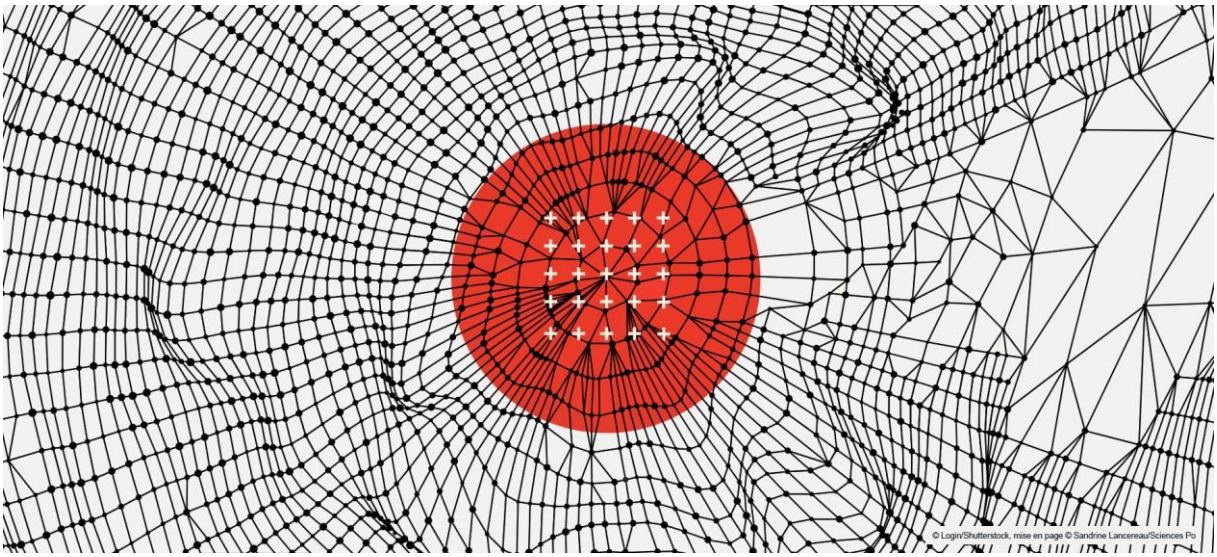
L'expérimentation capsule au cœur du projet ResPaDon : bilan et recommandations

Sara Aubry, Dorothée Benhamou-Suesser, Marie Cros

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 :

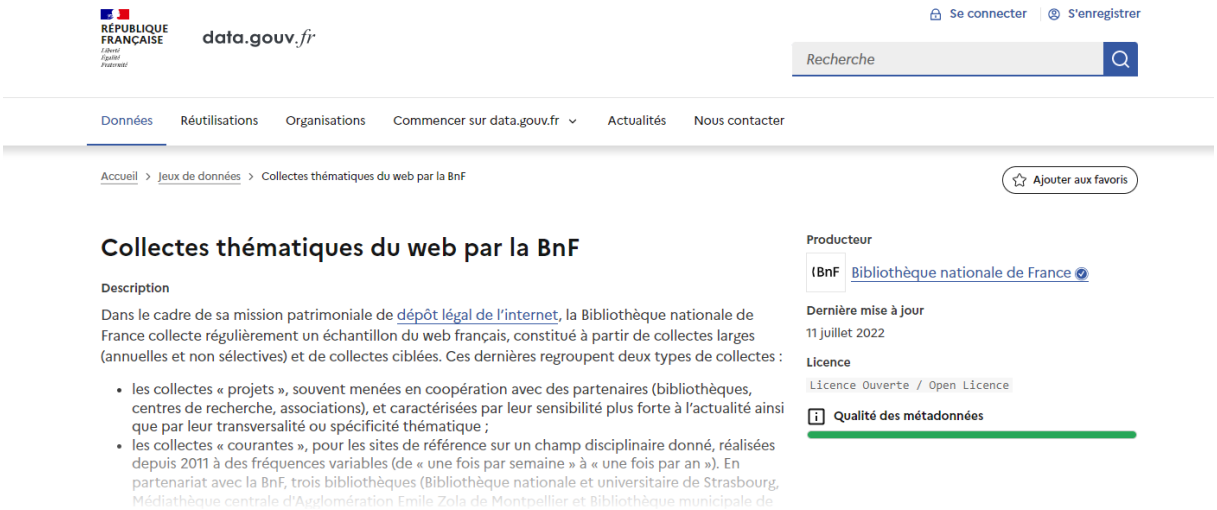
- certaines de ces données transformées ou enrichies sont entièrement la propriété du chercheur qui les a produites, on peut penser notamment au graphe de liens

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.



