Le projet Corpus et ses publics potentiels.

Eleonora Moiraghi

Le projet Corpus et ses publics potentiels.

UNE ÉTUDE PROSPECTIVE SUR LES BESOINS ET LES ATTENTES DES FUTURS USAGERS.

Eleonora Moiraghi, assistante de recherche
sur le programme de recherche Corpus porté
par la Bibliothèque nationale de France

Janvier 2018

{BnF

La Bibliothèque nationale de France a parmi ses objectifs de mettre en place un nouveau service de fourniture de données à destination de la recherche. Ce service s'appelle provisoirement Laboratoire d'étude et d'analyse de corpus numériques et fait l'objet du programme de recherche Corpus. La présente étude, en essayant de cerner les besoins des usagers potentiels, vise à fournir matière à réflexion pour contribuer à la conception du futur Laboratoire.

{BnF

{BnF

RÉSUMÉ EXÉCUTIF

RÉSUMÉ DE L'ÉTUDE ET DES PRINCIPAUX RÉSULTATS

Initié en 2016 dans le cadre du plan quadriennal de la recherche 2016-2019 de la Bibliothèque nationale de France (BnF), le **projet Corpus** a pour objectif d'adapter l'offre de services de la Bibliothèque aux nouveaux besoins de la communauté académique en construisant un **Laboratoire d'étude et d'analyse de corpus numériques**.

Situé dans les emprises de la Bibliothèque, à l'heure actuelle condition *sine qua non* de la mise à disposition de corpus sous droit, ce Laboratoire doit permettre à un public universitaire l'exploration, l'analyse et le traitement de corpus numériques issus des collections numérisées ou nativement numériques conservées à la BnF.

La BnF n'est pas la seule à avoir eu l'idée d'un espace permettant l'appropriation de technologies numériques et la fouille de données de type TDM (*Text and Data Mining*). Le tableau ci-dessous résume de manière partielle et synthétique un **contexte** encore mouvant dans lequel la BnF devra se positionner.

Le programme de recherche Corpus

2016	Pilote : Dépôt Légal du Web Manifestation (ateliers, journée de conférences)
2017	Pilote : Département de la Conservation Étude (ateliers, enquête qualitative)
2018	Pilote : Département des Métadonnées Plate-forme …
2019	**Bilan final**

Une orientation partielle et synthétique dans le contexte international

BIBLIOTHÈQUES NATIONALES	BIBLIOTHÈQUES UNIVERSITAIRES	CENTRES CONSTRUITS AUTOUR DE LA SCIENCE DES DONNÉES OU DES HUMANITÉS NUMÉRIQUES	LABORATOIRES PUREMENT VIRTUELS FONDÉS PAR PLUSIEURS UNIVERSITÉS ; INFRASTRUCTURES ET PROJETS DE RECHERCHE
KB LAB espace physique à l'accès restreint et virtuel	OXFORD Centre for Digital Scholarship	ALAN TURING INSTITUTE	HATHITRUST
LIBRARY OF CONGRESS LABS espace virtuel et physique à l'accès restreint	LEYDE Centre for Digital Scholarship	DATA INSTITUTE UNIVERSITÉ GRENOBLE ALPES	ALVEO
BRITISH LIBRARY LABS manifestation …	PITTSBURGH Digital Scholarship Commons	EPFL/ DHLAB	HUMA-NUM
	…	MEDIALAB	ISTEX
		STANFORD/ CESTA …	…
Cible : public universitaire d'origine variée et autres publics intéressés	**Cible : public universitaire « local »**	**Cible : public universitaire et expert**	**Cible : public universitaire d'origine variée**

Dans ce paysage changeant et inexploré, l'**étude** qui fait l'objet du présent rapport a été menée par Eleonora Moiraghi, assistante de recherche recrutée sur le projet Corpus entre août et décembre 2017 afin de cerner les besoins des publics potentiels du Laboratoire et de poursuivre la préfiguration du nouveau service.

Cette étude, intitulée « Le projet Corpus et ses publics potentiels », a été structurée en trois volets :

- une **enquête qualitative par entretiens** auprès de trois populations
  (seize universitaires, onze agents de la BnF et trois experts du domaine)
  pour cerner les besoins et mesurer les attentes
  en adoptant une méthodologie qualitative, compréhensive et inductive ;

- **deux ateliers** ou séances d'échange et de débat
  le 16 octobre « Décrire, transcrire et diffuser un corpus documentaire hétérogène : méthodes, formats, outils »[1]
  le 30 novembre « Géolocalisation et spatialisation de documents patrimoniaux »[2]
  pour inciter les échanges entre le monde des bibliothèques et le milieu de la recherche
  et effectuer des observations informelles ;

- un **atelier collaboratif utilisant la méthode des *personas***
  le 8 décembre, vingt-et-un participants (six universitaires et quinze agents de la BnF)
  pour vérifier les besoins et les attentes manifestées lors des entretiens individuels.

Les principaux **résultats** de l'étude sont les suivants :

1. La fouille de données de type TDM contribue à l'accroissement de la connaissance ;
   elle ne constituera pas l'approche exclusive de la recherche mais elle est appelée à se développer dans toutes les disciplines, même si le rythme et l'ampleur de cette évolution sont difficiles à déterminer dans le cadre de cette étude.(*)

2. L'utilisation ou la fourniture de corpus numériques impliquent des contraintes techniques, humaines, d'organisation et juridiques.

   Les contraintes techniques qui pèsent sur les projets sont dues souvent aux formats, à l'hétérogénéité des données, aux technologies et aux erreurs d'encodage.
   Les solutions envisagées sont un effort de transparence, d'explicitation des biais de la part des fournisseurs de données et d'accompagnement.

   Les contraintes humaines sont liées souvent à l'interdisciplinarité, difficile à gérer en termes de compréhension, collaboration et évaluation.
   Cependant l'interdisciplinarité est également source d'enrichissement mutuel.

   Les aspects d'organisation contraignent les équipes des départements de la Bibliothèque qui, sollicitées par des chercheurs, doivent parfois les accueillir dans leurs bureaux.
   La pénurie de locaux dans les universités parisiennes entraîne aussi des difficultés d'ordre logistique. Le Laboratoire à la BnF pourrait constituer une solution aux deux difficultés.

   Les aspects juridiques sont contraignants aussi bien pour le monde des bibliothèques que pour le milieu universitaire. Dans ce contexte, la Bibliothèque doit d'une part sécuriser les accès à travers l'établissement de conventions et la mise en place de mesures techniques, et d'autre part elle peut apporter sa connaissance du cadre juridique et satisfaire le besoin de formation des universitaires.

---

1 Compte-rendu : Eleonora Moiraghi, **Décrire, transcrire et diffuser un corpus documentaire hétérogène : méthodes, formats, outils**, Carnet de la recherche à la Bibliothèque nationale de France, Url : <https://bnf.hypotheses.org/2214>

2 Compte-rendu : Eleonora Moiraghi, **Géolocalisation et spatialisation de documents patrimoniaux : trois heures de partage autour de la cartographie numérique**, Carnet de la recherche à la Bibliothèque nationale de France, Url : <http://bnf.hypotheses.org/2299>

(*) Rappelons que les publics interrogés sont « acquis à la cause »

3. Le panorama de services similaires en France ou à l'étranger est difficile à saisir aussi bien pour les universitaires que pour les agents de la BnF.

4. La Bibliothèque doit trouver son rôle à jouer dans l'écosystème de la recherche. Outre le fait de remplir sa mission de fournisseur de données, la BnF pourrait beaucoup apporter au monde de la recherche en :
- offrant formation et accompagnement autour de son savoir-faire, autour de l'histoire de ses données et des politiques documentaires de l'Établissement ;
- animant une communauté interdisciplinaire qui n'a pas d'existence naturelle dans le milieu académique ;
- orientant les utilisateurs vers des réseaux académiques et des formations autour des outils et des techniques auxquelles elle ne doit pas se substituer ;
- valorisant éventuellement les outils existants et les résultats de la recherche.

5. Le Laboratoire, c'est-à-dire l'espace dans les emprises de la Bibliothèque, doit être facile d'accès, convivial, doté d'un réseau Wi-Fi rapide, compartimenté et capable d'évoluer au rythme de l'innovation et du progrès technologique. Il aurait intérêt à être utile aussi bien aux chercheurs qu'aux agents de la BnF.
Intéressant pour le partage d'expertise entre le milieu de la recherche et la Bibliothèque, pour l'aspect communautaire, la mise à disposition de locaux et la proximité des collections — mais sur fond d'une difficulté à se déplacer déjà vérifiée dans d'autres enquêtes (Les Archives de l'Internet, Public professionnel, 2016) — il ne doit pas être privé d'une dimension virtuelle qui garantit en mobilité le contrôle du flux de travail et donc la continuité du travail de recherche.

6. Compartimenté en fonction des activités, le Laboratoire devrait comprendre idéalement à terme :
- des espaces pour le personnel de la BnF ;
- des bureaux pour les agents de la BnF, des chercheurs en résidence et des enseignants ;
- une salle pour des événements, restitutions collectives ;
- un espace de restauration et détente ;
- des loges pour le travail individuel ;
- des salles pour le travail en groupe et des formations ;
- un espace de présentation/ exposition (résultats de la recherche, partage d'expériences).

7. Garante du contrôle du flux de travail en mobilité (de plus en plus importante pour les universitaires), l'infrastructure numérique devrait pouvoir être accessible idéalement depuis les ordinateurs portables.
Une infrastructure sécurisée de type *cloud* répondrait aux besoins des utilisateurs en fournissant essentiellement l'accès aux corpus numériques, à une palette d'outils, à des tutoriels et à des exemples d'usages ainsi qu'à plusieurs fonctionnalités comme la possibilité d'effectuer une demande en ligne de numérisation de corpus, d'océrisation ou de collecte de contenus web, de réservation d'une place dans le Laboratoire, de prise de rendez-vous avec des membres du personnel de la BnF.

8. Fournir les données ne dispense pas d'un accompagnement autour de cinq champs principaux : expertise sur les fonds, conseil juridique, soutien informatique et autour de l'analyse des données, orientation. Cet accompagnement pourrait être effectué par des membres du personnel BnF (conservateurs, juristes, ingénieurs, informaticiens, experts des formats des données) à travers du « service public » mais impliquerait également la présence permanente de deux personnes au moins : une spécialisée dans la science des données et ayant une formation en sciences humaines et sociales, et l'autre ayant une formation également hybride capable d'orienter les usagers, d'assurer la communication, le bon fonctionnement technique du Laboratoire et l'organisation d'événements.

Dans un univers aussi expérimental et indéfini, il est raisonnable de construire un service orienté usager et donc de tenir compte des exigences qui ressortent de la présente étude mais cela pourrait rentrer dans le rôle de la Bibliothèque aussi d'imaginer, stimuler et guider les usages.

Rappelons que la présente étude focalise son attention sur les besoins provenant du milieu académique

mais le projet d'un tel Laboratoire représente aussi l'opportunité pour la Bibliothèque de penser de façon organique et cohérente ses projets liés au numérique. Une articulation du Laboratoire avec d'autres services ou initiatives numériques comme le portail BnF API et jeux de données, Gallica Studio, le Hackathon ainsi que le projet de création d'une Chaire Bibli-Lab avec Télécom ParisTech serait bénéfique aussi bien pour la Bibliothèque que pour les usagers. De plus, l'ouverture progressive à des résidences créatives constitue une opportunité pour la Bibliothèque de stimuler des usages innovants et de mener aussi des projets collaboratifs autour des collections numérisées ou nativement numériques.

En conclusion, une synthèse de la préfiguration du futur Laboratoire basée sur les résultats de l'enquête est proposée ci-dessous :

LIEU	Dans les emprises de la Bibliothèque En Rez-de-jardin Accessible avec une carte de recherche Compartimenté, modulable et évolutif	Espaces pour les agents de la BnF Bureaux Loges pour le travail individuel Salles de groupe Zone de détente et restauration Salle pour événements Espace de présentation
COLLECTIONS NUMÉRIQUES	Numérisées Nativement numériques Dans le domaine public et sous droit	Documents numérisés (Gallica) Dépôt légal numérique Archives de l'Internet Métadonnées Logs
EXPLORATION, ANALYSE, TRAITEMENT	Fouille de données de type TDM Données brutes ou pré-travaillées	Corpus pré-constitués Extraction de données Constitution de corpus ou sous-corpus Demande de collecte de contenus web Demande de numérisation, d'océrisation
INFRASTRUCTURE NUMÉRIQUE	Machines virtuelles Plate-forme Puissance de calcul	Espace personnel sécurisé Accès sécurisé aux données Tutoriels ou exemples d'usages Extraction de données ou corpus Espace de stockage Demande de collecte ou numérisation Palette des outils plus utilisés Demande d'installation d'outils Notification e-mail de fin de processus FAQs Forum Réservation de place Prise de rendez-vous
PARTENAIRES	(Infra-)structures de la recherche Universités Laboratoires Réseaux Autres bibliothèques	CNRS TGIR Huma-Num TeraLab British Library Sciences Po/ Médialab Labex OBVIL DHLAB/ EPFL MATE-SHS ...

1. L'ÉTUDE ET SON CONTEXTE

LE PROGRAMME DE RECHERCHE CORPUS PORTÉ PAR LA BIBLIOTHÈQUE NATIONALE DE FRANCE : PRÉSENTATION, OBJECTIF, STRUCTURATION

La présente étude a été réalisée dans le cadre du **programme de recherche Corpus** porté par la Bibliothèque nationale de France (BnF). Initié en 2016, inscrit dans le plan quadriennal de la recherche[1] et porté par la Direction des services et des réseaux avec l'appui de la Délégation à la stratégie et à la recherche, ce projet d'une durée de quatre ans vise à préfigurer un nouveau service de fourniture de corpus numériques à destination de la recherche au sein de la Bibliothèque.

L'**origine** du programme réside dans le constat d'un besoin émergeant provenant du milieu académique. Durant les dernières années, plusieurs départements au sein de l'Institution ont remarqué une demande renouvelée de constitution et de fourniture de corpus numériques, notamment à des fins de fouille de textes et de données (TDM[2]). La Bibliothèque a su répondre à ces demandes en adoptant une approche *ad hoc,* c'est-à dire plus concrètement en établissant des partenariats avec les laboratoires de recherche demandeurs de corpus.

Comme les pratiques de fouille de données sont destinées à se développer à la Bibliothèque, il paraît nécessaire de mener une réflexion débouchant à terme sur un service capable de répondre de manière plus cadrée et efficace à l'augmentation de ce type de demandes.

Le programme de recherche Corpus s'inscrit dans la mission de la BnF d'« assurer l'accès du plus grand nombre aux collections[3] » en adaptant l'offre de services de la Bibliothèque au contexte actuel, c'est-à-dire aux nouvelles opportunités d'interrogation de masses de données, trop complexes pour que l'œil humain puisse les appréhender sans l'aide de nouvelles méthodes et outils d'analyse numériques.

Le programme de recherche Corpus vise à construire un nouveau service de fourniture de données à destination de la recherche

Afin de concevoir un service de fourniture de données à destination de la recherche, le projet a été structuré en quatre volets correspondant à chaque année du programme, et confiés à un service ou département impliqué dans le projet en raison de ses missions et de ses activités.

En 2016, le premier volet a vu le département du Dépôt légal numérique piloter le projet, travailler avec plusieurs équipes de recherche et organiser des ateliers ainsi qu'une journée de conférences[4]. Au fil de l'année 2017, la deuxième partie du programme a été assurée par le département de la Conservation qui a observé les pratiques de l'équipe du GRIPIC[5] travaillant sur le projet Giranium, et a organisé avec elle deux ateliers. En 2018, ce sera au département des Métadonnées de poursuivre la réflexion et l'expérimentation en travaillant sur un prototype de plate-forme d'outils.

La dernière année est destinée à un bilan final pour aboutir à la mise en place d'un service opérationnel à l'horizon 2020.

1 **Plan quadriennal de la recherche 2016-2019** de la BnF, Url : <http://c.bnf.fr/gPP>

2 Définition rapportée par Catherine Muller dans le billet de blog **#TDM : Fouille de textes et de données dans le contexte de la loi pour une République numérique - Journée d'étude ADBU du 13/12/16** : « La fouille de textes et de données, désignée sous le terme anglo-saxon de TDM pour *Text & Data Mining* désigne toute technique d'analyse automatisée visant à analyser des textes et des données sous forme numérique afin d'en dégager des informations stratégiques pour la recherche telles que des constantes, des tendances et des corrélations », Url : <http://www.enssib.fr/recherche/enssiblab/les-billets-denssiblab/tdm-fouille-de-donnees-istex-ist-text-and-data-mining-loi>

3 **Décret n°94-3 du 3 janvier 1994** portant création de la Bibliothèque nationale de France, Url : <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006082797&dateTexte=vig>

4 Manifestation scientifique **Il était une fois dans le web, 20 ans d'archives de l'internet en France**, 22 et 23 novembre 2016. Captations vidéo, Url : <http://c.bnf.fr/fse>

5 **Groupe de recherches interdisciplinaires sur les processus d'information et de communication du CELSA Paris-Sorbonne**, Url : <http://www.gripic.fr/>

UNE ORIENTATION PARTIELLE ET SYNTHÉTIQUE DANS LE CONTEXTE INTERNATIONAL

Le projet Corpus s'insère dans un contexte plus large, dans un paysage international indéfini et mouvant à cause du caractère émergeant du domaine du *data mining* et de l'absence d'une législation internationale commune et adaptée.

