour que les médiateurs soient à l'aise sur l'ensemble des outils et puissent les avoir manipulés suffisamment au cours d'ateliers pratiques. Des sessions d'entraînement à intervalles réguliers doivent être proposées, à partir de ressources mutualisées et mises en ligne. En complément des formations ciblées sur les archives du web, proposer avec l'aide des URFIST ou des CRFCB des formations d'initiation au traitement et à la fouille des données et aux outils de visualisation, semble intéressant pour développer les compétences des médiateurs. L'animation du réseau des médiateurs et de l'ensemble des acteurs concernés par un chef de projet mutualisé (1 ETP) serait une pièce centrale du dispositif pour permettre la formation continue des professionnels de l'IST sur les archives du web.

- **Recommandation n°14 : Créer des supports de médiation et d'accompagnement mutualisés et personnalisables et organiser des modalités d'échanges de pratiques et de retours d'expérience inter-établissements.**
  Le support réalisé pendant l'expérimentation à l'Université de Lille peut être décliné et enrichi à cette fin. Il serait en effet souhaitable de partager des pratiques et des questions/réponses entre établissements accueillant des capsules et d'enrichir au fil de l'eau une documentation détaillée et mutualisée sur les archives du web et les outils mais aussi sur les aspects techniques, juridiques, les procédures de connexion et de résolution de problèmes.

Consolider et enrichir les outils pour favoriser la découvrabilité des collections et répondre aux besoins de recherche.

- **Recommandation n°5 : Mettre en place à la BnF une infrastructure informatique permettant le passage à l'échelle du dispositif expérimental.**
  Il s'agit pour la BnF d'acquérir les serveurs, les licences et les espaces de stockage nécessaires pour déployer des capsules dans un réseau élargi d'établissements et d'organiser la supervision et le support adaptés à des usages hors les murs du système

d'information.

- **Recommandation n°6 : Améliorer la solution d'accès sécurisée aux collections de dépôt légal du web pour la rendre plus robuste et conforme à la politique de sécurité des établissements de l'ESR.**
  Les tests ont porté sur deux solutions d'accès distant différentes : inWebo et WALLIX. La solution cible doit prendre en compte les politiques de sécurité des systèmes d'information des Universités. Le choix d'une solution pourrait faire l'objet d'une validation conjointe par la BnF et des représentants des DSI des établissements partenaires. L'objectif est de faciliter l'installation et la maintenance de la solution tout en garantissant sa conformité aux procédures de gestion des identités et des postes.

- **Recommandation n°7 : Améliorer l'ergonomie de la solution d'accès distant sécurisé pour la rendre conforme aux usages de recherche, notamment en facilitant l'export et le copier-coller des données.**
  Les solutions d'accès distant actuelles ne permettent ni le copier-coller d'extraits de texte, ni l'export de données produites avec les outils d'analyse, de requêtage, de programmation ou de visualisation livrés dans l'environnement sécurisé, données qui ainsi enrichies et transformées constituent le résultat original de la recherche. Ces contraintes imposées par la solution technique doivent être levées dans le respect des conditions d'utilisation des collections.

- **Recommandation n°8 : Consolider et enrichir la palette d'outils d'exploration et d'aide à la fouille de texte et de données proposée dans la capsule, afin d'améliorer la découvrabilité des collections.**
  Les outils proposés dans la capsule (Archives de l'internet, SolrWayback, Jupyter Notebooks et Archives Unleashed Toolkit) font l'objet de corrections et d'évolutions régulières. Ces travaux réalisés par les communautés open source et la BnF devront également intégrer à un rythme régulier la capsule en fonction des retours d'usages.

- **Recommandation n°9 : Travailler sur le design et l'ergonomie de l'interface usager de la capsule en s'appuyant sur les retours de tests pour concevoir un parcours unifié.**
  La consolidation et l'amélioration du parcours usager et du design applicatif apparaissent comme un enjeu crucial pour faciliter l'appropriation des archives du web. Dans la phase d'expérimentation, et pour faciliter la mise en œuvre de la capsule, les outils d'exploration enrichie étaient proposés dans un dispositif technique autonome et distinct du dispositif BDLI existant. La construction d'un seul et unique environnement permettrait d'offrir une facilité et une fluidité dans l'utilisation de la capsule ainsi qu'une meilleure articulation et le renforcement des différentes fonctionnalités. Les fonctions de capture d'écran et de copier-coller étaient impossibles dans le dispositif WALLIX et jugées trop fastidieuses ou insuffisantes dans le dispositif BDLI. Elles doivent être ouvertes et plus intuitives.