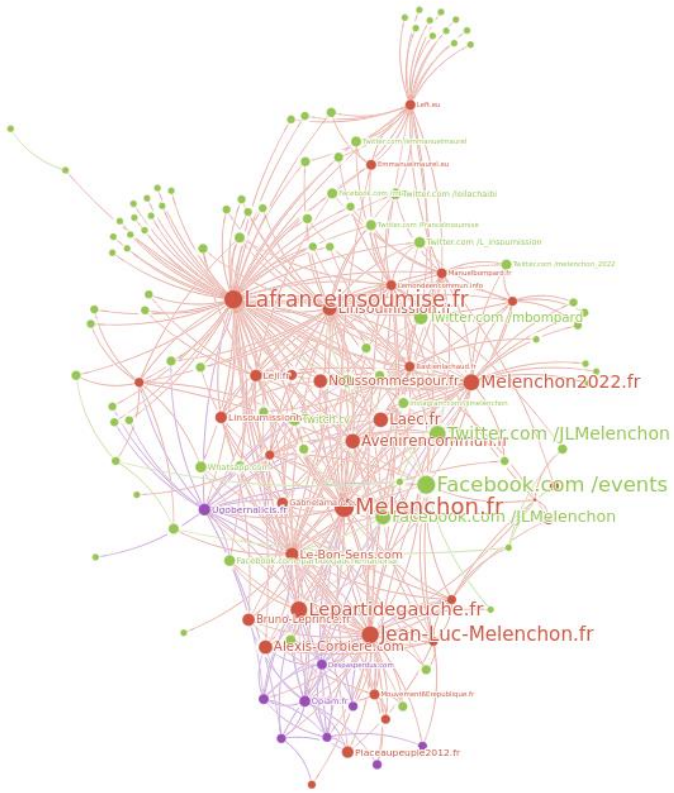
Exemple de métadonnées documentaires diffusées sous licence EtaLab : liste des sélections faites dans le cadre des collectes ciblées par la BnF

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

```

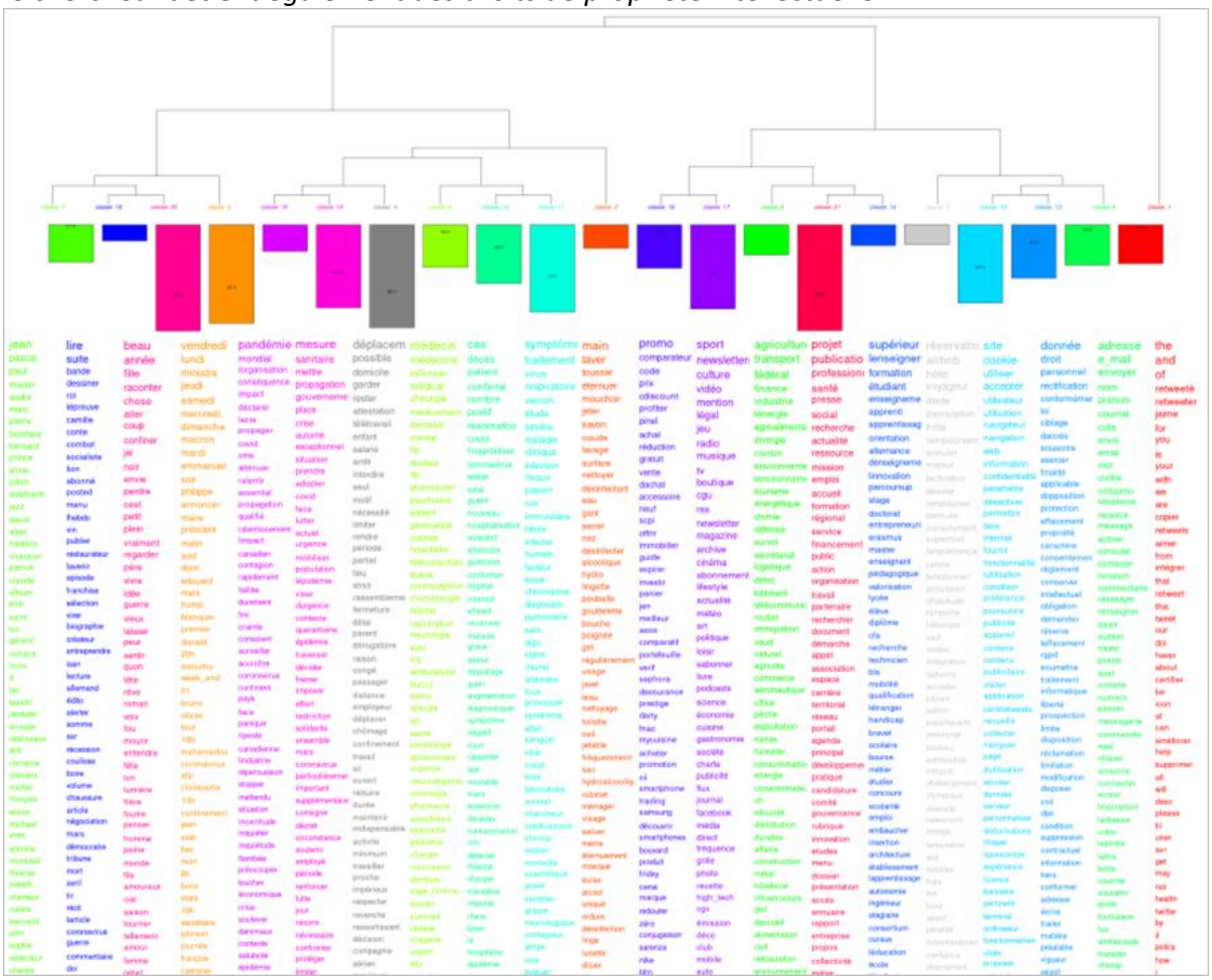
Fichier texte contenant les liens entrants et sortants de chaque site pouvant servir à produire un graphe de liens dans un logiciel dédié, par exemple Gephi, disponible dans la capsule. Donnée de la recherche / corpus intermédiaire, libre de droit, reproductible dans une publication



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

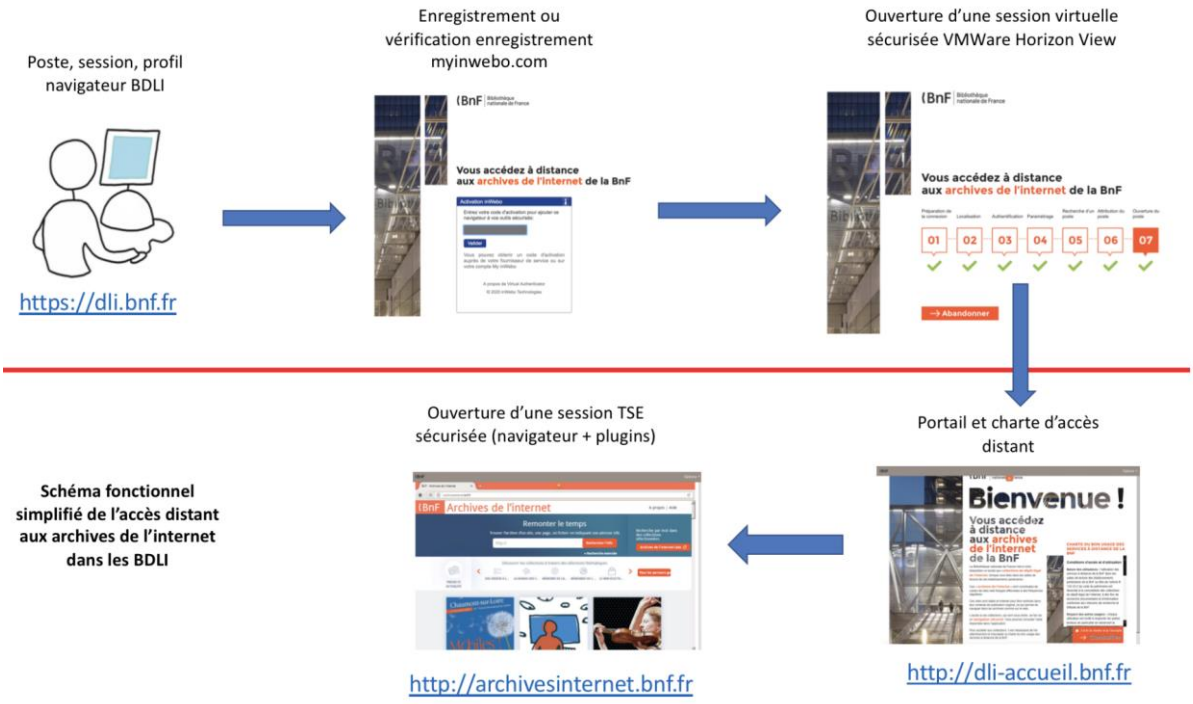


Schéma fonctionnel simplifié de l'accès distant aux archives de l'internet avec la solution inWebo

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

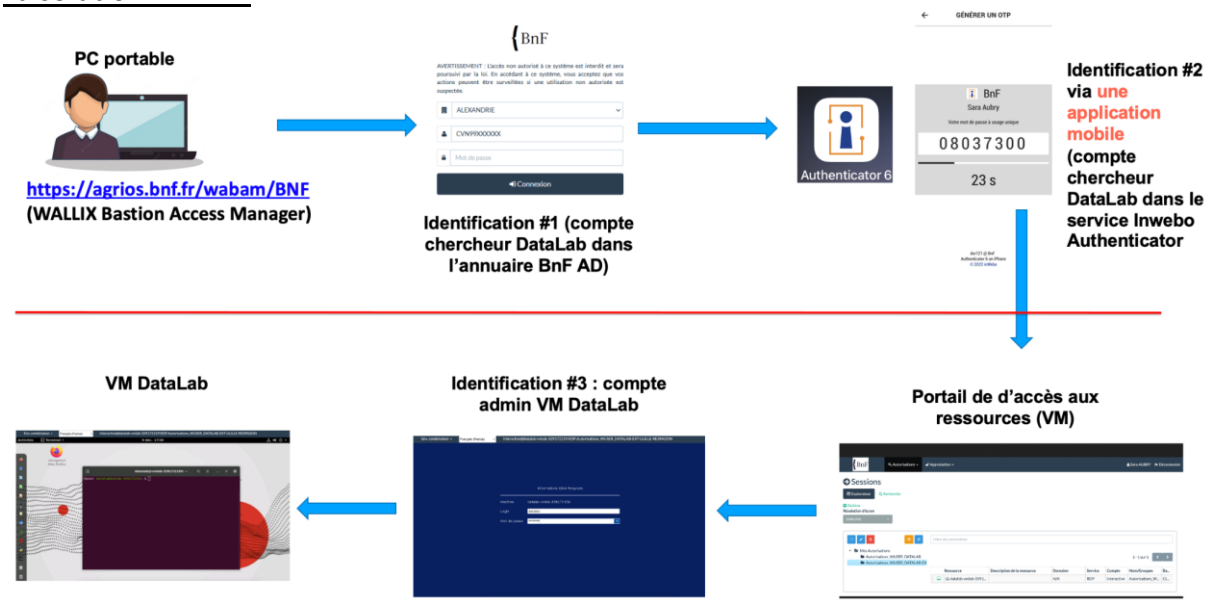
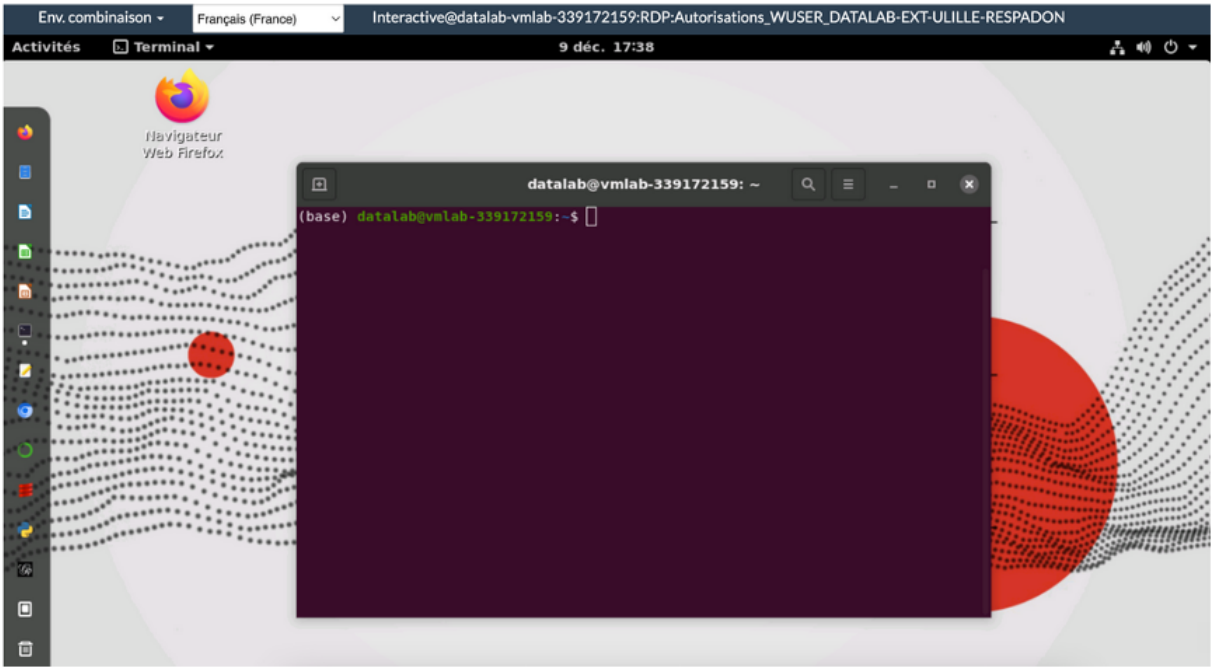


Schéma fonctionnel simplifié de l'accès distant à la capsule d'exploration des archives élections 2002 avec la solution WALLIX