BIBLIOTHÈQUES NATIONALES

Les résultats d'une enquête[1] menée par l'IFLA en 2016 montrent que l'ensemble des bibliothèques nationales est en train de réfléchir à comment répondre aux nouveaux défis imposés par le numérique et aux nouveaux besoins provenant du milieu de la recherche.

Une concrétisation de cette volonté de satisfaire les besoins naissants est à observer aux Pays-Bas, à la **Koninklijke Bibliotheek** (Bibliothèque nationale des Pays-Bas).

Fondé en 2014 au sein du département de la recherche de la Bibliothèque et reposant sur une équipe d'actuellement six personnes (deux conservateurs des collections numériques, deux développeurs et deux chargés de la formation au numérique ou *digital scholarship*), le **KB LAB**[2] est un service qui vise à donner un accès amélioré aux collections numériques et à promouvoir l'exploitation de contenus numériques.

Structuré autour des valeurs d'ouverture, d'expérimentation, de collaboration et de formation mutuelle, le laboratoire virtuel est ouvert à tout public. Cependant l'accès au laboratoire physique, c'est-à-dire à un espace de travail au sein de la Bibliothèque, est limité à un chercheur en résidence (six mois, chercheur en début de carrière, sélectionné suite à un appel à projet) et à un chercheur en poste (*fellowship,* 4 mois, chercheur confirmé, invité), qui pendant un an dialoguent un jour par semaine avec les deux développeurs sur des projets de recherche impliquant l'utilisation des collections numériques de la Bibliothèque. L'objectif de ce dialogue entre les développeurs rattachés au Laboratoire et les chercheurs accueillis est de produire des exemples d'approches et d'opérations possibles ainsi que des outils à intégrer ensuite dans la plate-forme virtuelle. En guise d'exemple, durant l'année 2017, deux chercheurs ont travaillé dans le Laboratoire sur la reconnaissance de forme et la fouille d'images.

Les travaux, les outils réalisés et les activités du Laboratoire font régulièrement l'objet d'événements, de billets de blogue et de *tweets*.

Deux ans après le lancement du KB LAB, en décembre 2016, un rapport[3] sur le projet d'un laboratoire pour la formation ou l'érudition autour du numérique (*digital scholarship)* a été produit par Michelle Gallinger et Daniel Chudnov — des cabinets de conseil Chudnov Consulting et Gallinger Consulting — à la demande du département National Digital Initiatives au sein de la **Bibliothèque du Congrès**.

L'objectif de ce rapport était d'explorer comment fournir les collections numériques conservées par la Bibliothèque sous forme de données brutes aux chercheurs. Pour atteindre cet objectif, les deux consultants ont conduit une série d'entretiens auprès d'universitaires et d'experts du domaine, et ils ont collaboré avec le personnel de la Bibliothèque pour développer un prototype.

La première partie du rapport met en avant les préconisations et les recommandations de l'enquête et sa méthodologie. Selon les deux auteurs, le Laboratoire de la Bibliothèque du Congrès devrait avoir trois missions principales : élargir l'accès aux collections numériques et fournir un accompagnement à l'utilisation du patrimoine numérique unique possédé par la Bibliothèque ; servir de ressource au personnel interne ; servir d'incubateur pour développer, tester et finaliser des services nouveaux.

Pour les deux consultants, le Laboratoire aurait donc intérêt à être utile aussi bien aux chercheurs qu'aux agents de la Bibliothèque.

Un service clé que le Laboratoire devrait offrir est un volet virtuel, c'est-à-dire un site internet exposant les collections numériques, donnant accès aux API et aux autres services mis en place par la Bibliothèque accompagnés de leur documentation. Le Laboratoire devrait également être équipé d'une infrastructure locale ou bien de type *cloud* pour permettre aux chercheurs de travailler de façon autonome et sécurisée. Selon Gallin-

1 Patrice Landry, **National libraries' functions: results from the 2016 survey of national libraries' functions**, Url : <http://library.ifla.org/1722/1/223-landry-en.pdf>

2 **KB LAB**, Url : <http://lab.kb.nl>

3 Michelle Gallinger et Daniel Chudnov, **Library of Congress Lab : Library of Congress Digital Scholars Lab Pilot Project Report**, Url : <http://digitalpreservation.gov/meetings/dcs16/DChudnov-MGallinger_LCLabReport.pdf>

ger et Chudnov, le Laboratoire impliquerait en outre le tissage de liens durables mais non exclusifs avec deux ou trois collaborateurs/ partenaires privilégiés tels que des universités, des instituts spécialisés dans l'analyse computationnelle et d'autres bibliothèques. Enfin, il devrait être un lieu de formation et de valorisation en stimulant l'envie de consulter d'autres collections conservées par la Bibliothèque.

La deuxième partie du rapport retrace le développement du prototype réalisé pour démontrer la faisabilité du transfert d'un corpus numérique, les archives du Web en l'occurrence, vers la plate-forme Amazon Web Services[4] (composée du service de stockage Simple Storage Service, la machine virtuelle Elastic Compute Cloud, la gestion d'accès sécurisés Identity and Access Management et l'outil Elastic MapReduce pour gérer la puissance de calcul répartie sur plusieurs machines), ainsi que la faisabilité de la conversion du format pour l'archivage du Web WARC, lourd et difficile à comprendre pour les chercheurs, au format CSV, léger et facile d'utilisation.

Suite à ce rapport, le 19 septembre 2017, la Bibliothèque du Congrès a ouvert en ligne un laboratoire virtuel[5], un espace pour encourager des usages innovants des collections numériques qu'elle conserve. Plus concrètement, ce site internet donne accès à des projets de production participative (*crowdsourcing)*, à des applications expérimentales ainsi qu'à des tutoriels et à différentes possibilités d'accès aux données pour les machines (*machine-readable*), comme des API par exemple.

Conçu autour d'une valeur d'expérimentation, le laboratoire virtuel est ouvert à tout public. Actuellement quatre personnes des National Digital Initiatives gèrent les Labs et les animent à travers l'organisation d'événements, l'alimentation d'un blogue et l'utilisation d'un compte Twitter.

En octobre 2017, un artiste a été accueilli en tant qu' « innovateur en résidence » par l'équipe des Labs pour explorer et faire des expérimentations à partir des collections numériques.

Le rapport de Gallinger et Chudnov mentionne à plusieurs reprises l'initiative d'une autre bibliothèque nationale : la **British Library**.

Initié quelques années auparavant, en 2013, British Library Labs[6] invite aussi bien les chercheurs que les enseignants, les développeurs, les artistes, en somme un large public à utiliser les collections et les données numériques de la British Library de façon innovante. C'est cet aspect d'ouverture que Gallinger et Chudnov soulignent dans leur rapport car ils voient dans le travail notamment des artistes des opportunités de création et d'inspiration inédites. British Library Labs développe de nouvelles idées ainsi qu'une communauté, notamment à travers des évènements, des projets collaboratifs et une compétition annuelle qui décerne des prix d'une somme modique.

Les bibliothèques nationales réfléchissent à un accès amélioré aux corpus numériques et à la mise en œuvre de services pour inciter les pratiques de fouille de données de type TDM et donc de nouveaux usages sur les collections numériques

Pour la British Library, les Labs représentent aussi l'opportunité de pouvoir étudier les utilisateurs et mieux comprendre comment construire des services autour des collections numériques en s'assurant qu'ils répondent aux besoins actuels.

Comme Adam Farquhar, directeur pour les études numériques (*digital scholarship*), l'avait dit en 2014 lors de la conférence OCLC Research Library Partnership Meeting : « nous avons besoin de faire venir les chercheurs, de nous asseoir à côté d'eux, de voir ce qu'ils font, de collaborer à l'évolution des méthodes, des techniques et des services avec eux ».

Des échanges entre la BnF et la British Library ont permis de comprendre que la proximité avec les chercheurs obtenue grâce aux British Library Labs devrait contribuer à terme au déploiement de services autour des collections numériques au sein de toutes les salles de lecture de la British Library. Dans la perspective britannique, les nouveaux services devront répondre aux besoins d'un large spectre d'usagers, de la découverte de collections, outils et méthodes à une analyse experte des données.

4 **Amazon Web Services**, Url : <https://aws.amazon.com>

5 **Library of Congress Labs**, Url : <https://labs.loc.gov>

6 **British Library Labs**, Url : <http://labs.bl.uk>

BIBLIOTHÈQUES UNIVERSITAIRES

Si pour ce qui concerne les bibliothèques nationales, la mise en place de services innovants à destination du monde académique semble encore au stade de l'expérimentation, les bibliothèques universitaires sont naturellement dans une phase plus avancée en couvrant une offre de services plus ample et adaptée à différents niveaux de compétences (des pratiques de base aux pratiques plus expertes comme le TDM).

Aux Pays-Bas, la **Bibliothèque de l'Université de Leyde**, comme la Bibliothèque bodléienne à Oxford parmi d'autres, abrite un Centre for Digital Scholarship[1] (Centre de formation au numérique).

Reposant sur les valeurs d'orientation, de collaboration et d'adaptabilité, le Centre met les compétences de son équipe à disposition de chercheurs, étudiants ou personnel académique souhaitant développer un projet impliquant l'utilisation de technologies et méthodes innovantes. Dans cette logique, chaque projet est traité comme étant unique et individuel. Le personnel du Centre effectue un travail sur demande et sur mesure mais il peut également orienter et accompagner ponctuellement les lecteurs les plus expérimentés qui peuvent trouver dans le Centre les outils les plus communément utilisés, un conseil juridique, une infrastructure de stockage contenant données et métadonnées interrogeables via des protocoles standard comme des API. Concernant plus spécifiquement la fouille de données de type TDM, le Centre offre plusieurs types d'accompagnement : constitution et fourniture de collections numériques ou numérisées conservées par la Bibliothèque ; conseil autour du nettoyage et de l'enrichissement des données ; accompagnement autour de l'analyse et de la visualisation des données ; support pour la gestion et la conservation des données. Deux scanners et un technicien spécialisé dans la numérisation sont également à disposition des chercheurs. Outre un accompagnement continu, le personnel du Centre organise des ateliers ponctuels et des cours notamment autour du *data management planning*, c'est-à-dire la planification et la gestion du cycle de vie des données de la recherche depuis leur création ou collecte jusqu'à leur utilisation, analyse, stockage, publication, archivage et conservation à long terme.

Services innovants au sein de bibliothèques universitaires, instituts spécialisés dans la science des données et laboratoires virtuels, infrastructures de recherche constituent un paysage divers, indéfini et changeant dans lequel la BnF devra se positionner pour mettre en place le Laboratoire d'étude et d'analyse de corpus numeriques

Aux États-Unis, des espaces consacrés à la découverte, à l'expérimentation et à l'appropriation de technologies numériques existent au sein de plusieurs bibliothèques universitaires. Ces locaux ne sont généralement pas définis comme laboratoires mais plutôt comme *digital scholarship commons*. Cette appellation met en évidence le caractère convivial, collaboratif et interdisciplinaire que ce type d'espace exige.

Un exemple parmi d'autres est celui de la **Bibliothèque Hillman**[2] rattachée à l'**Université de Pittsburgh**. Au rez-de-chaussée de la Bibliothèque, un espace est consacré à l'apprentissage et à l'expérimentation de méthodes et outils numériques pour la recherche et l'enseignement. L'espace est divisé en sous-sections : bureaux du personnel, salle flexible pour l'organisation de formations ou événements, laboratoire doté de douze postes informatiques, service de numérisation, galerie d'exposition. Le personnel accompagne les usagers sur de nombreux sujets : gestion des données de la recherche ou *research data management* ; acquisition et analyse des données (API, OCR, exploration, visualisation ...) ; géolocalisation ; modélisation, vocabulaires et *linked data* ; narrations et expositions multimédias.

Comme pour le Centre de la Bibliothèque de l'Université de Leyde, la vocation des Commons de Pittsburgh est d'offrir un accompagnement sur mesure aux chercheurs ainsi que des cours ponctuels et la possibilité d'une formation mutuelle entre pairs.

Le système de **bibliothèques de l'Université de Stanford**, comme celui de l'Université Columbia, offre également des services innovants aux chercheurs via le Center for Interdisciplinary Digital Research[3] (Centre

1 Bibliothèque de l'Université de Leyde, **Centre for Digital Scholarship**, Url : <https://www.library.universiteitleiden.nl/research-and-publishing/centre-for-digital-scholarship>

2 Bibliothèque Hillman de l'université de Pittsburgh, **Digital Scholarship Commons**, Url : <http://www.library.pitt.edu/digital-scholarship-commons>

3 Bibliothèques de l'Université de Stanford, **Center for Interdisciplinary Digital Research**, Url : <http://library.stanford.edu/research/cidr>

pour la recherche numérique et interdisciplinaire). Créé pour encourager et inspirer des recherches novatrices dans l'ensemble de l'Université, le Centre offre plusieurs services autour de la gestion et de l'analyse de données notamment dans le domaine des sciences humaines et sociales. Le Centre repose sur une équipe composée par neuf personnes : trois bibliothécaires avec des compétences dans le numérique (*digital librarian*), quatre spécialistes dans un domaine scientifique possédant des compétences dans le numériques (*digital humanist*), deux spécialistes dans la gestion des données en sciences humaines et sociales. L'équipe collabore étroitement avec les départements de l'Université et valorise auprès des lecteurs les outils développés par les laboratoires de recherche comme par exemple le célèbre CESTA[4]. Dans une logique pluridisciplinaire et dans une totale intégration aux dynamiques de recherche de l'Université, le Centre organise des formations, des ateliers et participe à la valorisation des projets via un blogue.

En France, le **Learning Center Innovation Lilliad**[5] est exemplaire de cette volonté de rencontre et de circulation des savoirs. Ouvert en septembre 2016 au cœur du campus de l'université de Lille, sciences et technologies, le centre comprend trois pôles : un pôle événementiel, un pôle médiation et une bibliothèque. Le pôle événementiel est conçu comme une interface entre le monde socio-économique et le milieu de la recherche, et il est composé de deux amphithéâtres, de plusieurs salles de commission ainsi que d'une zone d'exposition. Le pôle de démonstration ou Xperium est un espace de médiation, une vitrine qui vise à faire découvrir les principaux axes de recherche de l'université de Lille, sciences et technologies à un public plus large. La bibliothèque universitaire met à disposition ses collections documentaires ainsi que différents espaces : cinquante salles de travail en groupe (4-20 personnes), un espace de détente, une cafétéria (90 places) et une salle d'innovation pédagogique.

INSTITUTS CONSTRUITS AUTOUR DE LA DONNÉE

Si Lilliad encourage le croisement de différents publics et activités à travers le fil conducteur de l'innovation, d'autres centres de recherche ou instituts ne sont accessibles qu'à un public restreint et spécialisé. C'est le cas par exemple de l'Alan Turing Institute et du Data Institute de l'Université Grenoble Alpes.

Fondé en 2015 à Londres par l'Engineering and Physical Sciences Research Council et les universités de Cambridge, Edinburgh, Oxford, University College London et Warwick, l'**Alan Turing Institute**[6] est l'institut national pour la science des données et sa mission consiste à mener des recherches dans le champ de la science des données pour « construire un monde meilleur ».

L'Institut est hébergé par la British Library. Pour le moment la relation entre les deux institutions se limite à l'occupation des locaux, aucune collaboration n'existe entre l'Institut et la Bibliothèque. L'accroissement rapide et considérable de l'Institut commence à attirer l'attention non seulement de nouvelles universités et des entreprises mais aussi de la Bibliothèque. Toujours est-il que les humanités restent un champ de recherche secondaire au sein de l'Institut qui, dès sa fondation, a été centré sur les mathématiques et des domaines économiquement ou politiquement porteurs.

Ni fournisseur de données, ni rattaché à une seule université, l'Alan Turing Institute mérite d'être mentionné dans cette étude pour son statut de plus en plus indépendant, ses sujets de recherche autour des données et pour son organisation singulière.

Depuis cette année l'institut peut demander et obtenir des financements directement ; facteur qui le rend encore plus indépendant.