Faciliter la prise en main des collections par la construction d'outils pour et par la communauté universitaire.

Plusieurs pistes ont de plus été identifiées au cours du projet ResPaDon pour aller plus loin dans la démarche de facilitation et de médiation et abaisser encore le coût d'entrée méthodologique et technique dans les archives du web.

- **Recommandation n°4 : Décrire, préciser et faciliter les usages qui peuvent être faits des différents types de données relatifs au dépôt légal du web mis à disposition dans les capsules.**
  Un guide à l'usage des chercheurs et une Foire aux questions permettraient de décrire précisément les conditions de consultation, de traitement et d'exploitation des données disponibles dans la capsule, en fonction des différents types de données (données collectées, métadonnées et données dérivées techniques ou documentaires, données transformées ou enrichies) et des usages envisagés (accès, analyse et exploitation dont TDM, copie privée, publication et diffusion).

- **Recommandation n°15 : Créer un bac à sable à usage pédagogique et de recherche en accès libre regroupant et centralisant la documentation, les cas d'usages, des outils de recherche dans les métadonnées et d'exploration des collections, la présentation de résultats de recherche, une version open source des applications avec des corpus libres de droit ou dont les droits ont été négociés**. La conception de ce bac à sable s'appuierait sur l'ensemble des résultats des tests de la capsule déployée à Lille pour consolider l'ergonomie, le design et l'interactivité du parcours usager. Ce bac à sable serait complémentaire et articulé avec l'offre d'outils accessibles uniquement au sein des capsules. Il offrirait une diffusion plus large des méthodes et outils disponibles et permettrait ainsi de préparer ou de prolonger sa visite.

- **Recommandation n°21 : Permettre l'enrichissement collaboratif du bac à sable par le réseau d'établissements où seront déployées les capsules, afin de permettre un processus incrémental et la constitution d'une base de matériaux pédagogiques et de recherche.** Cet enrichissement pourrait reposer sur la formalisation de travaux de groupes d'étudiants sur une thématique précise, ou sur la mise en récit des méthodologies des chercheurs utilisateurs de la capsule ou sur des entretiens oraux. Il faciliterait le travail de prise en main des capsules par les établissements suivants et l'autonomie des chercheurs dans l'appropriation de ces collections pour des usages pédagogiques ou de recherche.

- **Recommandation n°16 : Développer un kit de communication mutualisé et construire des actions de sensibilisation en s'appuyant sur le réseau des établissements partenaires et en impliquant les chercheurs.**

- **Recommandation n°20 : Développer les usages pédagogiques des archives du web et en faire une des dimensions importantes de la stratégie de communication et de sensibilisation à cette source** : les usages pédagogiques des archives du web et la possibilité pour les enseignants-chercheurs de monter en partenariat avec le service documentaire des cours de découverte des archives du web ont rencontré un réel succès. Ces usages pédagogiques pourraient avoir une place centrale dans le dispositif de communication et de sensibilisation. Outre la sensibilisation des futurs chercheurs et l'inscription de la formation aux archives du web dans le développement de la

littératie numérique, l'expérience montre que cette offre de formation permettrait d'amener les enseignants-chercheurs à utiliser la source dans leurs propres recherches.

Accompagner les projets de recherche dans les établissements de l'ESR.

- **Recommandation n°17 : Encourager la participation des chercheurs aux collectes de corpus web dans leur domaine d'expertise, afin de favoriser une connaissance plus approfondie des collections proposées.** Les sélections faites dans ce cadre viendraient enrichir les collections de dépôt légal du web comme c'est déjà le cas sur certaines collectes et pourraient également faire l'objet d'une indexation spécifique.

- **Recommandation n°18 : Dans les capsules, déployer des outils d'exploration enrichie sur ces corpus co-produits ou extraits des archives avec les équipes de recherche** : la collection élections 2002 servirait de démonstrateur des outils et méthodes avancées d'analyse de corpus web pour permettre la découverte de la source. Dans un second temps, ces mêmes outils seraient déployés sur des corpus intéressant les équipes de recherche locales. Le déploiement d'une capsule pourrait ainsi s'appuyer sur des projets de recherche en lien avec des pôles d'expertise locaux et permettre dans les années suivantes des projets pédagogiques autour de ces corpus.