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 :

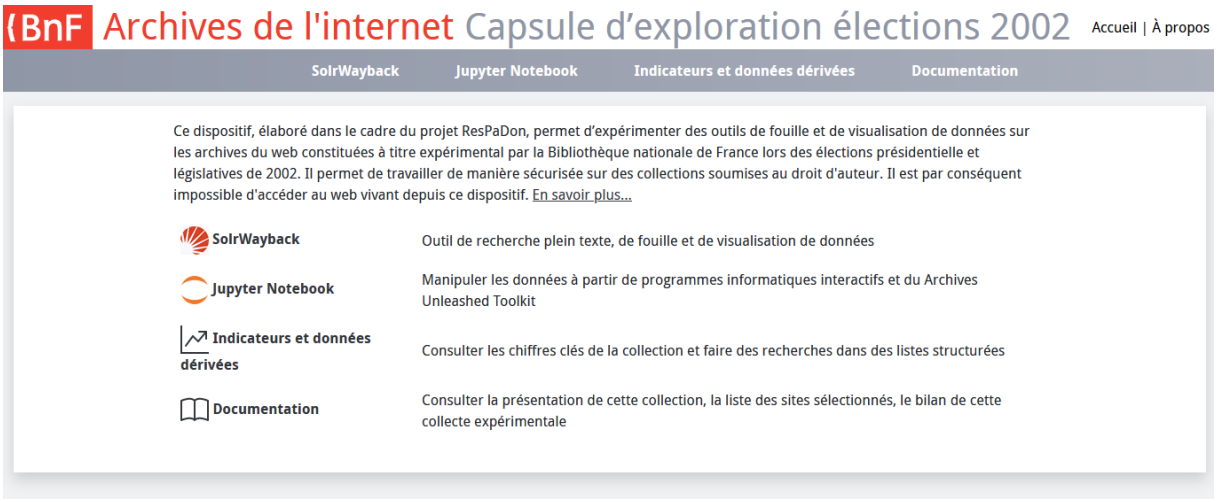
- l'application Archives de l'internet permettant de consulter l'ensemble des collections de dépôt légal du web (1996 à aujourd'hui, représentant 1,8 Po de données) à partir d'une recherche par URL, ou au travers de parcours guidés, tels que « Cliquer, voter : l'internet électoral » portant sur les élections 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 :