Le rapport annuel 2016-2017[7] fournit des informations concernant l'organisation de la recherche au sein de l'Institut :

- 7 experts dans la science des données qui constituent le comité scientifique ;
- 16 *full-time Research Fellows* sont recrutés pour mener à l'Institut des projets de recherche d'une durée de 3-5 ans ;
- 4 *Research Software Engineers* recrutés pour transformer les recherches en logiciels de haute qualité ;
- 93 *Faculty Fellows,* c'est-à-dire *senior computer scientists*, statisticiens, mathématiciens et chercheurs en sciences sociales travaillant dans les cinq universités fondatrices et à temps partiel à l'Institut ;
- 14 doctorants de différentes universités travaillant à leur thèse à temps plein à l'Institut ;
- 23 doctorants de différentes universités travaillant à leur thèse pendant un an à l'Institut ;

4 **Center for Spatial and Textual Analysis**, Url : <https://cesta.stanford.edu>

5 Université de Lille, sciences et technologies, **Lilliad**, Url : <https://lilliad.univ-lille.fr>

6 **The Alan Turing Institute**, Url : <https://www.turing.ac.uk>

7 **The Alan Turing Institute Annual Report 2016-2017,** Url : <https://aticdn.s3-eu-west-1.amazonaws.com/2017/07/Turing_AnnualReport_Members.pdf>

- 8 *leading data scientists* accueillis en tant que *Visiting Researchers ;*
- *12 interest groups* pour stimuler la collaboration.

Au niveau de l'aménagement de l'espace, les 18 126 m² contiennent 8 salles de réunion, 143 postes informatiques, des loges pour le travail individuel, des espaces de détente et une cuisine.

Environ 5 000 personnes ont assisté aux 120 événements (conférences, séminaires, ateliers) organisés dans l'année 2016-2017; environ 38 000 ont vu les captations vidéo en ligne.

Extrêmement différent en termes de conception, d'organisation et d'objectifs, le **Data Institute**[8] de l'**Université Grenoble Alpes** partage avec l'Alan Turing Institute la centralité de la donnée mais si le slogan de l'Alan Turing Institute est « changer le monde grâce à la sciences des données », celui du Data Institute est investiguer et comprendre « comment les données changent la science et la société ».
Ni indépendant, ni intégré au sein d'une bibliothèque universitaire mais rattaché à une seule université, le Data Institute représente donc un autre cas de figure dans ce paysage complexe et en cours de définition.

LABORATOIRES VIRTUELS

Un autre élément pour présenter une esquisse du panorama dans lequel cette étude s'insère est la présence de laboratoires virtuels. Deux sortes de laboratoires virtuels appartenant à des bibliothèques nationales ont déjà été mentionnées précédemment dans ce rapport, mais il existe également des laboratoires virtuels qui ne sont pas liés à une seule bibliothèque.

Un exemple majeur de ce type de ressource est le **HathiTrust**[9], ainsi nommé en référence à la mémoire particulièrement durable de l'éléphant Hathi en hindi.

Cette plate-forme a été fondée aux États-Unis en 2008 par treize universités. Aujourd'hui plus de soixante bibliothèques américaines, canadiennes et européennes sont partenaires et alimentent la bibliothèque numérique. Conçu autour des valeurs de partage, collaboration, standardisation et conservation, HathiTrust a pour missions principales de permettre la co-gestion et la conservation des collections numériques des bibliothèques partenaires et d'augmenter sensiblement l'accès aux contenus numériques.

Particulièrement pertinent pour la présente étude est le lancement en 2012 du HathiTrust Research Center[10]. Il s'agit d'un espace virtuel qui facilite les usages non commerciaux ou à des fins de recherche des contenus de la bibliothèque numérique en permettant des analyses computationnelles des matériaux dans le domaine public et, de façon plus limitée, de certains matériaux sous droit. L'infrastructure du HTRC met à disposition plus concrètement : la possibilité de télécharger et travailler sur sa propre machine sur des corpus extraits de données même sous droit (c'est le cas de corpus des

8 Université Grenoble Alpes, **Data Institute**, Url : <https://data-institute.univ-grenoble-alpes.fr>

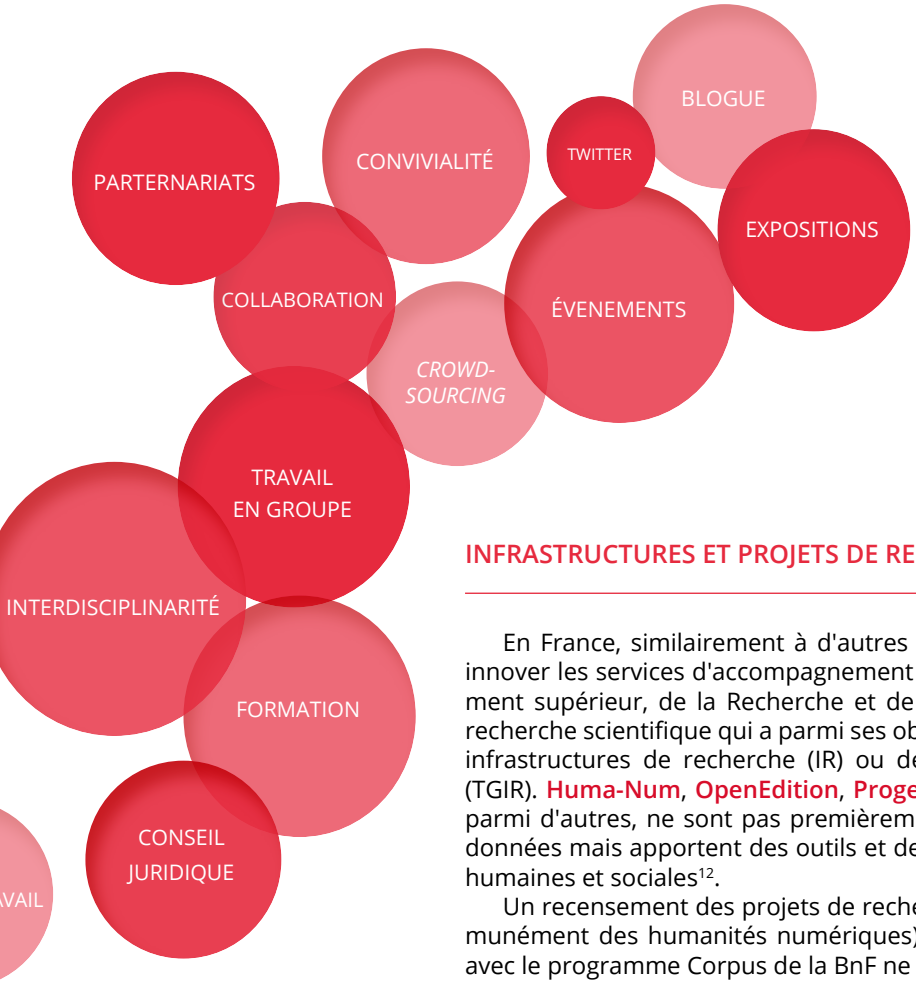
9 **Bibliothèque numérique HathiTrust**, Url : <https://www.hathitrust.org>

10 **HathiTrust Research Center**, Url : <https://www.hathitrust.org/htrc>

métadonnées de différentes granularités de description — du corpus à la page — et des comptages de mots) ; une série d'algorithmes qui permettent l'analyse et le téléchargement de sous-ensembles de données identifiés par l'usager ; un environnement de travail sécurisé, c'est-à-dire un espace personnel appelé « capsule », une machine virtuelle plus précisément où l'usager peut analyser ou faire des visualisations des données de la bibliothèque numérique en utilisant des outils fournis dans la plate-forme mais aussi d'autres outils. L'exportation des données dérivées est soumise à approbation.

Alveo[11] est un autre exemple de laboratoire virtuel qui permet de façon similaire au HTRC la fouille et l'analyse de données. Subventionné par le projet australien National eResearch Collaboration Tools and Resources, ce Laboratoire est le résultat de la collaboration de treize universités et trois organisations australiennes. La mission de ce Laboratoire est d'inciter à opérer la transition d'un modèle de recherche basé sur l'isolement (*desk-PC-lab-university-bound)* à un nouveau modèle de recherche basé sur le partage des analyses et des données de la recherche, sur le repérage et la combinaison d'outils grâce à une certaine sérendipité et sur le stockage de données et d'outils sur un espace public et collaboratif. L'idée qui sous-tend Alveo est donc de connecter les chercheurs, leurs bureaux, leurs ordinateurs, leurs laboratoires et universités pour accélérer la recherche et favoriser l'interdisciplinarité. Les caractéristiques principales de cette plate-forme sont : l'accessibilité à des chercheurs inexpérimentés à travers des interfaces interactives, des outils permettant d'établir un flux de travail et de suivre des protocoles ; l'interopérabilité des corpus provenant de différentes institutions ; la durabilité grâce aux financements, récoltés par les institutions fondatrices (treize universités et trois organisations) et quarante-sept investisseurs, qui s'élèvent à environ deux millions de dollars et permettent d'assurer la maintenance ainsi que des développements futurs.

11 **Above and Beyond Speech, Language and Music, A Virtual Lab for Human Communication Science**, Url : <http://alveo.edu.au>



INFRASTRUCTURES ET PROJETS DE RECHERCHE

En France, similairement à d'autres pays européens, un dernier acteur qui vise à innover les services d'accompagnement aux chercheurs est le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, tutelle du Centre national de la recherche scientifique qui a parmi ses objectifs d'assurer un déploiement territorial des infrastructures de recherche (IR) ou des très grandes infrastructures de recherche (TGIR). Huma-Num, OpenEdition, Progedo, RNMSH et E-RIHS, comme le projet ISTEX parmi d'autres, ne sont pas premièrement centrés sur l'analyse computationnelle de données mais apportent des outils et des moyens de recherche innovants en sciences humaines et sociales[12].

Un recensement des projets de recherche (notamment dans le champ appelé communément des humanités numériques) et des acteurs qui entreraient en résonance avec le programme Corpus de la BnF ne constitue pas l'objectif de la présente étude.
L'« état de l'art » dressé tout au long des pages précédentes est à considérer comme une orientation partielle et synthétique mais utile pour encadrer l'enquête qui fait l'objet de cette étude de besoins et pour fournir à la BnF des informations qui, croisées avec d'autres recherches, lui permettront de se positionner plus consciemment dans ce paysage divers, indéfini et changeant.

12 **Feuille de route Infrastructures de recherche 2016**, Url : <http://cache.media.enseignementsup-recherche.gouv.fr/file/Infrastructures_de_recherche/74/5/feuille_route_infrastructures_recherche_2016_555745.pdf>

SERVICES	KB (KB LAB)	BIBLIOTHÈQUE DU CONGRÈS (LoC LABS)	BRITISH LIBRARY LABS
ESPACE ET SERVICES PHYSIQUES	X (SI EN RÉSIDENCE)	X (*INNOVATOR-IN-RESIDENCE*)	(PARFOIS ACCUEIL DANS LES BUREAUX DU PERSONNEL)
ESPACE ET SERVICES VIRTUELS ACCESSIBLES À DISTANCE	X (DONNÉES OUVERTES ET SOUS DROIT SUR DEMANDE)	X (DONNÉES OUVERTES)	X (DONNÉES OUVERTES)
TOUT PUBLIC INTÉRESSÉ	X (LABO VIRTUEL)	X (LABO VIRTUEL)	X
PUBLIC UNIVERSITAIRE	X (LABO VIRTUEL ET PHYSIQUE MAIS ACCÈS RESTREINT)	X (LABO VIRTUEL ET PHYSIQUE MAIS ACCÈS RESTREINT)	X
COLLECTIONS	X	X	X
FOURNITURE DE DONNÉES	X	X	X
ENRICHISSEMENT D'UN CORPUS (NUMÉRISATION, COLLECTE)	?	?	?
ACCOMPAGNEMENT SUR MESURE	X (SI EN RÉSIDENCE)	X (SI EN RÉSIDENCE)	X (SI ACCUEIL DANS LES BUREAUX)
ACCOMPAGNEMENT AUTOUR DU DMP(*)	?	?	NON
ACCOMPAGNEMENT JURIDIQUE	?	?	NON
DÉVELOPPEMENT D'OUTILS AVEC LES CHERCHEURS	X (SI EN RÉSIDENCE)	?	?
PALETTE D'OUTILS	X	X	(INDICATION D'OUTILS EXISTANTS)
FORMATIONS	X	?	?
ÉVÉNEMENTS	X	X	X
ÉQUIPE	6 PERSONNES 2 DÉVELOPPEURS 2 SPÉCIALISTES DS (**) 2 CONSERVATEURS "NUMÉRIQUES"	4 PERSONNES 1 CHEF 1 CHEF DE PROJET 1 SPÉCIALISTE DE L'INNOVATION 1 COORDONNATEUR DES ÉVÉNEMENTS	3 PERSONNES 1 CHEF DE PROJET 1 RESPONSABLE TECHNIQUE 1 AGENT POLYVALENT

(*) ***DATA MANAGEMENT PLANNING*** = PLAN DE GESTION DES DONNÉES

(**) ***DIGITAL SCHOLARSHIP*** = FORMATION, ÉTUDES, CULTURE, ÉRUDITION NUMÉRIQUE

BU LEYDE (CENTRE POUR DS(**))	ALAN TURING INSTITUTE	HATHITRUST (HTRC)	BNF
X (SI UNIVERSITÉ DE LEYDE)	X (SI CONTRAT)	NON	X (PROJET CORPUS)
X (SI UNIVERSITÉ DE LEYDE)	?	X	X (DONNÉES OUVERTES)
NON	NON	(OUVERTURE PLUS LARGE SUR DEMANDE)	? (À TERME)
X (SI UNIVERSITÉ DE LEYDE)	X (SI CONTRAT)	X	X
X	NON	X	X
X	NON	X	X
X	NON	?	X (PROJET CORPUS OU CONVENTIONS)
X	X (ENTRE PAIRS)	NON	X (PROJET CORPUS OU CONVENTIONS)
X	?	NON	?
X	?	NON	X
?	X (PAR LES CHERCHEURS)	NON	X (PROJET CORPUS OU CONVENTIONS)
X	X (PLUTÔT DÉVELOPPEMENT)	X	X (PROJET CORPUS)
X	X	X	X (À PRÉVOIR ET ACCUEILLIR)
X	X	X	X (PROJET CORPUS)
5 PERSONNES 1 RESPONSABLE DS(**) 4 BIBLIOTHÉCAIRES DS(**)	n > 50 PERSONNES 6 COMMUNICATION 4 SERVICE INFORMATIQUE 3 ASSISTANTS ACADÉMIQUES 11 ASSISTANTS DE RECHERCHE 10 RELATIONS ACADÉMIQUES 9 DÉVELOPPEURS, etc ...	7 PERSONNES 2 CO-DIRECTEURS 1 RESP. RECHERCHE - TECHNOLOGIE 1 BIBLIOTHÉCAIRE 1 COMMUNICATION 1 DIRECTEUR IT 1 SPÉCIALISTE HUMANITÉS NUMÉRIQUES	2 PERSONNES ? 1 AGENT POLYVALENT ET SPÉCIALISTE DS(**) 1 SPÉCIALISTE ANALYSE DES DONNÉES

L'ÉTUDE DE BESOINS : VOLETS ET OBJECTIFS

La présente étude découle de la nécessité de cerner les besoins des usagers potentiels du Laboratoire d'étude et d'analyse des corpus numériques que le projet Corpus vise à mettre en place. Malgré les échanges que la BnF perpétue avec le monde universitaire, une enquête prospective auprès des publics potentiels du nouveau service est apparue utile afin de préciser et mesurer leurs besoins, leurs désirs et leurs attentes.

L'enquête, d'une durée de cinq mois (août – décembre 2017), a été structurée en trois volets :

- une **enquête qualitative** par entretiens ;
- **deux ateliers** ou séances d'échange et de débat ;
- un **atelier collaboratif**.

Le **premier volet** a consisté à conduire des entretiens qualitatifs auprès de trois populations distinctes: universitaires (chercheurs, enseignants-chercheurs, professeurs, ingénieurs de recherche), membres du personnel de la BnF et experts du domaine travaillant au sein d'autres établissements. Le choix des trois terrains d'enquête a permis non seulement de récolter les visions des usagers potentiels mais aussi de relativiser, mesurer et pondérer ces points de vue grâce à l'apport d'autres perspectives.

L'échantillon d'interlocuteurs a été défini en plusieurs étapes. Un noyau initial a été constitué sur la base de la proximité. Onze agents de la BnF, seize universitaires et trois experts extérieurs ont été sollicités et constituent un échantillon total de trente interlocuteurs interrogés sur un laps de temps de trois mois (août-octobre 2017).

Le **deuxième volet** a consisté à organiser deux ateliers ou plus précisément deux séances d'échange et de débat de type *symposium*. Le premier atelier a eu lieu le 16 octobre 2017 sur le site François Mitterrand et le deuxième atelier a eu lieu le 30 novembre 2017 sur le site Richelieu. Conçus comme des moments catalyseurs d'échanges et de partage d'expertise entre le milieu de la recherche et le monde des bibliothèques, ces deux ateliers ont été structurés autour de deux thèmes découlant des questions posées par l'équipe de recherche du GRIPIC travaillant sur le projet Giranium, l'équipe de recherche témoin du projet Corpus pour l'année 2017.

Le matin du 16 octobre, devant une audience composée d'acteurs du monde de la recherche et d'agents de la BnF, l'atelier « Décrire, transcrire et diffuser un corpus documentaire hétérogène : méthodes, formats et outils » a mené une réflexion sur la gestion de l'hétérogénéité des données.