- **Recommandation n°19 : Accompagner au mieux l'émergence de projets de recherche sur les archives du web en formant un ingénieur d'études ou de recherche dédié au réseau des établissements accueillant des capsules** : cet ingénieur fournirait une aide méthodologique dans la construction d'une démarche de recherche sur les archives du web, et jouerait un rôle de passeur des compétences et d'acquis méthodologiques pour ces différents projets, capable d'orienter plus rapidement les chercheurs vers tel outil ou telle démarche spécifique, en complément de l'accompagnement de premier niveau fourni par les médiateurs.

Annexe 4 :

Le livrable WP4

- Site web « Explorer les archives du web avec Hyphe » (retour d'expérience et travaux du datasprint): https://ResPaDon.medialab.sciencespo.fr/

- Réflexions méthodologiques pour étudier de manière complémentaire les archives du web et le web vivant à des fins de recherche : https://drive.google.com/file/d/11jfRnaZFum0eRPetxLb8G0JwuoOBFtps/view

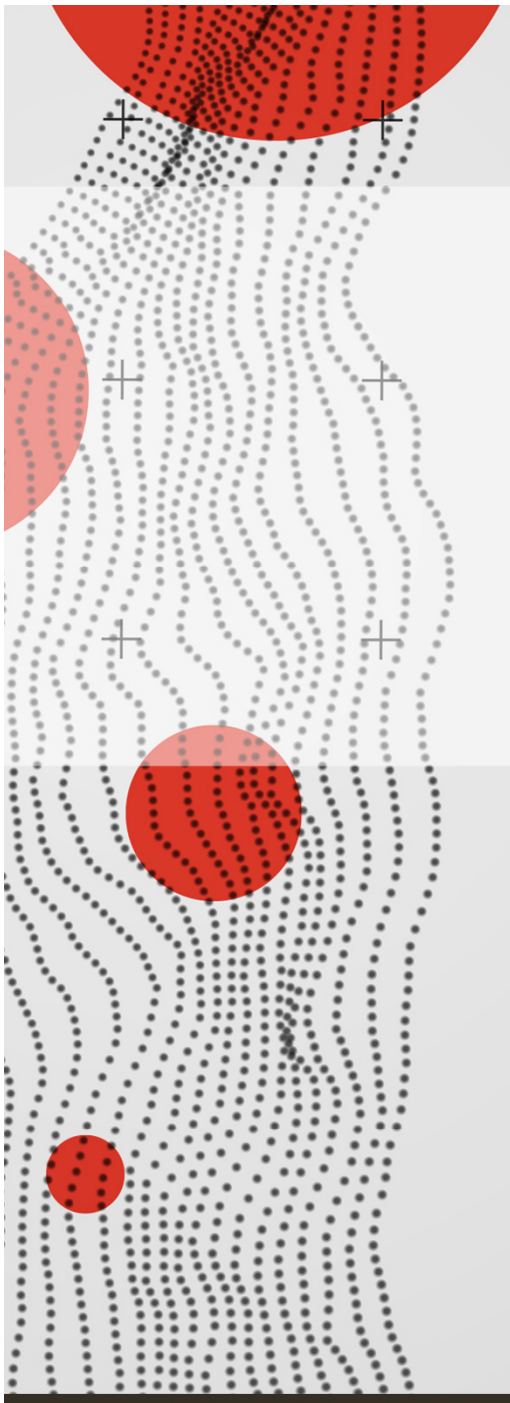
- Logiciel libre Hyphe et ses évolutions pour fonctionner sur le web archivé : https://github.com/medialab/hyphe/issues/372

Annexe 5 :

Programme de la journée d'étude :

« Faire réseau autour des archives du web : Perspectives du projet ResPaDon »

13mars 2023,BnF,Paris



Faire réseau autour des archives du web

Bilan et perspectives du projet ResPaDon

13/03/2023

10:00 - 17:00

Petit auditorium, hall d'accueil (Entrée principale - EST)
Sur inscription

BnF François Mitterrand

Université de Lille

SciencesPo

Gérico

CAMPUS CONDORCET PARIS - AUBERVILLIERS

CollEx-Persée

Faire réseau autour des archives du web : Perspectives du projet ResPaDon

Lundi, 13 mars 2023, BnF.