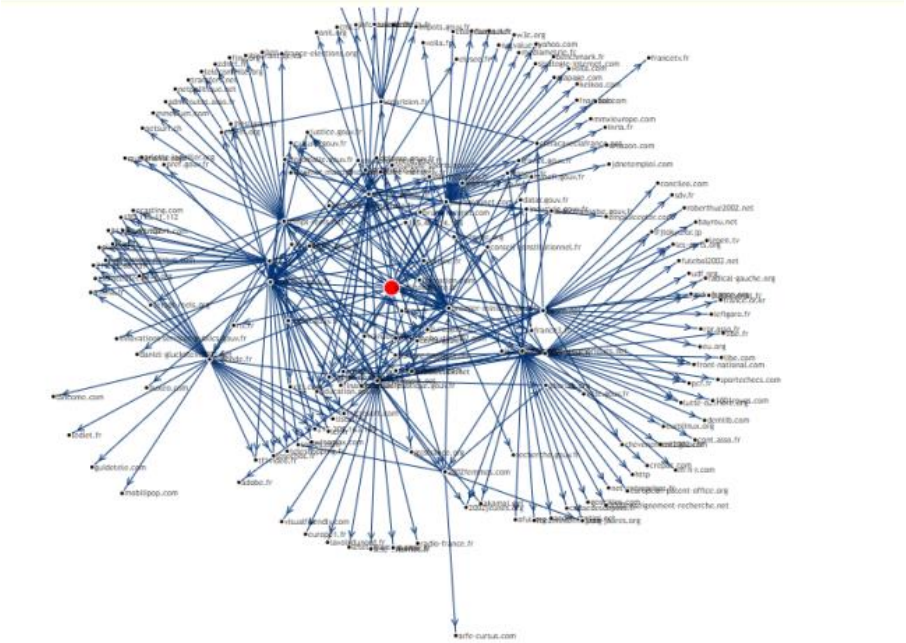
- SolrWayback (https://github.com/netarchivesuite/solrwayback) est un outil open source développé par la Bibliothèque Royale du Danemark et au développement duquel la BnF contribue depuis 2022 qui propose des fonctionnalités d’exploration : recherche par adresse URL, recherche plein texte, navigation 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

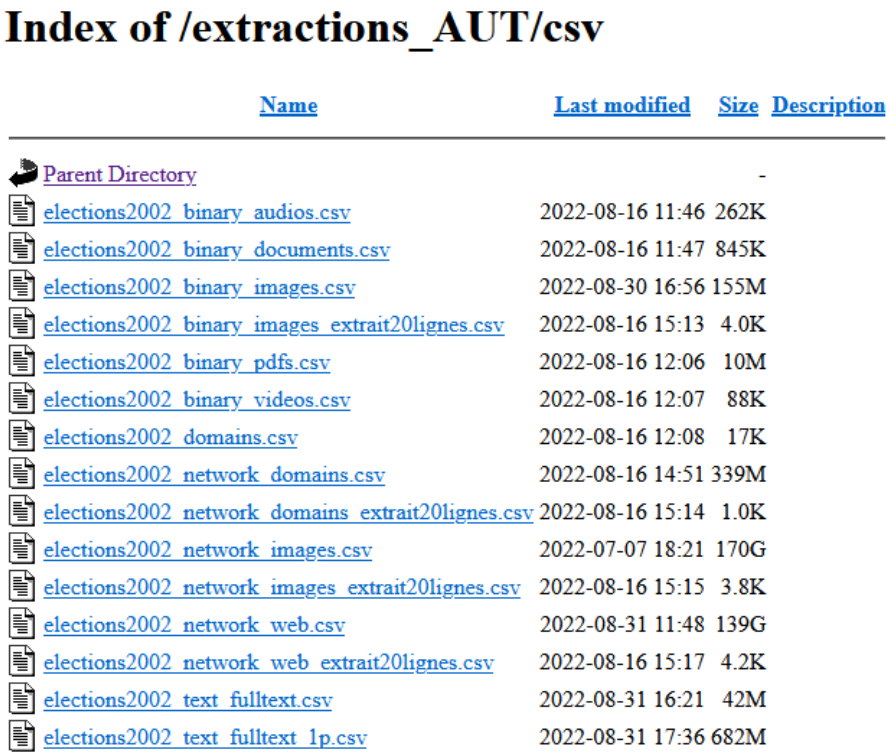
	2002-01	2002-02	2002-03	2002-04	2002-05	2002-06	2002-07	2002-08