L'après-midi du 30 novembre, une quarantaine de représentants du monde de la recherche et des institutions patrimoniales ont participé à l'atelier « Géolocalisation et spatialisation de documents patrimoniaux » qui a porté essentiellement sur différents projets prenant en compte les enjeux de la cartographie numérique.

Les comptes-rendus[1] des deux ateliers ont été publiés sur le blogue « Carnet de la recherche à la Bibliothèque nationale de France » pour valoriser le projet Corpus et diffuser les échanges auprès d'un public plus large.

Le **troisième volet** a consisté à préparer et à organiser un atelier collaboratif utilisant la méthode des *personas*.

Sur la base de la recherche menée dans le cadre du volet I, et suite à une analyse approfondie du terrain et de la matière récoltée, cinq *personas*, c'est-à-dire cinq profils d'usagers potentiels, ont été dessinés. Les *personas* ont été ensuite utilisés pour imaginer et esquisser des parcours d'usagers au sein et en dehors du futur Laboratoire d'étude et d'analyse de corpus numériques. À l'atelier, qui a eu lieu le 8 décembre 2017 sur le site Richelieu, ont été conviés six chercheurs et cinq agents de la BnF sollicités lors

1 Eleonora Moiraghi, **Décrire, transcrire et diffuser un corpus documentaire hétérogène : méthodes, formats, outils**, Carnet de la recherche à la Bibliothèque nationale de France, Url : <https://bnf.hypotheses.org/2214>

Eleonora Moiraghi, **Géolocalisation et spatialisation de documents patrimoniaux : trois heures de partage autour de la cartographie numérique**, Carnet de la recherche à la Bibliothèque nationale de France, Url : <http://bnf.hypotheses.org/2299>

de la première phase de l'étude (volet I) ainsi que dix agents de la BnF non sollicités précédemment mais impliqués dans les axes de recherche du projet Corpus.

Si les entretiens avaient pour objectif de cerner les besoins des usagers potentiels et de discuter leurs propos à travers le contre-champ donné par les experts, les ateliers avaient pour but d'entrevoir le hors champ. Autrement dit, l'objectif des trois volets de l'étude est de synthétiser les besoins et les attentes des universitaires en les mettant en regard de l'expérience des experts et avec les comportements observés lors des ateliers pour contribuer à terme à la définition d'un service opérationnel orienté usager et donc adapté aux besoins des utilisateurs futurs.

ENQUÊTE QUALITATIVE	DEUX ATELIERS/ DÉBATS	ATELIER COLLABORATIF
- UNIVERSITAIRES - AGENTS DE LA BNF - EXPERTS	- UNIVERSITAIRES - AGENTS DE LA BNF	- UNIVERSITAIRES - AGENTS DE LA BNF
=	=	=
30 INTERLOCUTEURS	**60** INVITÉS	**21** PARTICIPANTS

2. MÉTHODOLOGIE

L'ENQUÊTE QUALITATIVE ET LES OBSERVATIONS INFORMELLES

L'**approche** adoptée pour l'enquête relève des sciences humaines et sociales. Elle est **qualitative, compréhensive, inductive** et elle combine techniques formelles et informelles de collecte d'informations.

Elle est qualitative car elle consiste à conduire des entretiens individuels d'une durée d'une heure chacun en moyenne.

Elle est compréhensive car elle prend en compte le point de vue des interlocuteurs.

Elle est inductive car elle laisse émerger les informations du terrain plutôt que les canaliser, les faire rentrer dans des cases préétablies.

Elle combine des techniques formelles et informelles de collecte d'informations car, outre les entretiens, des observations informelles ont été effectuées à l'occasion des ateliers.

Chacun des trente entretiens a été conduit par la même intervieweuse qui a été recrutée par la BnF spécifiquement pour réaliser la présente étude. Des grilles d'entretien différentes et adaptées à chaque population ont été rédigées avant de commencer l'enquête par Eleonora Moiraghi (assistante de recherche sur le projet Corpus et intervieweuse) avec l'aide d'Emmanuelle Bermès (Adjointe scientifique et technique au Directeur des Services et des Réseaux à la BnF) et Philippe Chevallier (Responsable des études à la BnF). Cependant les principaux thèmes abordés sont communs à tous les guides d'entretien :

- La fouille de données de type TDM ;
- Les expériences d'utilisation ou bien de fourniture de corpus numériques ;
- Les problèmes rencontrés dans ces expériences ;
- La définition du rôle de la BnF dans l'écosystème de la recherche ;
- L'idée d'un Laboratoire d'étude et d'analyse de corpus numériques au sein de la BnF ;
- L'infrastructure nécessaire ;
- Les outils, les pratiques, les temporalités ;
- La structuration de l'espace et la place de la convivialité ;
- Les modalités d'accompagnement ;
- Les services similaires en France ou à l'étranger.

Comme l'objectif premier de l'enquête était de définir les besoins liés à un lieu physique dans les emprises de la BnF, les thèmes explorés pendant les entretiens prennent en compte d'autres aspects du projet Corpus mais servent à porter l'attention sur l'idée d'un espace physique.

Au début de chaque entretien une demande d'autorisation d'enregistrement a été formulée. La plupart des entretiens ont donc pu être enregistrés et transcrits. Il y a deux avantages à cette approche : l'enregistrement garantit une fidélité considérable aux propos et la transcription permet une analyse qui peut tenir compte de la complexité.

LA MÉTHODE DES *PERSONAS*

La méthode des *personas* est une technique utilisée dans le domaine du **design d'expérience utilisateur (UX)** qui a été vulgarisée dans les années **1990** par Alan Cooper, designer d'interfaces et développeur américain. Selon certains, la méthode des *personas* « est une technique relativement ancienne qui a notamment été utilisée dès 1993 chez Apple »[1]. Inventeur ou pas, Alan Cooper a sûrement le mérite d'avoir diffusé et popularisé cette méthode, d'abord en l'anticipant dans son ouvrage « *About face : the essentials of user interface design*[2] » publié en 1995 et ensuite en l'expliquant dans sa mythique publication

1 Sylvie Daumal, **Design d'expérience utilisateur : principes et méthode UX**, Paris, 2012, p. 89

2 Alan Cooper, **About face: the essentials of user interface design**, Foster City, 1995

« *The Immates Are Running the Asylum*[3] » parue en 1999.

Dans la XVIII partie du premier ouvrage, Cooper évoque son secrétaire imaginaire Rodney, il trace son comportement dans le classement de fichiers et il le compare avec celui d'Elliot, un autre secrétaire imaginaire. En effet, les *personas* sont des personnages imaginaires qui reflètent différents profils types d'utilisateurs. Tous les deux secrétaires, Rodney est « consciencieux », Elliot en revanche est un complet « idiot, pas du tout consciencieux ». En personnifiant différentes catégories, ces « **archétypes spécifiques et représentatifs** »[4] servent à aider une équipe à construire un produit ou un service en développant de l'**empathie** afin d d'éviter de « se référer à l'utilisateur comme à une abstraction ». Les *personas* ont comme toute personne réelle un prénom ou un nom, une image, un âge, des comportements, des caractéristiques, des motivations et des objectifs mais ils ne sont pas le seul fruit de l'imagination, ils sont le résultat de deux phases, piliers de la méthode : une première phase de **recherche** et une deuxième d'**analyse**.

La phase de recherche consiste à mener une enquête auprès des utilisateurs potentiels car, à travers des entretiens, il est possible de mieux comprendre les publics potentiels et donc d'identifier des profils similaires.

La phase d'analyse consiste à rassembler tous les éléments récoltés durant la recherche et à définir les profils à partir des similarités qui ont émergé.

En l'occurrence, l'enquête qualitative comprenant les trente entretiens précédemment mentionnés a constitué la phase de recherche.

La phase d'analyse a ensuite consisté à relire attentivement les transcriptions des entretiens et à identifier des critères en termes de caractéristiques, comportements et objectifs. Sur la base de ces critères, les interlocuteurs ont été placés par l'intervieweuse sur des échelles (voir Annexe D) qui ont permis à sept collègues d'identifier des regroupements correspondants à **cinq profils types d'utilisateurs** et donc à cinq *personas*.

Dans le respect de la recommandation générale qui prescrit le nombre des *personas* entre trois et sept par projet, les cinq personas ont été ensuite utilisés lors de l'**atelier collaboratif**, volet III de la présente étude. Durant trois heures, vingt-deux participants divisés en cinq groupes — un groupe par *persona* — se sont d'abord approprié les *personas* à travers les fiches préalablement préparées (voir Annexe E) et grâce à l'exploration de deux questions posées par les animatrices autour du point de contact entre le *persona* et la BnF et autour du rapport avec un autre *persona*.

Suite au stade de l'appropriation, le cœur de l'atelier a été constitué par le remplissage d'un tableau visant à faire des hypothèses sur les activités des personas non seulement au sein du futur Laboratoire mais également en dehors (à la maison, à l'université, à la BnF, ailleurs) afin de vérifier, de confirmer ou remettre en cause les résultats de l'enquête qualitative.

Les groupes ont été établis en fonction de la proximité au *persona*. Les six chercheurs ont été notamment placés dans les groupes travaillant sur le *persona* le plus proche dans la limite du possible (les deux *personas* plus extrêmes, Alexis et Cécile, n'avaient pas de représentant) afin de favoriser le développement de l'empathie et des mécanismes de projection.

Une place importante a été accordée dans l'atelier à la restitution pour inciter la critique et l'échange entre les participants, et donc pour permettre le relevé et l'observation des réactions.

3 Alan Cooper, **The Immates Are Running the Asylum**, Indianapolis, 1999

4 S. Daumal, p. 89

3. RÉSULTATS

LES RÉSULTATS DE L'ENQUÊTE QUALITATIVE

Les résultats de l'enquête sont présentés dans ce chapitre à travers huit axes de recherche contenus dans les grilles d'entretiens.

Les huit axes sont les suivants :

1. LA FOUILLE DE DONNÉES DE TYPE TDM
2. LES EXPÉRIENCES DE FOURNITURE OU D'UTILISATION DE CORPUS NUMÉRIQUES, LES PROBLÈMES RENCONTRÉS
3. LES SERVICES SIMILAIRES EN FRANCE OU À L'ÉTRANGER
4. LE RÔLE DE LA BNF DANS L'ÉCOSYSTÈME DE LA RECHERCHE
5. L'IDÉE D'UN LIEU PHYSIQUE DANS LES EMPRISES DE LA BNF
6. LA CONVIVIALITÉ ET L'AGENCEMENT DE L'ESPACE
7. L'INFRASTRUCTURE, LES OUTILS, LES PRATIQUES ET LES TEMPORALITÉS
8. L'ACCOMPAGNEMENT

1 Comment définiriez-vous la fouille de données ? Quelle importance a la fouille de données dans la recherche ? Est-elle appelée à se développer ? (Quelle évolution dans votre discipline ?)

À la demande d'une définition de la fouille de données, chaque universitaire a évidemment répondu avec une définition différente dépendant de ses sujets de recherche et de son niveau d'expertise dans le domaine.

De manière similaire, les experts ont donné des définitions diverses en fonction de leurs missions, leurs activités et leur niveau de connaissance du sujet.

Malgré la variété des réponses, la fouille de données de type TDM est communément identifiée comme étant une réponse au défi posé par les gros volumes de données : certaines choses ne sont pas visibles à l'œil nu.

Les chercheurs ayant une formation uniquement dans une branche de l'informatique (licence, master et doctorat) mettent plutôt l'accent sur les aspects techniques. La définition d'un docteur spécialisé dans le traitement du signal s'articule autour des différentes phases qui constituent la fouille de données : le *scraping*, c'est-à-dire la récupération des données, l'organisation de la donnée, le nettoyage « car généralement on récupère plus de données que nécessaire », l'enrichissement (« on va chercher des données extérieures pour grossir notre base de données ») et le développement d'algorithmes « qu'on définit en fonction d'un objectif ».

L'attention portée à l'expertise technique ressort également dans une affirmation du type : « La fouille de données reste un truc d'ingénieurs » ; phrase prononcée par une jeune ingénieure informatique.

Du point de vue d'un professeur en sciences du langage, c'est-à-dire quelqu'un qui travaille dans le traitement automatique des langues, la définition de fouille est évidemment plus circonscrite au texte :
« Mon matériau, il va être langagier par rapport à de la fouille de données qui serait quantitative. Moi je m'intéresse à la fouille de données textuelles et la fouille de données textuelles consiste à rechercher dans

un matériau langagier soit un motif lexical, c'est-à-dire une suite de termes lexicaux, soit un motif morphosyntaxique et éventuellement des motifs plus sémantiques ».

Pour un maître de conférences en linguistique informatique la distinction entre l'extraction de données et le *data mining* est aussi importante. L'extraction implique qu' « on sait ce qu'on veut extraire et on va trouver des moyens automatiques pour le faire, alors qu'avec *text mining* et *data mining* on va essayer de trouver des choses qu'on ne connaît pas forcement au départ. Le *data mining* sert à découvrir des choses latentes, invisibles même. On voit bien la séparation ».

Selon une enseignante-chercheuse en sciences sociales, pour élaborer une définition de fouille de données il faut opérer une distinction entre extraction d'informations et *data mining*, *text mining*, *web mining, stream mining* et *machine learning*, mais aussi entre deux approches de l'apprentissage automatique : supervisée et non supervisée. Si l'objectif de la première approche est de vérifier une hypothèse faite a priori, celui de la deuxième est à l'opposé de faire émerger « de la gangue des données le pur diamant de la véridique nature[1] ».

« Dans un gros volume de données, il y a des choses qui ne sont pas visibles à l'œil nu »

La plupart des littéraires et des « humanistes » sollicités quant à eux ont eu une approche plus critique dans l'élaboration d'une définition. Pour le directeur d'un Labex et professeur de littérature française, il ne faudrait pas parler de « fouille de données au singulier mais de fouilles de données au pluriel ».

Selon un ingénieur d'études travaillant depuis plusieurs années dans les sciences humaines, il n'y a pas seulement la question de la fouille mais surtout la question du corpus « qui n'est pas plat. Quand on fait de la fouille de données, il est nécessaire de limiter un corpus et de considérer qu'il y a des trous dans les collections patrimoniales, des biais. La fouille est une question scientifique, elle naît d'une attente, d'une envie de savoir avant tout, même avant une hypothèse ou une intuition. Le corpus n'est pas figé, il doit pouvoir se transformer au fur et à mesure que la question évolue ».

L'œil critique est aussi celui d'un enseignant-chercheur en sciences de l'information et de la communication qui considère que la fouille de données est « juste une méthode » et qui porte plutôt l'attention sur les techniques comme par exemple le *topic modeling*[2] ou le *word embedding*[3].

Il serait excessivement réducteur de tirer des conclusions à partir de cette première question dont l'objectif était simplement de clarifier le terrain de jeu, d'identifier le point de vue et de rentrer dans le vif du sujet.

Si la première question a suscité des réponses diverses et variées, les interlocuteurs (chercheurs, agents de la BnF et experts confondus) s'accordent à l'unanimité sur l'importance croissante de la fouille de données de type TDM dans le monde de la recherche, et sur son inévitable développement et impact dans tout domaine de la connaissance notamment grâce au *machine learning* et au *deep learning*. Les interlocuteurs sont également d'accord sur le fait que les techniques de fouille de données ne constitueront pas l'approche exclusive mais que les approches plus traditionnelles ou non outillées continueront à perdurer et à accompagner ce type de recherches plus innovantes et outillées qui visent dans tous les cas à

1 Jean-Paul Benzécri, **Histoire et préhistoire de l'analyse des données**, Les Cahiers de l'analyse des données, n° 1, 1976, p.144

2 En apprentissage automatique et en traitement automatique du langage naturel, un *topic model* ou modèle thématique est un modèle probabiliste qui permet de repérer des thèmes ou des sujets dans un document ou un ensemble de documents. Il repose sur l'idée qu'un sujet est une distribution probabiliste d'un ensemble de mots.

3 En apprentissage automatique et en traitement automatique du langage naturel, un plongement de mots est une technique qui consiste à représenter chaque mot d'un dictionnaire par un vecteur de nombres réels correspondant. Elle permet de faciliter l'analyse sémantique et elle est utilisée notamment dans les domaines de l'analyse des sentiments et de la traduction.

l'aboutissement de productions savantes.

Il est pourtant important de souligner que l'échantillon était constitué d'interlocuteurs ayant une familiarité plus ou moins accrue avec la pratique ou le concept de fouille de données et que donc ils restituent inévitablement une vision positive quant à l'évolution de ces méthodes.

2 Avez-vous utilisé (ou fourni) des corpus de données ?
Pouvez-vous me décrire ces expériences brièvement ?
Quelles sont les problèmes que vous avez rencontrés sur votre chemin ?

Les expériences d'utilisation ou de fourniture de corpus numériques pourrait être analysées suivant quatre aspects différents : techniques, juridiques, organisationnels et humains.