Présentation

L'équipe de ResPaDon organise une journée d'échange autour des résultats et des perspectives de ce projet débuté en mars 2021. Des conditions d'accès aux collectes collaboratives, des enjeux juridiques à la coopération entre acteurs nationaux et établissements de proximité, le Réseau de Partenaires pour l'exploration et l'analyse de données numériques (ResPaDon) propose de partager les résultats de sa démarche pour développer les usages scientifiques de la collection des archives du dépôt légal du web.

Le programme de la journée propose de découvrir les différents résultats du projet, d'en apprendre davantage sur les expérimentations menées, de comprendre comment ResPaDon envisage son évolution.

RES
PA
DON

PROGRAMME

10:00	Conférence d'ouverture • Alain Colas, directeur du Groupement d'intérêt scientifique CollEx-Persée • Isabelle Nyffenegger, directrice générale adj. et Directrice des Services et des Réseaux de la BnF • Julien Roche, directeur du Service commun de documentation de l'Université de Lille.
10:30	« *Le projet ResPaDon : retour sur l'origine et les coulisses d'un projet d'envergure nationale* » Emmanuelle Bermès (Ecole nationale des Chartes) et Marie-Madeleine Géroudet (Université de Lille, SCD) Cette présentation reviendra sur les origines du projet ResPaDon et sur les coulisses de ce projet original par sa forme et son envergure.
11:00	« *Les archives du web : présentation et enjeux pour la recherche* » Vladimir Tybin (BnF, service du Dépôt légal numérique) et Dorothée Benhamou-Suesser (BnF, DLN). Quelles sont les particularités des archives du web conservées à la BnF ? Quels défis techniques, méthodologiques et épistémologiques pose leur exploitation ?
11:15	Pause
11:30	<u>Session 1. La fabrique des archives du web</u> Petit Auditorium animée par les équipes en charge du dépôt légal du web à la BnF.
	<u>Session 2. Explorer les archives du web</u> Salle 70 animée par Audrey Baneyx (Sciences Po, médialab) et Eleonora Moiraghi (Sciences Po, Direction des Ressources et de l'Information Scientifique).
12:15	« *Le vivant et les archives : l'expérience du DataSprint ResPaDon* » Audrey Baneyx (Sciences Po, médialab), Eleonora Moiraghi (Sciences Po, DRIS) et Fabienne Greffet (Université de Lorraine, IRENEE). Cette présentation propose de revenir sur le DataSprint ResPaDon et de partager les enseignements tirés de cette expérimentation notamment d'un point de vue méthodologique.
12:45	Déjeuner-buffet offert au Foyer du Petit auditorium

RES
PA
DON

14:00 « *Embarquez dans la capsule : retour sur l'expérimentation d'un accès distant* »

Sara Aubry (BnF, département des Systèmes d'information) et Marie Cros (Université de Lille, SCD).

Que contient la capsule d'accès aux archives du web ? Comment l'installer et la faire vivre ? Quel dispositif de médiation mettre en place ? Comment la valoriser auprès des chercheurs ? Quels sont les premiers enseignements de l'expérimentation ?

14:30 « *Les préconisations du projet* »

Emmanuelle Bermès (Ecole nationale des chartes) et Marie-Madeleine Géroudet (Université de Lille, SCD).

Ces préconisations sont issues des travaux du cycle d'ateliers du groupe de réflexion qui s'est réuni tous les deux mois depuis le début du projet. Elles ont fait l'objet d'une démarche de co-construction progressive.