Size in kilobytes	0	53,172	2,136,652	7,768,931	8,736,618	6,426,155	0	0
Pages	0	2,692	157,467	579,287	477,825	327,113	0	0
Ingoing links	1	6	43	60	58	72	7	0

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

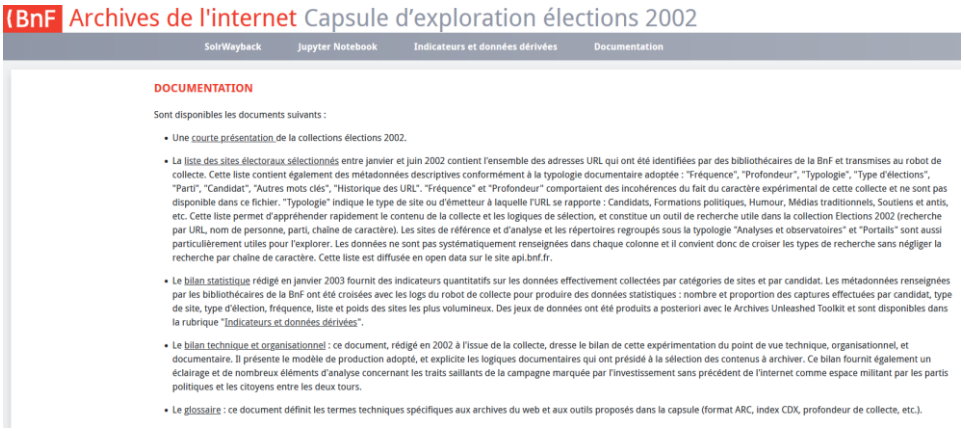
Max. node degree correspond au nombre maximum de liens entrants (ingoing) ou sortants (outgoing) d'un noeud

- 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…

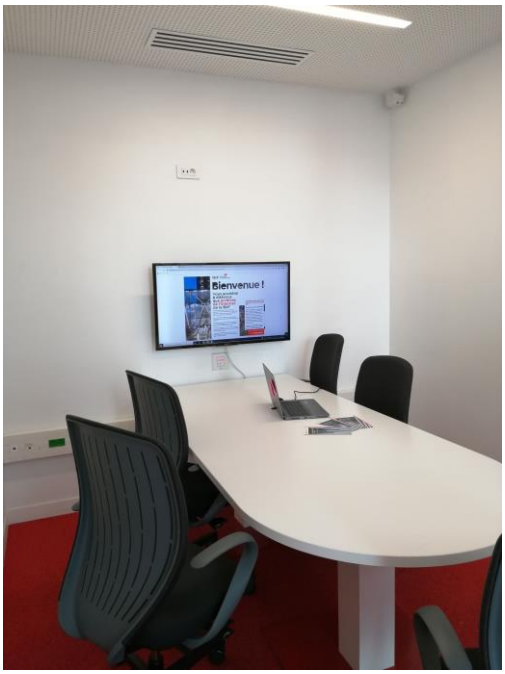
Les salles