ASPECTS TECHNIQUES

Qu'ils aient utilisé des corpus de presse, d'images, de textes scientifiques, extraits de réseaux sociaux ou d'archives de l'Internet, de Persée, Europeana, data.bnf.fr ou Gallica, qu'ils soient débutants ou plus experts, les chercheurs ou les ingénieurs de recherche qui font eux-même de la fouille de données rencontrent tous quelque part au long de leur parcours des contraintes d'ordre technique.

Les aspects techniques peuvent être au cœur même de la recherche ou bien dus à un manque d'aisance numérique mais dans certains cas, ils dépendent des fournisseurs de données.

Du point de vue d'un docteur spécialisé dans le traitement du signal, la fourniture de formats type dictionnaire comme JSON est préférable à la fourniture de formats type arbre comme XML (qui reste pourtant important pour des chercheurs travaillant dans d'autres disciplines, notamment pour ce qui concerne la

manipulation de documents).

La question des formats est problématique pour des ingénieurs de recherche qui travaillent avec des sérialisations de RDF. Selon un ingénieur de recherche, l'auto complétion des propriétés est indispensable dans les SPARQL *endpoints* et il souhaiterait obtenir automatiquement une requête SPARQL suite à une requête traditionnelle dans le catalogue de la BnF.

L'utilisation de corpus d'images implique aussi des contraintes techniques en matière de format ou plus précisément de définition. Le format JPEG n'est pas énormément problématique en soi (sauf dans le cas d'une forte compression), ce qui est contraignant est la fourniture d'images en basse définition qui rend plus difficile le travail d'historiens de l'art sur des inscriptions par exemple ou la qualité de l'OCR. Même si IIIF permet d'obtenir les images en haute définition en modifiant l'URL, les chercheurs ou ingénieurs de recherche préféreraient pouvoir télécharger dans Gallica les images en haute définition sans devoir passer par la modification de l'URL.

Pour un enseignant-chercheur en sciences de l'information et de la communication, les fournisseurs de données devraient aussi fournir des données plus « propres », de son point de vue : « Souvent on doit passer un temps de fou à nettoyer les données. Il y a énormément d'erreurs d'OCR [...] Souvent on constate que les données sont disponibles dans un format de type TXT qui est non-structuré et qui contient énormément d'erreurs d'encodage ». La fourniture des deux formats ALTO et TXT est préférable et l'explicitation du taux d'erreur de l'OCR fondamentale.

En effet, des interventions lors du premier atelier l'ont confirmé, encore plus qu'un éventail de différents formats, la communauté scientifique demande aux fournisseurs de données plus de « transparence » , c'est-à-dire l'explicitation des biais, des taux d'erreurs, des limites d'un corpus, de son histoire, des phases de sa constitution ainsi que de la chaîne de traitement, les outils et les méthodes adoptés.

ASPECTS ORGANISATIONNELS ET HUMAINS

Du point de vue des **agents de la BnF**, les contraintes techniques résident dans l'hétérogénéité de la qualité des données (défauts d'OCR, différentes qualités de numérisation) et dans la complexité de leur gestion (cas d'absence/ disparition de documents) qui constituent de véritables défis.

Ces aspects techniques sont accompagnés par des **aspects organisationnels** lors de demandes complexes qui impliquent l'accueil des chercheurs dans les bureaux réservés au personnel de la BnF. Les projets de recherche auxquels la Bibliothèque décide de contribuer sont chronophages et avides de ressources humaines à cause de la durée limitée des projets et souvent du manque de ressources financières qui ne permet pas l'investissement dans des compétences humaines des ingénieurs de recherche. Les équipes de recherche vont donc se reposer sur les compétences internes à la Bibliothèque, non sans difficulté de repérage du bon interlocuteur.

Mais pour les équipes de recherche qui ont la possibilité d'être entourées par un ou plusieurs ingénieurs de recherche, si les questions strictement techniques ne suscitent pas autant de préoccupations, la gestion de l'**interdisciplinarité** entraine souvent d'autres contraintes. Le constat d'une enseignante-chercheuse en sciences sociales est que « les ingénieurs ont perdu la connexion avec les humanités et à l'inverse les littéraires ont une certaine méfiance envers la machine. Et les critères entre les disciplines sont très différents ». Plus qu'une méfiance c'est un regard critique qu'un directeur de Labex et professeur de littérature française porte sur l'interdisciplinarité et l'utilisation de technologies numériques : « On travaille dans une forte complicité entre les informaticiens, les ingénieurs et les chercheurs en littérature. Il ne faut pas penser que c'est n'importe quel outil qui va répondre à nos besoins. On est dans une complémentarité, quand on écrit un article l'ingénieur rédige la partie technique et moi j'assume l'interprétation, la partie littéraire. Il nous arrive d'avoir des problèmes avec les enseignants-chercheurs en informatique parce que les visées d'un informaticien ne sont pas les mêmes que celles d'un littéraire. Le problème qui se pose dans toute interdisciplinarité, c'est une instrumentalisation de l'autre discipline. Le principal pour l'informaticien c'est qu'il teste l'efficacité de son logiciel. Or pour nous, la perspective n'est pas là ».

La question des publications interdisciplinaires n'est pas simple. Plusieurs interlocuteurs mettent en avant l'intérêt de mener une recherche en binôme (chercheur en SHS et ingénieur) mais nous alertent sur la difficulté pour l'ingénieur de la publier du fait qu'elle ne soit pas suffisamment valorisante d'un point de vue technique.

Parmi les personnes interpellées dans le cadre de cette étude, le dialogue entre chercheur en SHS et

ingénieur informatique semble être aussi fascinant que compliqué. Selon une professeure spécialisée dans le traitement de l'image « les gens en SHS au début pensent qu'on ne sait rien faire et c'est vrai parce qu'on n'a jamais 100% de résultats. Quand on demande quelque chose, il y a toujours des erreurs. Par ailleurs ils n'imaginent pas ce qu'on pourrait faire, c'est-à-dire qu'ils ne mesurent pas la difficulté de certaines choses qui sont évidentes pour l'œil que nous on a beaucoup de mal à faire automatiquement et puis il y a des choses qui nous ne demandent pas trop de travail mais qu'ils n'imaginent pas ».

Malgré ces aspects humains de part et d'autre, l'apport de l'interdisciplinarité à la recherche est reconnu comme étant considérable. Le numérique ralliant les disciplines pourrait reporter à la conception antique de la connaissance comme savoir unique et indivisible.

ASPECTS JURIDIQUES

Si les aspects techniques et relationnels sont présents mais diffèrent beaucoup dans leur nature en raison des habitudes et de la particularité de chaque individu, les **aspects juridiques** contraignent la recherche de manière transversale à différents niveaux : du travail sur des corpus protégés en raison de l'application de plusieurs droits à l'obligation dans certains cas de l'anonymisation des données ; des citations aux publications. La BnF de son côté n'est pas immune de contraintes juridiques et c'est justement à cause de l'obligation de ne communiquer les corpus sous droit que dans ses emprises qu'est née l'idée d'un espace physique au sein de la Bibliothèque pour l'étude et l'analyse de corpus numériques.

3 Services similaires ?

À la question : avez-vous connaissance de services similaires en France ou à l'étranger ? **Les interlocuteurs hésitent ou repondent par la négative**. Un docteur en sciences de la communication et de l'information mentionne HathiTrust ; un enseignant-chercheur en sciences de l'information et de la communication suggère les Labs de la New York Public Library, bibliothèque publique mais pas nationale ; un agent de la BnF pointe la Bibliothèque du MIT.

Le manque de réponse s'explique probablement par le fait que des services similaires sont en train d'être imaginés par différentes structures mais ils n'ont pas encore vu le jour.

Dessiner un paysage qui est en cours de définition, à un stade d'avancement embryonnaire et *in progress*, est donc une tâche difficile à réaliser.

4 Quel rôle doit avoir la BnF dans l'écosystème de la recherche ? Comment voyez-vous la relation entre les laboratoires de recherche et la Bibliothèque ?

La BnF contribue à la recherche en remplissant avant tout un rôle de **fournisseur de données**. Selon un maître de conférences en linguistique informatique « elle doit donc faciliter l'accès aux données, la recherche des données, et d'autre part être à jour de ce qui se fait en recherche pour pouvoir intégrer des nouveaux outils et puis collaborer éventuellement à la création d'outils ».

Selon un enseignant-chercheur en sciences de l'information et de la communication, la BnF a un rôle à jouer aussi « à essayer de standardiser ou à proposer des chartes ou des méthodes [...]. Dans le contexte

du Web, il est nécessaire qu'en tant que bibliothèque publique elle donne accès via des plate-formes comme Github sans qu'une personne ou une organisation doive s'identifier, si les données sont libres de droit. [...] Mais à côté de ça, il y a une sorte de service *premium* à offrir à des chercheurs avec qui potentiellement il y a une relation de confiance : on voit l'intérêt de la BnF aussi d'avoir des chercheurs en résidence. C'est ainsi que la BnF peut apprendre de leurs méthodes ». En effet, selon un ingénieur de recherche en histoire de l'art, les premiers bénéficiaires d'un éventuel laboratoire ce seraient les agents mêmes de la BnF.

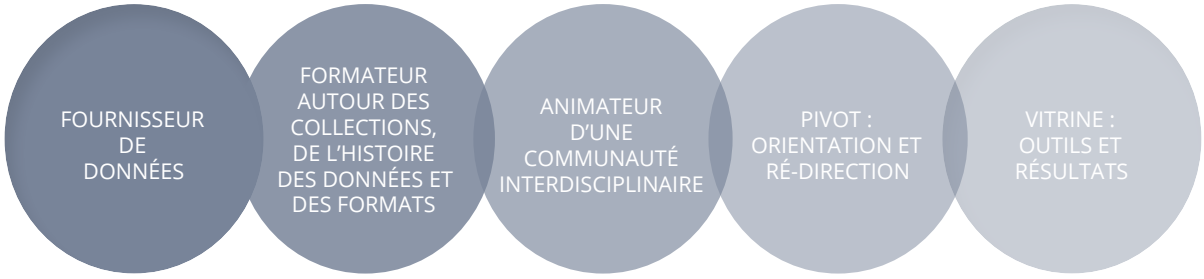
La Bibliothèque aurait un rôle à jouer aussi dans la **constitution d'une « communauté qui n'a pas d'existence naturelle dans les laboratoires** ». Une communauté qui serait également internationale grâce à la notoriété dont la BnF jouit à l'étranger.

De plus la Bibliothèque, aux yeux d'un directeur de Labex et professeur de littérature française, possède « un savoir-faire extraordinaire qui est une chance pour la recherche ». Ce savoir-faire, qui réside dans la pratique du catalogage, dans la normalisation et dans la connaissance des strates, de l'histoire des données, est en effet très apprécié par le milieu académique ; l'atelier « Décrire un corpus documentaire hétérogène » en a donné une évidente confirmation.

Selon la plupart des membres du personnel BnF sollicités, la Bibliothèque doit assurer son rôle de réservoir de données, apporter du conseil autour de l'utilisation des données et offrir des **formations** aux publics comme à ses agents. Elle devrait donc expliciter les conditions de consultation et de fouille et mettre en avant son expertise sur les données et sur les collections sans forcement donner des cours de méthodologie de fouille.

Quasiment tous les interlocuteurs, universitaires aussi bien qu'agents de la BnF et un expert français notamment, mettent en évidence l'avantage pour la BnF d'élaborer des formations sur les chaînes de traitement, sur l'organisation et la gestion des collections. La BnF doit toutefois éviter de se substituer aux formations sur les outils et les méthodes déjà nombreuses dans le milieu universitaire. À cause de l'importante pénurie de locaux dans les universités parisiennes, la BnF pourrait mettre ses espaces à disposition pour des formations données par d'autres entités et jouer un **rôle de pivot** en orientant, en redirigeant les chercheurs vers des séminaires et des enseignements existants par ailleurs comme ceux animés par le réseau Mate-SHS du CNRS par exemple. En d'autres mots, ceux d'un professeur en sciences du langage expérimenté : « On ne peut pas confiner des individus dans la production de la recherche, ça donne un compartimentage entre les gens qui fournissent les données, ceux qui les traitent et ça finit toujours mal. La BnF, il faut qu'elle mette à disposition ses données mais je pense qu'elle doit aussi s'approprier certains problèmes de la recherche pour qu'elle comprenne sans, et c'est un jeu d'équilibriste, dépasser son rôle. [...] À la fois le Laboratoire de la BnF c'est un laboratoire de recherche mais il ne se trompe pas sur son rôle dans la recherche, et donc à la fois il accompagne, il peut être promoteur d'idées, il peut être formateur mais ce n'est pas le laboratoire d'excellence en sociologie ou de traitement de fouille de textes et c'est cet équilibre-là qui est difficile à trouver ».

Une dernière mission mise en avant par un expert étranger consiste non pas dans le développement d'outils mais dans la **maintenance d'outils existants** que la BnF, en vertu de son rôle d'institution patrimoniale, pourrait remplir.



Dans le domaine de la conservation, une autre question se pose : la BnF doit-elle ingérer les résultats de la recherche ?

La plupart des universitaires interviewés considèrent que la conservation des résultats est en dehors du périmètre de la BnF. Cependant une petite partie des interlocuteurs voit l'intérêt réciproque de la conservation des résultats en tant qu'enrichissement ou correction des données existantes tels que l'in-

gestion d'un OCR corrigé, d'une édition numérique encodée en XML/ TEI, d'annotations d'images. Cela demanderait à la Bibliothèque non seulement l'établissement de nouvelles chaînes d'entrée, mais aussi une validation. Pour alléger le travail chronophage d'un processus de validation, un fonctionnement par calques type Shared Canvas[4] permettrait par exemple d'ingérer dans Gallica les données produites par les laboratoires de recherche ou d'autres institutions mais en les identifiant clairement comme produites à l'extérieur de l'établissement et en les proposant manifestement comme ajouts hors BnF.

Deux agents de la BnF soulignent alors la nécessité de prêter attention non seulement à la fourniture des données mais aussi à l'ingestion des données produites par la recherche. Dans cette perspective, il serait alors important d'élaborer de nouvelles chaînes d'entrée pour anticiper une demande de réintégration qui pourrait augmenter rapidement même si elle est marginale aujourd'hui.

5 Est-ce que vous venez à la BnF ? Si oui, pourquoi ? Si non, pourquoi ? Quelles sont les chances et limites d'un laboratoire à la BnF ? Qu'est-ce que vous gagneriez ? Qu'est-ce que vous perdriez ?

Au premier abord les universitaires, notamment ceux qui ont une formation approfondie dans une branche de l'informatique, perçoivent l'idée d'un **lieu physique** dans les emprises de la BnF comme une contrainte importante et un verrou considérable.

Un maître de conférences en linguistique informatique manifeste la position la plus radicale dans ce sens : « Maintenant j'ai tout ce qu'il me faut sur mon ordinateur et sur Internet ». Point de vue absolument partagé par un docteur spécialisé dans le traitement du signal : « Je n'ai jamais été client de la BnF. Les ressources sont en ligne. Un accès à distance est inévitable. De plus les gens qui s'intéressent à ces données sont soit étrangers soit potentiellement pas à Paris » .

En effet, les non-parisiens interviewés souhaitent disposer des données à distance à travers des modalités d'accès sécurisé. Et même les parisiens mettent en avant les contraintes d'un lieu physique : « Moi je suis parisien donc ça va bien mais je travaille avec des enseignants chercheurs qui sont à Aix-Marseille. [...] Donc ça c'est un verrou pour les gens avec lesquels je travaille. [...] Je prends l'exemple de la TGIR[5], tout se fait par mail, ils me donnent des codes et je ne les ai pas vus physiquement. Être sur place, forcement vous allez rajouter un verrou ».

La même vision est portée par une professeure spécialisée dans le traitement de l'image : « Un espace virtuel ce serait plus intéressant pour nous. Par exemple cela voudrait dire que vous avez une famille de données et vous posez un challenge. Plusieurs laboratoires peuvent s'entraîner là-dessus, ils vont pouvoir faire des choses avec ça et se comparer sur le même corpus. Pour les archives du Web il y a une chose qui est intéressante : trouver l'écrit dans ce qui est image car il y a beaucoup de texte caché dans le Web ».

Du point de vue d'une chargée de recherche en sciences de la communication travaillant sur les archives du Web justement « ce qui est fastidieux c'est d'avoir le corpus enfermé dans le lieu. » Mais elle comprend, comme tous, les raisons dues au cadre juridique.

Un autre frein dû à un espace physique est l'accessibilité. Un ingénieur de recherche craint non seulement des verrous informatiques mais aussi des problèmes d'accès dus à « la réputation de la BnF. Si je dois attendre trois ou quatre contrôles... ! Il faut que ce ne soit pas compliqué d'accéder à la salle ».

Bien qu'au premier abord le désir de disposer d'un **laboratoire virtuel** soit partagé par tous les universitaires, les attraits d'un lieu physique émergent dans un second temps.