15:00 « *ResPaDon : et la suite ?* » (Table-ronde)

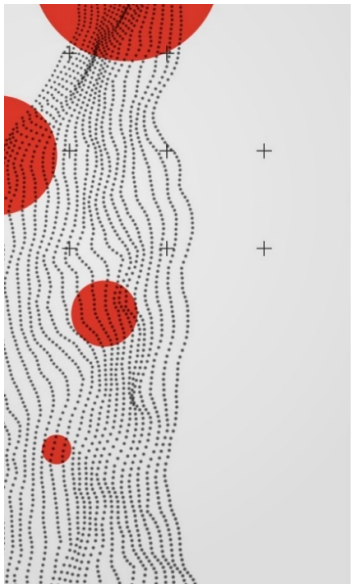
- Sophie Gebeil, maîtresse de conférences, Aix-Marseille Université, TELEMMe
- Stéphanie Groudiev, directrice de l'Humathèque du Campus Condorcet
- Grégory Miura, directeur du SCD de l'Université Bordeaux-Montaigne et Président du TC46 Information Documentation de l'Organisation internationale de normalisation (ISO)
- Claude Mussou, directrice de l'Inathèque
- Benoît Tuleu, directeur du département du dépôt légal de la Direction des Services et des réseaux de la BnF.

animée par Julien Roche, directeur du SCD de l'Université de Lille.

16:15 Conclusion de la journée

« *Continuer à faire réseau : le regard d'une chercheuse sur le projet Respadon* »

Valérie Schafer, professeure à l'Université du Luxembourg (C²DH).



Le projet ResPaDon (Réseau de Partenaires pour l'analyse et l'exploration de données numériques) vise à développer et à diversifier les usages par les chercheurs des archives du web collectées et conservées par la Bibliothèque nationale de France. Soutenu par le GIS CollEx-Persée, le projet Respadon est porté par l'Université de Lille et la Bibliothèque nationale de France, en partenariat avec Sciences Po Paris et le Campus Condorcet. Il mobilise les équipes de recherche du médialab et le laboratoire GERiiCO.

https://respadon.hypotheses.org
respadon@groupes.renater.fr
@Respadon_Projet

RES
PA
DON

Annexe 6 :

Programme du colloque international :

« Le web : source et archive »

3, 4, 5 avril 2023, Université de Lille

RES
PA
DON

PROGRAMME

Colloque international | 3 - 5 avril 2023

Le web : source et archive

Le web : source et archive
Le web : source et archive
Le web : source et archive

LILLIAD | Université de Lille | Villeneuve d'Ascq

Bienvenue

Le web : source et archive

Ce colloque international propose d'interroger la place des sources issues du web dans la recherche et de situer les pratiques d'archivage du web dans des démarches et des questionnements scientifiques pluriels.

Site web du colloque : respadon.sciencesconf.org
Site web du projet ResPaDon : respadon.hypotheses.org

Découvrez la collection des archives de l'Internet de la BnF

Les archives du web sont des collections soumises au droit d'auteur, dont la consultation est strictement encadrée par le Code du Patrimoine et s'effectue exclusivement dans les salles de recherche de la BnF et de ses partenaires en région.

Les expérimentations conduites dans le cadre du projet ResPaDon, labellisé par le dispositif national Collections d'excellence CollEx-Persée, co-porté par la BnF et l'Université de Lille, en partenariat avec Sciences Po et le campus Condorcet, offrent cette année aux chercheurs de l'Université de Lille volontaires l'opportunité d'**explorer l'ensemble des archives du web**, et de découvrir plus particulièrement la **collection « élections 2002 »** depuis des postes dédiés situés dans les bibliothèques de l'Université de Lille.

Le dispositif vise à tester des outils d'exploration et de fouille de données, ainsi que des services d'accompagnement à l'exploitation de ces collections, dans l'objectif d'accroître et diversifier l'usage des archives du web par les chercheurs.

Deux sessions de découverte de la collection des archives de l'Internet de la BnF sont proposées en Salon Pi (LILLIAD) :

4 avril
13h - 14h

5 avril
13h - 14h

Lundi 3 avril 2023

Relations entre dispositif technique et données scientifiques : l'archive web en réseau

Relations between technical devices and scientific data: the networked web archive

APRÈS-MIDI

14:00 Discours d'ouverture *Opening speeches*

Olivier Colot, Vice-Président Recherche, Pr., *Université de Lille*
Isabelle Nyffenegger, Dir. Adj. et Dir. des Services et des Réseaux, *BnF*
Laurence Favier, *GERiiCO, Université de Lille*
Emmanuelle Bermès, *École nationale des chartes*
Marie-Madeleine Géroudet, *SCD, Université de Lille*

14:30 Interpreting the web : the critical role of historical context in web archival research