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



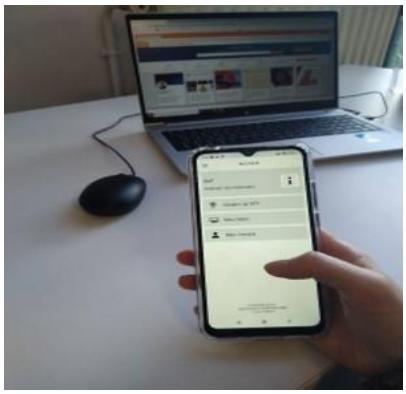
Lilliad Salle 1S01

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…".

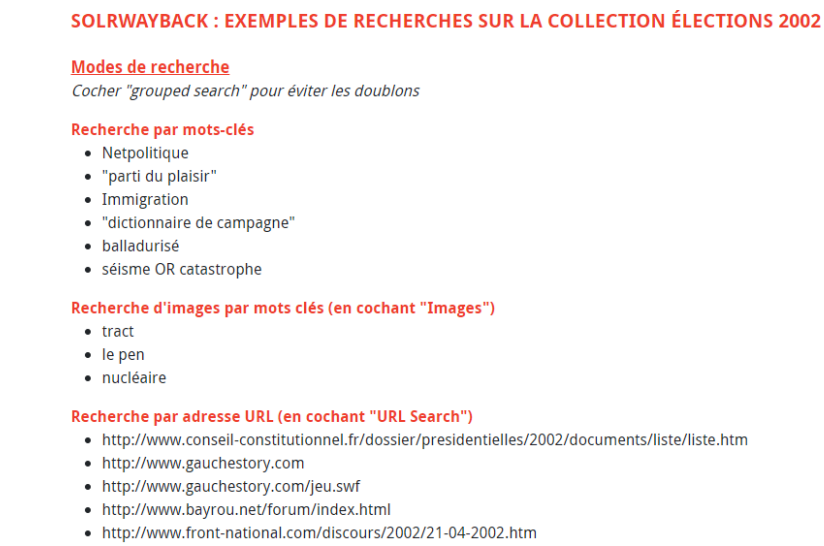
- **Créer un document qui récapitule toutes les étapes de l'accompagnement** pour que les médiateurs puissent le consulter si besoin avant chaque nouvelle séance (en particulier si les sessions ne sont pas régulières dans le temps). Le document créé à l'université de Lille peut être repris et adapté.