Du point de vue d'une enseignante-chercheuse en sciences sociales, cet espace « pour les TD ce serait génial, moi je prends tout de suite. Il faut que tous les environnements soient installés (R, base de données, etc.) et qu'il y ait la possibilité de construire rapidement un corpus ».

4 Modèle de données sur lequel se base l'API Présentation de IIIF (International Image Interoperability Framework). Il permet la superposition de plusieurs calques dans le périmètre d'une même toile, autrement dit il permet d'empiler des strates distincts d'information sur une même image, Url : <http://iiif.io/model/shared-canvas/1.0>

5 Très Grande Infrastructure de Recherche

L'enseignement, le fait de « permettre à un professeur d'amener quelques étudiants de master, à un directeur de thèse de pouvoir y aller avec son doctorant » est reconnu par plusieurs interlocuteurs comme un grand atout. « Il faut avoir les deux : un laboratoire virtuel et un espace physique. Le premier permet de donner un accès direct ou ouvert aux données mais en parallèle dans le laboratoire physique je serais le premier à venir avec mes étudiants. C'est une excellente idée d'avoir un espace physique où on peut rencontrer la personne ».

Au désir d'échange s'ajoute l'intérêt de bénéficier de locaux dans un contexte de pénurie dans les universités parisiennes.

Un ingénieur de recherche rappelle que « les chercheurs parisiens n'ont quasiment plus de bureaux. Il y a nécessité qu'ils puissent travailler à un endroit. Moi je n'ai pas de lieu pour pouvoir me réunir avec mes chercheurs. Mais cela veut dire travailler ensemble donc cela veut dire des *open space* où on peut parler devant l'écran, on a un accès wifi, on peut sortir, on peut faire du FTP, SSH. Il faut prévoir un espace virtuel, on rentre chez soi et on peut télécharger. Sinon le risque est grand ».

Un autre ingénieur de recherche met en évidence l'isolement dû aussi à la pénurie de locaux : « La BnF peut créer un véritable dialogue entre les universités et les chercheurs qui maintenant sont isolés chacun dans son coin. Beaucoup de chercheurs en SHS sont seuls. Cet espace serait important pour eux ».

Pour le docteur spécialisé dans le traitement du signal, la rencontre constitue effectivement le seul attrait d'un lieu physique : « La seule vraie raison pour se déplacer aujourd'hui est d'aller voir l'humain, l'expert. On passe beaucoup de temps à se documenter alors qu'un expert pourrait nous diriger beaucoup plus rapidement ».

De manière similaire, un ingénieur de recherche aurait « intérêt à y aller au début d'un projet pour avoir un humain en face de moi. J'aimerais bien avoir en face de moi l'ingé qui a mis en place les manifestes JSON. Limite, d'avoir un service public d'ingénieurs comme il y a des services publics en salle avec des conservateurs. Parce que quoi qu'on dise, et je suis le premier à écrire la doc, personne ne lit la doc. Il faut qu'elle soit là mais la doc ne peut pas être la première étape. Il faut d'abord qu'il y ait une médiation d'un humain qui connaisse bien la doc, qui écoute la personne et qui dise " ah mais ça, c'est dans le sous-sous-menu de cette doc " ».

Le souhait d'un échange humain est également exprimé par un philologue : « Ce Laboratoire peut être intéressant parce qu'on n'est pas si nombreux que ça dans ce domaine, on ne se connaît pas forcement tous et ça peut être intéressant de croiser des gens ».

La professeure de reconnaissance de forme partage elle aussi ce besoin d'échange : « La raison pour se déplacer ça pourrait être qu'avec différents labos on a travaillé sur une base commune et on se retrouve pour présenter les résultats. C'est l'aspect communauté parce qu'on aurait travaillé sur la même thématique, on fait un rapport de nos méthodes et on peut les mettre en petit peu en commun ».

Et même pour un ingénieur de recherche basé sur Lille, la « raison pour se déplacer est l'humain, le fait de pouvoir échanger, trouver des solutions à mes problèmes et aux problèmes des autres. Apprendre des autres ».

« La seule vraie raison pour se déplacer aujourd'hui est d'aller voir l'humain, l'expert. On passe beaucoup de temps à se documenter alors qu'un expert pourrait nous diriger beaucoup plus rapidement »

En synthèse, qu'ils soient lecteurs BnF ou pas, qu'ils aient une aisance numérique ou pas, les universitaires préféreraient avoir un accès à un laboratoire virtuel mais le manque de locaux à Paris, l'envie de dispenser des enseignements sur des données conservées à la BnF et le besoin de dynamiques communautaires et interdisciplinaires représentent les éléments qui détermineraient le succès du laboratoire physique.

La phrase suivante prononcée par un professeur en sciences du langage d'expérience est tout à fait emblématique de cet idéal d'ouverture et de partage :

« Moi j'ai une expérience qui remonte à mes débuts d'enseignant-chercheur. On travaillait dans un espace ouvert parce qu'il se trouvait que le dispositif technique, l'ordinateur, y était physiquement. C'était l'époque des cartes perforées. Il fallait aller près de la machine, il fallait mettre les cartes dans le lecteur de

cartes et de ce fait il y avait un grand nombre de salles où les chercheurs de différents laboratoires venaient et comme vous attendez vos résultats après une heure, deux heures de traitement, vous ne partez pas et donc vous discutez avec les autres chercheurs et à cette époque-là, qui a duré entre 1976 et 1986, rétrospectivement je me dis que j'ai appris énormément de choses. Et donc si ce lieu à la BnF peut être aussi un lieu d'échange c'est sûrement très fructueux. [...] C'est intéressant de voir qu'après une époque de totale virtualisation et d'isolement, on arrive à discuter. Cela peut être un effet de bord mais qui, pensé, peut être très important dans l'apport de la BnF qui, en tant qu'entité qui n'est pas au cœur de la recherche mais qui est à la périphérie, peut servir de lieu de passage. C'est le rôle qu'elle n'a pas joué ces dernières années en étant comme une sorte de forteresse ».

Du côté des agents de la BnF, pour les conservateurs travaillant sur les collections, un laboratoire physique est perçu comme un moyen à inciter la fréquentation et la consultation des collections papier. Mise en avant seulement par une professeure en sciences de l'information et de la communication, la continuité entre le Laboratoire et les autres salles de lecture comme celle entre le numérique et le papier apparaît comme intéressante aux yeux des experts des collections qui voient un nouveau moyen de valoriser les fonds.

La plupart des membres du personnel sollicités voient la nécessité d'un espace physique aussi bien que d'un volet virtuel. L'espace physique donnerait une réponse à la difficulté actuelle d'accueillir des équipes de recherche au sein des bureaux et soulagerait en partie les départements d'une charge de travail trop importante. Le volet virtuel permettrait de réduire la fourniture de corpus et d'assurer l'accès à un plus grand nombre. Les deux possibilités comportent des inconvénients. Un lieu physique impliquerait l'allocation, voire le recrutement, de personnels permanents et représenterait une contrainte pour les usagers hors Paris notamment. Le laboratoire virtuel aurait l'avantage d'offrir une accessibilité plus grande aux données mais ne dispenserait pas la Bibliothèque de l'investissement en compétences humaines. Plusieurs agents mettent en avant le besoin de conseil même dans le cas d'ouverture des données. Selon l'expérience de la cheffe de produit data.bnf.fr : « On s'aperçoit de plus en plus que l'ouverture des données ne consiste pas juste à lâcher des données sur le Web. Même si vous postez un CSV, les données n'ont forcement pas été produites pour les chercheurs. Il y a des politiques et tout ça c'est lié à l'histoire de l'établissement et il faut la connaître pour comprendre les données. [...] Le fait qu'il y ait beaucoup de documentation en ligne par exemple ne suffit pas, il faut une personne. Il faut des ressources humaines et logistiques avec une organisation en interne pour mettre les gens en contact avec d'autres gens dans d'autres départements ».

Physique et virtuel, le Laboratoire d'étude et d'analyse de corpus numériques doit selon le personnel avoir des retombées internes. Il doit premièrement permettre aux agents d'apprendre des chercheurs et des collègues pour monter en compétences car les outils et les méthodes du TDM peuvent servir à une meilleure gestion des collections numériques ; il doit apporter des retours sur investissement sous forme d'outils d'exploration des fonds et d'enrichissements des données brutes de la BnF tels que la reconnaissance d'entités nommées, de forme, l'extraction de thème. Des enrichissements qui peuvent être utilisés dans les systèmes d'information de la Bibliothèque afin d'améliorer ses services aux usagers.

« On s'aperçoit de plus en plus que l'ouverture des données ne consiste pas juste à lâcher des données sur le Web. Même si vous postez un CSV, les données n'ont forcement pas été produites pour les chercheurs. [..] Il faut des ressources humaines et logistiques avec une organisation en interne pour mettre les gens en contact avec d'autres gens dans d'autres départements »

LABORATOIRE PHYSIQUE

-

1 - **UN SEUL LIEU**

2 - **ACCESSIBILITÉ**
queue
contrôles de sécurité
carte de recherche

3 - **DONNÉES RENFERMÉES**

4 - **HORAIRES**

+

1 - **TRAVAIL EN GROUPE**

2 - **DONNÉES SOUS DROIT**

3 - **CONVIVIALITÉ**

4 - **ENSEIGNEMENT**

5 - **COLLECTIONS**

6 - **LOCAUX**

LABORATOIRE VIRTUEL

-

1 - **DONNÉES SOUS DROIT**

2 - **PROXIMITÉ DES COLLECTIONS PAPIER**

3 - **CONVIVIALITÉ**

4 - **LOCAUX**

+

1 - **ACCESSIBLE PARTOUT**

2 - **ACCESSIBILITÉ**
nom d'utilisateur
mot de passe

3 - **DONNÉES OUVERTES**

4 - **DISPONIBILITÉ 24H/24**

5 - **TRAVAIL EN GROUPE À DISTANCE**

6 Quel type de lieu ? Quelle convivialité ? Quel agencement de l'espace ?

Tous les interlocuteurs s'accordent sur le type de lieu que le Laboratoire devrait être et le niveau de convivialité qu'il devrait y avoir.

Ni espace de détente ni salle de lecture traditionnelle, le Laboratoire devrait être convivial, compartimenté, modulable et évolutif.

Fréquenté par un public d'habitués qui vont constituer une communauté, ce lieu doit permettre de « rencontrer des gens, se donner rendez-vous, échanger autour d'un café ». Les boissons et la nourriture constituent un lien social qui n'est pas négligeable et qu'il ne faudrait pas sous-estimer. Quasiment tous les interlocuteurs mentionnent la nécessité du café au cours de l'entretien et plusieurs chercheurs rapportent le lieu commun : « On le sait tous, lors d'un colloque les moments les plus importants sont les pauses café, le déjeuner, le dîner. Si on veut catalyser les interactions entre les êtres humains ça passe toujours bien par un repas ou par un café ».

Comme le Laboratoire vise à inciter les échanges et le travail collectif mais également le travail individuel, l'espace devrait être compartimenté et modulable. Une idée proposée par un ingénieur de recherche est de créer des « zones qui seraient en fonction de l'avancement du projet, de la vie du projet de recherche». Mais la plupart des interlocuteurs modulent l'espace plutôt en fonction des activités. Une configuration répondant aux besoins des futurs utilisateurs comprendrait : des espaces pour le personnel, une zone de détente et de restauration, des espaces équipés d'écran pour le travail en groupe et les formations, des logettes individuelles bien isolées au niveau phonique, des bureaux pour les enseignants ou pour les chercheurs en résidence et une salle pour des événements capable d'accueillir une soixantaine de personnes. Malgré cette configuration assez définie et claire, plusieurs interlocuteurs mettent en garde sur sa stabilité et immuabilité. L'innovation ne devrait pas seulement passer par les services mais aussi par l'agencement de l'espace qui devrait être idéalement capable d'évoluer en harmonie au rythme de la technologie et de la recherche.

Convivial, compartimenté, modulable et évolutif

Un dernier élément pour donner de la vie à ce lieu vient d'une enseignante-chercheuse en sciences sociales : « Ce serait une super idée d'ouvrir aux artistes. Ce serait très enrichissant. Ils ont une façon de parler du numérique absolument géniale. Ils auraient plein d'idées mais il faudrait leur laisser de la liberté. Ce serait vraiment chouette de pouvoir faire des posters, des productions visuelles à exposer dans le Laboratoire. Vu que les productions ne peuvent pas sortir de la BnF qu'elles soient valorisées dans cet espace, que le Laboratoire s'enrichisse de productions créatives. On est dans un monde fermé mais on peut tout montrer à l'intérieur. Dans l'enceinte on peut tout faire et on peut montrer ce qu'on fait. Cela donne des idées, c'est comme ça qu'on crée une communauté ».

7 Quelle infrastructure ? Quels outils ? Quels défis ?

Au mot « infrastructure » les universitaires réagissent quasiment tous de la même manière, avec des phrases comme :

- « Il faudrait pouvoir travailler sur l'ordinateur personnel et il faudrait qu'il y ait une communication entre les deux. Les environnements sont quand même très personnalisés ».
- « On doit pouvoir travailler sur nos ordinateurs portables ».
- « Ne pas pouvoir installer tel truc alors que je découvre que j'en ai besoin ? Non, il faut que j'ai la main complète sur la machine ».
- « Hors de question de prendre un ordinateur d'ici ou de n'importe où. Sur mon portable j'ai toutes mes archives, mes manips. Puis les ordinateurs c'est une telle galère à maintenir, à mettre à jour, à nettoyer ».
- « Je ne travaille que sur mon portable. Tout mon travail est complètement centralisé sur un seul poste. C'est insupportable pour moi de devoir réinstaller le logiciel que j'ai déjà installé ailleurs ».

Une seule voix est nettement dissonante : « Moi je ne fais aucun traitement sur mon ordinateur personnel, il me sert uniquement pour me connecter soit sur un serveur qui héberge mes données soit depuis la TGIR. Ce qui n'était pas le cas il y a cinq ou six ans ».

La nette préférence des chercheurs à travailler sur leur propre machine ne fait que renforcer l'idée d'un espace à la fois physique en même temps virtuel, un *unicum* sans césure évidente.

Les universitaires comprennent pourtant les contraintes imposées par le cadre juridique et ils demandent alors qu'il y ait une continuité entre l'environnement proposé par la BnF et l'environnement personnel en permettant l'export de données dérivées.

Un expert français, directeur technique de la TGIR Huma-Num[6], partage la solution pour laquelle la TGIR a opté : quarante machines virtuelles et un énorme *firewall*. À ses yeux : « C'est cher, mais c'est le prix de la liberté ». Cette liberté semble être bien appréciée par les utilisateurs de la TGIR. Un professeur en sciences du langage fait part de son expérience : « J'ai lancé un processus sur la machine de la TGIR vendredi et il continue à tourner…J'ai calculé il va tourner une semaine mon script et la TGIR m'a mis à disposition une machine à 8 cœurs, donc il y a 4 processeurs qui tournent en même temps pendant une semaine».

La puissance de calcul est un élément qui préoccupe les agents de la BnF mais qui n'est pas prioritaire pour la plupart des universitaires. Selon un directeur de Labex et professeur de langue française la puissance de calcul pourrait être facilement externalisée. Selon un maître de conférences en linguistique informatique le calcul peut tourner ailleurs : « Le problème est que moi les trucs que j'aurais à faire pour traiter du texte ça va mettre 10 heures quoi. Je ne vais pas attendre la machine, je lance et je m'en vais quoi. Maintenant existent des interfaces web où on lance un processus sur telle machine je ne sais pas où sans être sur place en fait ».

Autour des outils, les universitaires répondent avec une grande réticence à l'idée d'une boîte à outils standardisée : « Les choses évoluent tellement rapidement… Je pense ce serait une erreur de vouloir proposer une boîte à outils standardisée. L'essentiel c'est d'investir dans des compétences humaines. Si on force des chercheurs à se limiter ça ne marche pas. Pour un chercheur se limiter à des outils, à une place, à une infrastructure c'est un frein énorme ».

Cependant, une fois interrogés sur les outils qu'ils utilisent, les logiciels et langages qu'ils mentionnent sont assez récurrents.

6 Très Grande Infrastructure de Recherche Huma-Num, Url : <https://www.huma-num.fr>

Un environnement de développement pour le langage informatique **R** comme R Studio :	https://www.r-project.org/about.html https://www.rstudio.com
Le langage de programmation **Python** et un éditeur comme PyCharm :	https://www.python.org https://www.jetbrains.com/pycharm
Iramuteq, logiciel libre, interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires :	http://www.iramuteq.org
TXM, plate-forme libre pour la textométrie :	http://textometrie.ens-lyon.fr/?lang=en
Mallet, logiciel libre programmé en Java pour l'analyse computationelle de textes :	http://mallet.cs.umass.edu
Voyant tools, application web libre pour l'analyse computationnelle de textes :	https://voyant-tools.org
Gephi, logiciel libre pour la visualisation de données :	https://gephi.org
Un **éditeur de code** comme Atom ou Oxygen :	https://atom.io https://www.oxygenxml.com
Bibliothèques de modèles pré-entrainés de **réseaux de neurones** (pour la reconnaissance de forme, de l'écriture manuscrite, etc.)	
QGIS, système libre pour la gestion de l'information géographique :	https://www.qgis.org/en/site
Suite Elastic, suite logicielle libre pour le traitement de données	https://www.elastic.co/fr
Un **environnement** de type LAMP, WAMP, MAMP, XAMPP pour la création et la gestion de **bases de données** :	https://doc.ubuntu-fr.org/lamp http://www.wampserver.com https://www.mamp.info https://www.apachefriends.org

Même si une boite à outils ne sera jamais exhaustive et complètement adaptée, un cœur, une palette d'outils pourrait être définie et mise à disposition.