Ian Milligan, *University of Waterloo*

15:15 Méthodologie pour l'élaboration d'un corpus et d'une archive du web littéraire *Methodology for the elaboration of a corpus and an archive of the French-speaking literary web*

Christian Cote, *MARGE, Univeristé Jean Moulin, Lyon III*

15:45 Table ronde : Les sources de l'étude du web militant
Roundtable: The sources of the study of the activist web

Léna Bouillard, *Enssib*
Alicja Jaworska-Kaska, *Centre de civilisation française et d'études francophones, Université de Varsovie*
Animation : **Eleonora Moiraghi**, *Sciences Po, DRIS*

16:30 Pause *Break*

16:45 Cartographie de la critique en ligne dans les arts du spectacle : entre approche synchronique et diachronique *Mapping online criticism in the performing arts: between synchronic and diachronic approaches*

Cristina Tosetto, *CLARE, Université Bordeaux Montaigne*
Introduction : **Eleonora Moiraghi,** *Sciences Po, DRIS*

17:15
17:45 Du corpus télévisuel au corpus web à l'aide de l'outil visuel automatique : méthodes du projet CROBORA
From the television corpus to the web corpus using the automatic visual tool: methods of the CROBORA project

Shiming Shen, *SIC.Lab Méditerranée, Université Côte d'Azur*
Eric Kergosien, *Université de Lille, GERiiCO*
Matteo Treleani, *SIC.Lab Méditerranée, Université Côte d'Azur*

Mardi 4 avril 2023

Le web à l'intersection de la mémoire et du savoir : enjeux épistémologiques

The web at the intersection of memory and knowledge: epistemological issues

MATIN

09:20 Introduction *(Accueil dès 9h)*

Marie-Madeleine Géroudet, *SCD, Université de Lille*

09:30 Le goût de l'archive numérique et les archives du web

The taste of the digital archive and the web archives

Caroline Muller, *EA Tempora, Université Rennes 2*
Frédéric Clavert, *C²DH, Université du Luxembourg*

10:15 Des émissions culinaires aux vidéos conseils : transformation des recettes audiovisuelles de la télévision au web et la question du faire à manger *digital*

From cooking shows to advice videos: transformation of audiovisual recipes from the television to the web and the question of « digital » cooking

Christian Bonah, *SAGE, Université de Strasbourg*
Solène Lellinger, *SPHERE, Université Paris Cité*
Caroline Sala, *Université de Strasbourg*

10:45 Web vivant et web archivé, aux sources de l'histoire nativement numérique

Live web and archived web, at the sources of natively digital history

Sophie Gebeil, *TELEMMe, MMSH, INSPE, Aix Marseille Université*

11:15 Pause *Break*

11:30 Étudier les usages des archives du web
12:30 *Studying the uses of web archives*

Laurence Favier, *GERiiCO, Université de Lille*
Antoine Henry, *GERiiCO, Université de Lille*
Irène Bastard, *BnF, Délégation à la stratégie et à la recherche*
Alexandre Faye, *BnF, service du Dépôt légal numérique*

APRÈS-MIDI

13:00 Découverte de la collection des archives de
14:00 l'Internet de la BnF - Salon Pi

14:00 Table ronde : Recherches et méthodes mobilisant les archives du web *Research and methods using web archives*

Clément Bert-Erboul, *ULB RESI*
Grégoire Clémencin, *Autonoia*
Amira Dahmani, *LARIME, ESSECT, ISGB, Université de Carthage*
Brice Demars, *CY Cergy Paris Université*
Jean Finez, *Université Grenoble Alpes, PACTE*
Amélie Macaud, *Climas, Université Bordeaux Montaigne*

Animation : **Marie-Madeleine Géroudet**, *SCD, Université de Lille*

15:15 Exploration des archives du Web par un public étudiant : contribution à l'analyse critique des sources

Exploration of web archives by students: contribution to the sources critical analysis

Joana Casenave, *GERiiCO, Université de Lille*
Laurence Favier, *GERiiCO, Université de Lille*

15:45 Pause *Break*

16:00 Table ronde : Enjeux épistémologiques et didactiques
17:00 des sources web

Epistemological and didactic issues of web sources

Sophie Gebeil, *TELEMMe, MMSH, INSPE, Aix Marseille Université*
Grégory Miura, *SCD, Université Bordeaux Montaigne*