- **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

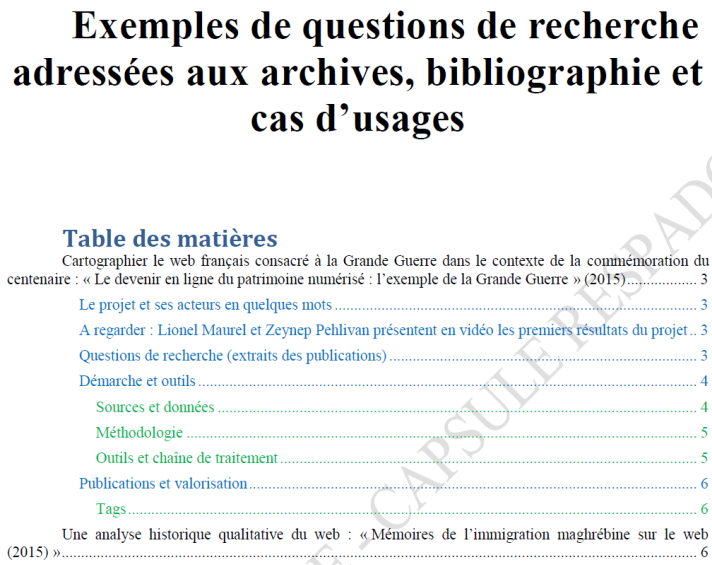
Recherche d'images par mots clés (en cochant "Images")

- 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.

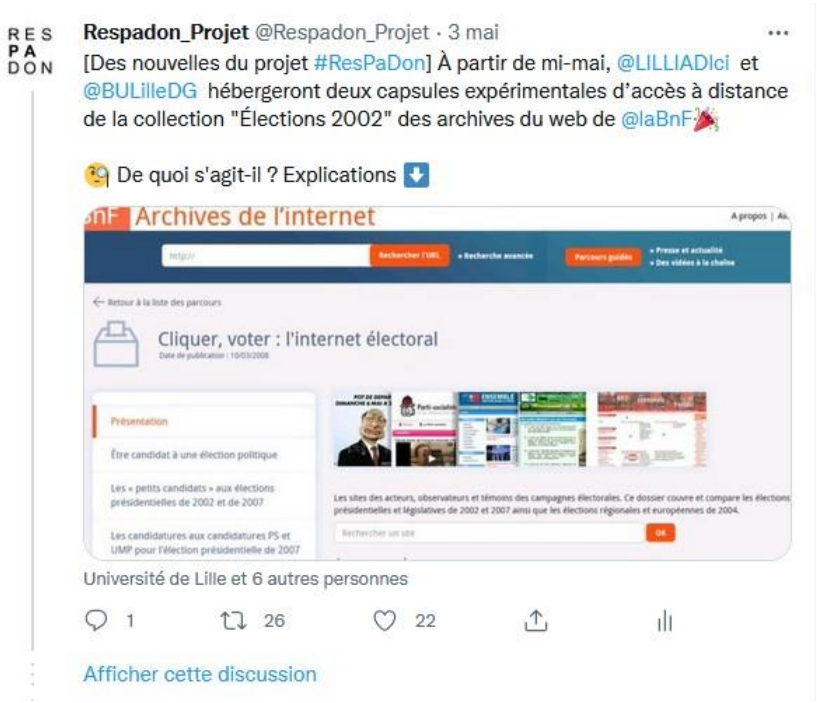
Toute la communication visuelle s'est appuyée sur la charte graphique du projet ResPaDon ce qui a apporté une cohérence aux affiches, kakemono, vignettes et visuels développés.

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.