En résumé, la solution idéale qui ressort de l'enquête serait constituée par un système comprenant des machines virtuelles et une **plate-forme sécurisée** comme par exemple TeraLab[7], accessible aussi bien dans les emprises de la Bibliothèque que depuis l'extérieur. Au niveau des fonctionnalités de la plate-forme, elle devrait essentiellement contenir une palette d'outils mais aussi pouvoir gérer des **niveaux d'habilitations** permettant l'installation de logiciels, elle devrait en fournir la documentation, des **tutoriels** sous forme de MOOC, des **exemples d'usages**, l'accès via des connecteurs à des **corpus pré-constitués** (API et jeux de données, Archives de l'Internet Labs) et notamment la possibilité d'effectuer une demande en ligne de numérisation et d'océrisation d'un corpus ou collecte Web, de réservation d'un espace dans le laboratoire physique et de prise de rendez-vous avec un membre du personnel de la Bibliothèque.

7 TeraLab, Url : <https://teralab-datascience.fr>

8 Quel type d'accompagnement ? Quels profils ? Quelles compétences ?

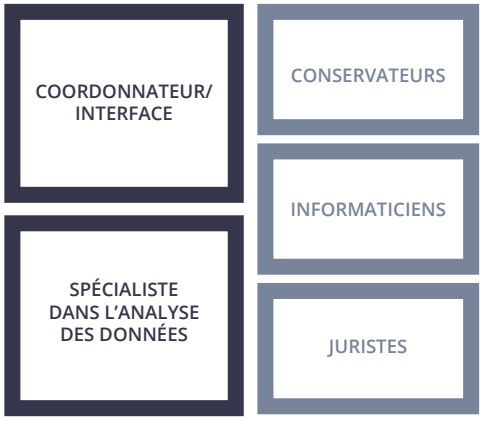
L'échange, le conseil, l'accompagnement constituent une plus-value capable d'attirer les universitaires sur place. L'accompagnement souhaité par les universitaires demanderait idéalement cinq figures autour de cinq pôles d'expertise : le conservateur pour son **expertise sur les fonds**, le juriste pour le **conseil juridique**, l'ingénieur pour le **soutien purement informatique**, le spécialiste dans l'**analyse des données** ayant une formation hybride (SHS-informatique et science des données) et le responsable de la communication/ coordonnateur qui aurait un rôle d'**interface**, de pivot en étant capable d'orienter les utilisateurs et de frapper à la bonne porte dans la Bibliothèque.

Pour une enseignante-chercheuse en sciences sociales : « La collaboration entre le conservateur et le chercheur en sciences humaines et sociales est essentielle. Il faudrait pouvoir travailler en triptyque : le chercheur ou l'équipe de recherche externe, le conservateur et le soutien informatique (qui n'est pas le point fort de la BnF) ».

Aux yeux d'une professeure en sciences de l'information et de la communication : « La formation, je n'y crois pas trop mais au déblocage, à la sensibilisation, à l'indication d'outils oui. [...] Le dialogue sur la numérisation est important ».

Selon des ingénieurs de recherche, il faudrait « un service de développeurs en banc de salle » mais « par contre un chercheur qui ne serait pas accompagné par un profil un peu plus interface comme moi il lui faudrait une personne d'interface ».

Les agents de la BnF, comme les universitaires, comme les experts s'accordent tous sur le fait que, même si les compétences internes peuvent déjà être facilement mises à disposition du public, « il faut vraiment investir dans le personnel » pour que le Laboratoire puisse fonctionner. En effet, l'orientation démontre que même derrière les laboratoires virtuels travaillent des équipes de quatre personnes minimum. Si le conservateur, le juriste et le technicien peuvent être repérés au sein de l'Établissement, présents de façon semi-permanente et tourner, les figures du spécialiste dans l'analyse des données et du coordonnateur devraient être les piliers du Laboratoire et assurer une présence permanente pour pouvoir suivre les aspects logistiques (réservation d'espaces, organisation d'événements et formations), la communication et les partenariats, les aspects techniques (signalement de pannes, gestion de la maintenance), l'accueil et l'orientation des usagers.



LES APPORTS DE LA MÉTHODE DES *PERSONAS*

La méthode des *personas* a aidé à l'identification et à la définition des profils des usagers potentiels. La phase notamment de l'analyse constituée par la disposition des interviewés le long des axes (voir annexe D) et le repérage de regroupements de comportements, objectifs et caractéristiques similaires a fourni un cadre utile à la fidélité aux propos.

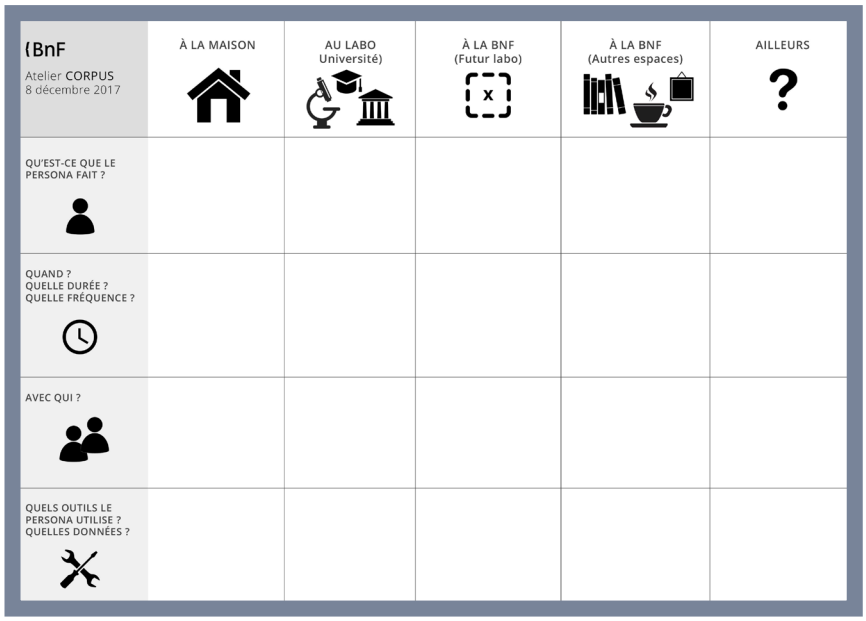
L'identification de regroupements représentait le risque de la standardisation et de l'établissement de cases trop rigides. Le travail en groupe de huit agents de la BnF réunis autour des échelles a limité ce risque en établissant un équilibre entre la complexité, la variété des interlocuteurs et la représentativité, l'exemplarité des *personas*.

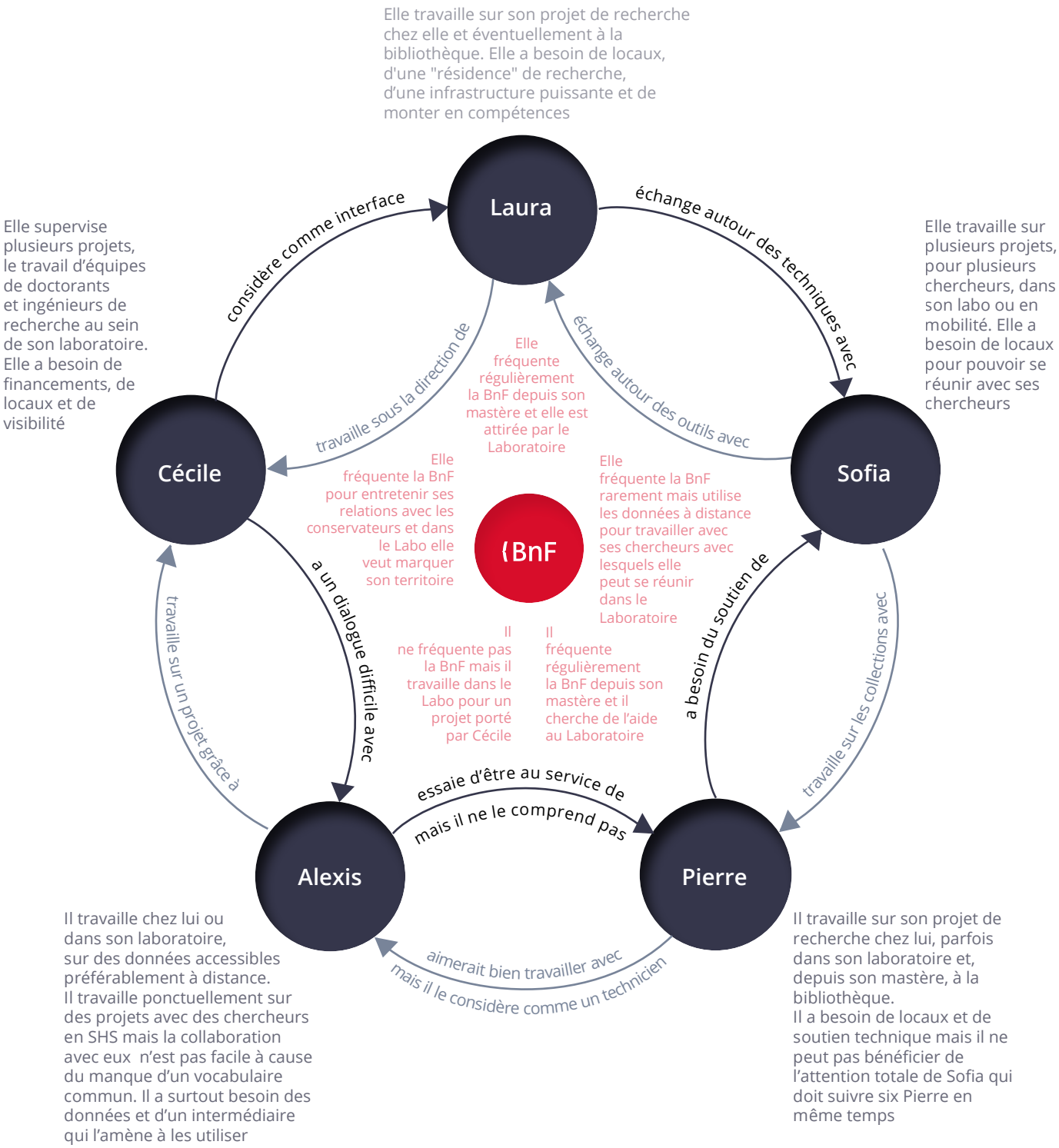
L'utilisation des *personas* lors de l'atelier collaboratif a apporté la définition des relations entre les *personas* et des *personas* avec la BnF (voir le schéma à la page suivante), des confirmations ainsi que quelques interrogations.

Représentatifs et exemplaires, les profils des *personas* élaborés et utilisés étaient constitués par des éléments caricaturaux qui ont contribué à l'appropriation et au développement de l'empathie. Les six universitaires invités ne se sont pas forcement identifiés dans un profil, mais ils ont retrouvé les cinq profils dans leur entourage professionnel, ce qui démontre la plausibilité des *personas* qui, sans être authentiques, sont tout de même vraisemblables.

L'appropriation du profil a permis aux participants de remplir le tableau fourni (voir figure suivante) sans difficulté. Les informations qui ressortent de ces tableaux constituent des confirmations par rapport aux résultats de l'étude mais alertent en offrant un aperçu du hors champ. En effet, le tableau élaboré proposait en abscisse les lieux fréquentés par le *persona* et en ordonnée la convivialité, le temps et les outils utilisés. L'objectif était donc de tracer où, quand et quelles activités le *persona* effectue afin d'avoir une idée de sa routine et de la place que le Laboratoire d'étude et d'analyse de corpus numériques pourrait avoir dans son travail de recherche. Les tableaux remplis contiennent les résultats de l'étude mais posent des interrogations sur les contraintes de lieu et de temps. Dans des agendas aussi remplis, dans un contexte d'un milieu de la recherche de plus en plus tourné vers la nouveauté, la rapidité, la mobilité et la performance, y aura-t-il le temps de se déplacer à la BnF ? Dans cette dimension contrainte par les limites spatio-temporelles y aura-t-il de la place pour l'échange tant demandé lors des entretiens ?
La pénurie de locaux dans les universités parisiennes peut constituer un élément qui contribuerait à une réponse positive à la première interrogation. Les deux autres ateliers organisés en 2017 dans le cadre du projet Corpus contribuent à donner une réponse positive à la deuxième question en raison du fort intérêt, de l'affluence considérable (jusqu'à 60 personnes), de la participation active et de l'appréciation constatée à ces deux occasions.

Tableau fourni aux cinq groupes




Elle travaille sur son projet de recherche chez elle et éventuellement à la bibliothèque. Elle a besoin de locaux, d'une "résidence" de recherche, d'une infrastructure puissante et de monter en compétences
Laura
considère comme interface
échange autour des techniques avec
Elle supervise plusieurs projets, le travail d'équipes de doctorants et ingénieurs de recherche au sein de son laboratoire. Elle a besoin de financements, de locaux et de visibilité
Elle travaille sur plusieurs projets, pour plusieurs chercheurs, dans son labo ou en mobilité. Elle a besoin de locaux pour pouvoir se réunir avec ses chercheurs
travaille sous la direction de
échange autour des outils avec
Elle fréquente régulièrement la BnF depuis son mastère et elle est attirée par le Laboratoire
Cécile
Sofia
Elle fréquente la BnF pour entretenir ses relations avec les conservateurs et dans le Labo elle veut marquer son territoire
Elle fréquente la BnF rarement mais utilise les données à distance pour travailler avec ses chercheurs avec lesquels elle peut se réunir dans le Laboratoire
BnF
a un dialogue difficile avec
a besoin du soutien de
travaille sur un projet grâce à
travaille sur les collections avec
Il ne fréquente pas la BnF mais il travaille dans le Labo pour un projet porté par Cécile
Il fréquente régulièrement la BnF depuis son mastère et il cherche de l'aide au Laboratoire
essaie d'être au service de
mais il ne le comprend pas
Alexis
Pierre
aimerait bien travailler avec
mais il le considère comme un technicien
Il travaille chez lui ou dans son laboratoire, sur des données accessibles préférablement à distance. Il travaille ponctuellement sur des projets avec des chercheurs en SHS mais la collaboration avec eux n'est pas facile à cause du manque d'un vocabulaire commun. Il a surtout besoin des données et d'un intermédiaire qui l'amène à les utiliser
Il travaille sur son projet de recherche chez lui, parfois dans son laboratoire et, depuis son mastère, à la bibliothèque. Il a besoin de locaux et de soutien technique mais il ne peut pas bénéficier de l'attention totale de Sofia qui doit suivre six Pierre en même temps

3. CONCLUSIONS

LE «PORTRAIT-ROBOT» DU LABORATOIRE

Dans son contrat de performance, la BnF s'engage à « offrir aux chercheurs, dans les emprises de la Bibliothèque, des outils de fouille et d'exploration de textes et de données sur des corpus numériques de la BnF.[1]»

Le programme de recherche Corpus d'une durée de quatre ans (2016-2019) a pour objectif de définir, en se basant sur des études de cas concrets, l'organisation cible d'un service destiné à permettre à un public de niveau recherche d'accéder à des corpus numériques, notamment à des fins de fouille de texte et de données (TDM – *Text and Data Mining*).

Les premières années du programme ont permis d'esquisser la nature du besoin et une première préfiguration du service à construire. La présente étude de besoins a apporté d'autres éléments qui permettent d'enrichir le dessin d'un « portrait-robot » du futur Laboratoire d'étude et d'analyse de corpus numériques proposé ci-dessous.

UN LIEU PHYSIQUE DANS LES EMPRISES DE LA BNF SUR LE SITE FRANÇOIS MITTERRAND ...

Un espace physique est une condition *sine qua non* de la mise à disposition de collections sous droit.

L'image innovante que le site François Mitterrand renvoie en fait le site le plus adapté à abriter un laboratoire tourné vers le numérique et les opportunités inédites de découverte qu'il offre.

Accessible depuis les espaces publics grâce à une carte de recherche, ce lieu s'annonce convivial, compartimenté, modulable et évolutif. La direction des Collections a proposé le réaménagement de la salle X pour abriter ce nouveau service en raison du faible taux de fréquentation de cette salle destinée à l'heure actuelle à la Recherche bibliographique. Des convergences avec le projet de réaménagement de la mezzanine de la salle P du département de l'Audiovisuel ont été également constatées.