Discutante : **Emmanuelle Bermès**, *École nationale des chartes*

Mercredi 5 avril 2023

Politiques, pratiques et techniques archivistiques et archives web : du document aux corpus

Archival policies, practices, techniques and web archives: from document to corpus

MATIN

09:20 **Introduction** (Accueil dès 9h)

Laurence Favier, *GERiiCO, Université de Lille*

09:30 **Entre inscription éphémère et donnée pérenne : peut-on archiver le Web au-delà de son enregistrement ? Quelques remarques méthodologiques et critiques**

Between ephemeral inscription and perennial data: can we archive the web beyond its recording ? Some methodological and critical remarks

Bruno Bachimont, *Costech, Université de technologie de Compiègne*

10:15 **Explorer les archives de l'internet à l'Université de Lille : regards croisés sur un dispositif expérimental au service des chercheurs**

Exploring the internet archives at the University of Lille: transversal perspectives on an experimental system for researchers

Dorothée Benhamou-Suesser, *BnF,* Service du dépôt légal numérique
Marie Cros, *SCD, Université de Lille*
Gwladys Hadjimanolis, *Clersé, Université de Lille*

10:45 **Pause** *Break*

11:00 **Table ronde : Question(s) de droit(s)**
12:30 *Roundtable: Legal issues*

Lucien Castex, *chercheur, Université Sorbonne Nouvelle-Paris 3, co-responsable du groupe de recherche Gouvernance et régulation d'Internet, CIS GDR CNRS, membre de la Commission nationale consultative des droits de l'Homme (CNCDH)*
Nathalie Mallet-Poujol, *directrice de Recherche au CNRS, CEPEL (UMR 5112), Université de Montpellier*

Animation : **Marie-Madeleine Géroudet**, *SCD, Université de Lille* et **Arnaud Laborderie**, *BnF, Service de la coopération numérique et de Gallica*

13:00
14:00 **Découverte de la collection des archives de l'Internet de la BnF** - Salon Pi

APRÈS-MIDI

14:00 **Archiver le web littéraire. Défis méthodologiques et conceptuels**

Archiving the literary web. Methodological and conceptual challenges

Servanne Monjour, *CELLF, Sorbonne Université*
Nicolas Sauret, *Université Paris 8, Paragraphe*

14:30 **Collectes du confinement et archives du web : exploration croisée des archives de BNF et de l'INA**

Collection of the lockdown and web archives: cross-exploration of BnF and INA archives

Louis Gabrysiak, *Labex Les Passés dans le Présent, Université Paris Nanterre*
Sarah Gensburger, *Labex Les Passés dans le Présent, Université Paris Nanterre*
Marta Severo, *Labex Les Passés dans le Présent, Université Paris Nanterre*

14:45 **Harlem Shake à la BnF ... À la recherche d'un phénomène viral dans les archives du Web**

Harlem Shake at the BnF... Looking for a viral trend in the web archives

Alexandre Faye, *BnF, service du Dépôt légal numérique*
Fred Pailler, *C²DH, Université du Luxembourg*
Sara Aubry, *BnF DSI*
Antoine Silvestre de Sacy, *Huma-Num*
Valérie Schafer, *BnF, C²DH, Université du Luxembourg*

15:15 **Échanges et questions** *Discussion and questions*

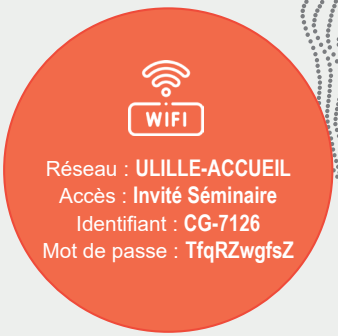
15:25 **Le cycle d'ateliers ResPaDon : bilan et préconisations**

The ResPaDon cycle of workshops: results and recommendations

Emmanuelle Bermès, *École nationale des chartes*
Marie-Madeleine Géroudet, *SCD, Université de Lille*

16:10
16:30 **Discours de clôture** *Closing remarks*

RES
PA
DON



Un colloque organisé par :

Université de Lille

BnF Bibliothèque nationale de France

CollEx-Persée

SciencesPo

CAMPUS CONDORCET
PARIS - AUBERVILLIERS

Gériico