L'implémentation du service dans une salle de recherche doit se faire de manière progressive. Cependant une première sectorisation de l'espace en fonction des usages est envisageable :

- salle pour les événements (ateliers, conférences, journées d'études, etc., ...) capable d'accueillir au moins cinquante personnes ;
- bureaux du personnel et pour accueillir des chercheurs en résidence ainsi que des professeurs souhaitant recevoir leurs étudiants ou doctorants ;
- loges bien isolées phoniquement pour le travail individuel et les vidéo-conférences ;
- salles équipés d'écran ou vidéoprojecteur pour le travail en groupe et des formations;
- zone de détente et de restauration à proximité de la zone pour l'événementiel ;
- zone de présentation pour valoriser l'avancement des projets ;
- espaces pour assurer l'accueil et l'accompagnement.

La possibilité de réaménager cette configuration en mettant en place des cloisons amovibles et en optant pour du mobilier modulable serait une vraie valeur ajoutée.

Le « service public » dans cet espace pourrait être assuré en partie par les agents de la BnF ayant une familiarité avec les problématiques liées à la gestion et à la fouille de données. Des experts de l'OCR, de la géolocalisation, des formats utilisés notamment dans le monde des bibliothèques, des archives de l'Internet, de la reconnaissance de forme, du Web sémantique représenteraient une ressource importante pour les usagers potentiels. Les conservateurs en raison de leur expertise sur les fonds constitueraient une ressource autant importante qu'attendue. En effet, les résultats de la présente étude de besoins suggèrent la nécessité d'au moins cinq champs d'expertise : la connaissance des collections, le conseil juridique, le soutien informatique, l'expérience dans l'analyse de données et la coordination de la communication autour des évènements, de l'orientation des usagers. Le personnel travaillant pour ce Laboratoire serait donc constitué par des profils variés : conservateur, juriste, informaticien, spécialiste dans l'analyse de données mais ayant une formation en SHS et un profil à la formation également hybride capable de faire dialoguer les différentes compétences.

1 Contrat d'objectifs et de performance 2017-2021

L'agencement, l'accompagnement sont essentiels mais pour que ce lieu prenne réellement forme il faut également y insuffler de la vie.

La convivialité est un facteur nécessaire au succès de cet espace. La possibilité d'échanger, de parler à voix haute, d'y boire des boissons, du café notamment, est fortement souhaitée par les usagers potentiels. Des horaires d'accès étendus sont également souhaitables mais pas prioritaires.

OÙ SERONT MISES A DISPOSITION LES COLLECTIONS NUMÉRIQUES DE LA BNF ...

Les collections issues du dépôt légal numérique, les archives de l'Internet, les documents numérisés et les métadonnées doivent rester, en parfaite continuité avec les missions de la Bibliothèque, au cœur du Laboratoire. Ces collections suscitent premièrement un grand intérêt pour la recherche en sciences humaines et sociales mais elles pourraient attirer aussi l'attention de la recherche dans certaines branche de l'informatique en raison de leur caractère varié et massif qui les rend intéressantes pour le développement et la mise au point d'algorithmes et solutions logicielles.

La mise à disposition des données ne dispense pas de l'accompagnement et du conseil autour des collections au niveau scientifique, juridique et technique mais aussi au niveau institutionnel-documentaire, c'est-à-dire le partage de la connaissance des politiques de l'Établissement quant aux aspects documentaires et de la connaissance des strates de l'histoire des données.

Un accompagnement continu est préféré par les utilisateurs potentiels à un accompagnement ponctuel ou de type formation. Au moins deux membres du personnel devraient donc assurer une présence permanente.

Les moyens à mobiliser et les modalités de cet accompagnement seront à définir.

La question de la continuité entre le Laboratoire et les autres salles de lecture ainsi que celle de l'accès aux documents papier restent ouvertes.

Une articulation fluide avec les autres services de la Bibliothèque autour des collections numériques (Gallica Studio, Chaire Bibli-Lab, Hackathon, API et jeux de données, Archives de l'Internet Labs, etc.) reste également à définir.

À DES FINS D'EXPLORATION DE CES DONNÉES ...

La mise à disposition des données ne vise pas seulement à la consultation mais en priorité à l'analyse et au traitement de corpus numériques constitués de données « brutes » ou bien pré-travaillées. Une plateforme sécurisée devrait donc être mise à disposition et, outre l'accès via des connecteurs à des corpus pré-constitués (API et jeux de données, Archives de l'Internet Labs), elle devrait inclure des fonctionnalités d'extraction de données, de constitution de corpus ou sous-corpus, de demande de collecte web et de numérisation d'un corpus ou documents.

La TGIR Huma-Num constitue un partenaire privilégié en raison de son expérience dans le stockage, la gestion et le traitement des données de la recherche, mais aussi en raison de l'infrastructure mise en place qui comprend plusieurs machines virtuelles, un *firewall* imposant et la possibilité de demander l'installation de logiciels à distance.

La question des moyens à mobiliser et de l'identification de partenaires reste à explorer.

VIA UNE INFRASTRUCTURE NUMÉRIQUE ...

L'analyse et le traitement de masses de données demandent une puissance de calcul relativement importante qui pourrait être externalisée au centre de calcul de Paris-Sorbonne par exemple. Des interfaces web permettent le lancement de processus à distance mais la question de la notion d'emprises reste problématique. En étendant la notion d'emprises physiques à la notion d'emprises numériques, une infrastructure de type *cloud* comme Teralab - de l'Institut Mines Télécom - répondrait aux besoins des usagers potentiels et assurerait la continuité entre le laboratoire physique et virtuel et donc la continuité du travail pour les utilisateurs. Une telle infrastructure devrait comprendre idéalement plusieurs fonctionnalités pour l'usager :

- Espace personnel avec un tableau de bord ;
- Accès sécurisé aux collections et aux corpus pré-constitués ;

- Tutoriels et exemples d'usages qui pourraient être réalisés par d'autres établissements et mutualisés ;
- Possibilité d'extraction de données, de constitution et sauvegarde de corpus ou sous-corpus ;
- Espace de stockage ;
- Demande de collecte et de numérisation de corpus ou documents ;
- Palette des outils les plus utilisés ;
- Demande d'installation d'outils ;
- Possibilité d'exporter ou partager des données ou des résultats avec d'autres usagers ;
- FAQs ;
- Forum (moyen d'alléger la charge de travail pour le personnel et de souder la communauté) ;
- Notification par courriel électronique de la fin d'un processus de plusieurs heures ou jours;
- Possibilité de réserver une place dans le laboratoire ;
- Possibilité de contacter le personnel, de faire une réclamation et de prendre un rendez-vous.

À la BnF l'infrastructure devrait permettre essentiellement :
- Gestion de différents niveaux d'habilitations en fonction des données ;
- Sécurisation des données notamment sous droit ;
- Traçabilité des opérations effectuées par les usagers ;
- Validation des exports ;
- Administration des installations et blocage de comptes ou opérations suspectes.

Une infrastructure de ce type ne fonctionne pas sans le travail et les compétences humaines dans lesquelles la BnF devra dans tous les cas investir pour la gestion du Laboratoire.

EN PARTENARIAT AVEC DES STRUCTURES DE RECHERCHE.

L'appui d'universités ou de laboratoires de recherche sera indispensable pour attirer les chercheurs ou les étudiants à travailler sur place. Plusieurs acteurs de la recherche ont déjà exprimé leur intérêt. Ce serait en outre un atout que d'en faire un lieu de formation en mettant à disposition les locaux pour des formations dispensées par d'autres structures et en fournissant des formations BnF autour des collections, de la constitution des fonds, du classement, du catalogage, des formats, des silos de données, en un mot autour du savoir-faire extraordinaire de la Bibliothèque qui suscite un grand intérêt auprès du milieu académique.

La forme juridique du Laboratoire et sa gouvernance restent encore à imaginer. Des liens avec les instances nationales de la recherche pourraient notamment être un moyen de financer les emplois et éventuellement des projets qui pourraient avoir des retombées, des retours sur investissement pour la Bibliothèque.

L'apport de ressources humaines ne servirait pas seulement à la gestion de l'infrastructure et de l'espace mais aussi au déploiement d'une politique de communication via des outils tels qu'un portail, un blogue et un compte Twitter. À travers ces canaux pourrait passer les informations relatives à la valorisation scientifique des travaux réalisés au sein du Laboratoire constituée par des journées d'études, des conférences, des ateliers, des publications.

Enfin, au niveau de l'accessibilité, la BnF doit garantir la non-exclusivité en ouvrant le Laboratoire à tout chercheur ou équipe de recherche mais à condition que le projet de recherche soit agrée et financé. Une ouverture à des groupes non-académiques et notamment à des artistes serait très fructueuse mais probablement à évaluer dans un deuxième temps en fonction du taux d'affluence et de la charge de travail du personnel.

Dans un contexte où toutes les bibliothèques réflechissent à une nouvelle offre de services aux chercheurs, Le Laboratoire préfiguré ci-dessus représente en conclusion une opportunité pour la BnF de ne pas se limiter à remplir sa mission de fournisseur de données mais de jouer un véritable rôle dans l'écosystème de la recherche en stimulant la création d'une communauté qui n'a pas d'existence naturelle dans le milieu académique, en la sensibilisant à son savoir-faire, en suscitant de nouveaux usages et découvertes, bénéfiques aussi bien pour l'univers de la recherche que pour le monde des bibliothèques.

Phases de l'étude

Besoins identifiés

ENTRETIENS

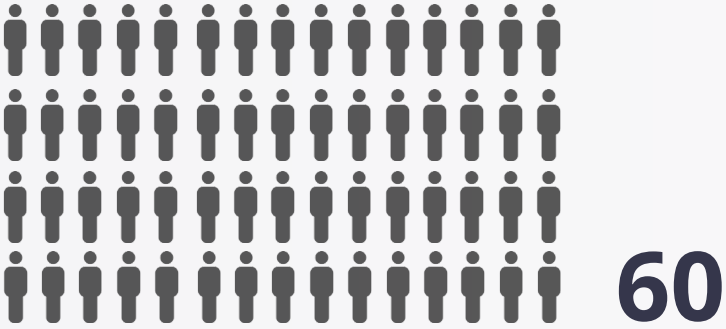
- accès facile au lieu physique
- accès aux données aussi à distance
- accompagnement autour des collections, juridique et technique
- convivialité, partage, inspiration mutuelle, interdisciplinarité
- ordinateur personnel, puissance de calcul et palette d'outils
- développement et installation d'outils
- locaux et programmes en résidence
- formations et événements

PERSONAS

ATELIER COLLABORATIF

- mobilité et contrôle du flux de travail
- travail individuel et collectif
- numérisation
- visibilité
- formation
- locaux

2 ATELIERS



- formation mutuelle
- partage et débat
- interdisciplinarité
- standardisation
- convergences


Espace pour événements
pour satisfaire le besoin de
PARTAGE
FORMATION MUTUELLE
DÉBAT
INTERDISCIPLINARITÉ
STANDARDISATION
CONVERGENCE
Bureaux
pour satisfaire le besoin de
PROGRAMMES EN RÉSIDENCE
LOCAUX
VISIBILITÉ
Espace de présentation
pour satisfaire le besoin de
PARTAGE
FORMATION MUTUELLE
INSPIRATION
Loges
pour satisfaire le besoin de
TRAVAIL INDIVIDUEL
LOCAUX
Collections
Numérisées ou nativement numériques
Salles de groupes
pour satisfaire le besoin de
TRAVAIL COLLECTIF
LOCAUX
FORMATION
Espaces du personnel
pour satisfaire le besoin de
ACCOMPAGNEMENT PERMANENT
SERVICE PUBLIC
RENDEZ-VOUS
Zone détente
pour satisfaire le besoin de
CONVIVIALITÉ
ÉCHANGE
Infrastructure numérique
pour satisfaire le besoin de
ACCÈS À DISTANCE
MOBILITÉ
CONTRÔLE FLUX DE TRAVAIL
ORDINATEUR PERSONNEL

ANNEXES

LISTE DES THÈMES ABORDES AVEC LES UNIVERSITAIRES

1. Définition de fouille de données

1.1 Comment définiriez-vous la fouille de données ?

1.2 Quelle importance la fouille de données a-t-elle dans la recherche ?

1.3 Est-elle appelée à se développer ? À quelle vitesse ?

1.4 Quelle évolution dans votre discipline ?

1.5 Le TDM constituera-t-il l'approche exclusive ou les pratiques traditionnelles perdureront-elles ?

2. Expériences d'utilisation de corpus numériques

2.1 Avez-vous déjà exploité des corpus de données fournies par une institution ?

2.2 Si oui, pouvez-vous me décrire brièvement votre expérience ?

2.3 Quelles collections de la BnF connaissez-vous ?

2.4 Quel type de fouille avez-vous effectué ? Quels en étaient les objectifs ?

2.5 Quel a été le produit de votre utilisation ? Quel résultat avez-vous obtenu ?

2.6 La fouille a-t-elle conduit à la réalisation d'une production savante ?

3. Attentes

3.1 Est-ce que vous venez à la BnF ?

3.2 Si oui, pourquoi ? Sinon, pourquoi ?

3.3 Comment voyez-vous la relation entre un laboratoire de recherche et une bibliothèque pour la fouille de données ? Quel rôle doit avoir la Bibliothèque ?

3.4 Qu'est-ce que vous gagneriez dans ce Laboratoire ? Qu'est-ce que vous perdriez ?

4. Infrastructure et outils

4.1 Quels outils ou logiciels avez-vous utilisé précédemment ?

4.2 De quels outils et infrastructure auriez-vous besoin pour faire de la fouille de données ?

5. Services similaires

5.1 Êtes-vous au courant d'expériences similaires en France ou à l'étranger ?

LISTE DES THÈMES ABORDES AVEC LES EXPERTS

1. Définition de fouille de données

1.1 Comment définiriez-vous la fouille de données ?

1.2 Quelle importance la fouille de données a-t-elle dans la recherche ?

1.3 Est-elle appelée à se développer ? À quelle vitesse ?

1.4 Quelle évolution dans votre discipline ?

1.5 Le TDM constituera-t-il l'approche exclusive ou les pratiques traditionnelles perdureront-elles ?

2. Expériences de fourniture de corpus numériques

2.1 Avez-vous déjà fourni des corpus ?

2.2 Si oui, pouvez-vous me décrire brièvement votre expérience ?

2.3 Quels sont les problèmes que vous avez rencontrés sur votre chemin ?

2.4 Avez-vous pu obtenir des retours de la part des usagers ?

2.5 Qu'est-ce que vous auriez pu améliorer, a posteriori ? C'est-à-dire, si vous pouviez refaire cette expérience, qu'est-ce que vous changeriez ?

3. Conseils

3.1 Quel rôle doit avoir la Bibliothèque dans la fouille de données ?

3.2 Comment voyez-vous la relation entre laboratoire de recherche et bibliothèque ?

3.3 Considérez-vous intéressante l'idée d'un espace physique consacré à cette activité ?

3.4 Quelles sont les chances et limites d'un espace dédié à la BnF ?

3.5 Quelles modalités d'accompagnement envisageriez-vous ?

3.6 Quelles sont à votre avis les raisons qui freineraient les chercheurs à utiliser un espace consacré à la fouille de masses de données dans une bibliothèque nationale ?

4. Infrastructure et outils

4.1 De quels outils et infrastructure les chercheurs auraient-t-ils besoin pour faire de la fouille de données ?

5. Services similaires

5.1 Êtes-vous au courant d'expériences similaires en France ou à l'étranger ?

LISTE ANONYMISÉE DES PERSONNES RENCONTRÉES

1. Docteur spécialisé dans le traitement du signal
2. Ingénieure
3. Enseignante-chercheuse en sciences sociales
4. Ingénieur de recherche SHS
5. Docteur en sciences de l'information et de la communication
6. Directeur de Labex et professeur de littérature française
7. Professeure en sciences de l'information et de la communication
8. Maître de conférences en linguistique informatique
9. Philologue
10. Maître de conférences en sciences sociales
11. Ingénieur de recherche SHS - littérature
12. Professeure spécialisée dans le traitement de l'image
13. Chargée de recherche en sciences de la communication
14. Ingénieur de recherche SHS - histoire de l'art
15. Professeur en sciences du langage
16. Enseignant-chercheur en sciences de l'information et de la communication

1. Expert - chef de projet OCR et formats éditoriaux
2. Cheffe de service du Dépôt Légal Web
3. Adjointe au directeur du département des Cartes et Plans
4. Adjointe à la cheffe du service Diffusion des métadonnées et responsable de l'équipe Services aux professionnel
5. Coordonnateur de la recherche
6. Adjoint au directeur du département des Systèmes d'information
7. Adjoint au chef du service Ingénierie des métadonnées et responsable de l'équipe Analyse et traitement de données
8. Cheffe de produit Data.bnf.fr
9. Adjointe au directeur des Collections pour les questions scientifiques et techniques
10. Directrice du département Droit, Économie et Politique
11. Directrice du département des Monnaies, médailles et antiques

1. Digital Scholarship Advisor, KB LAB
2. Head of Research Services, British Library
3. Directeur technique, TGIR Huma-Num

{BnF

PROJET CORPUS
Préparation de l'atelier du
8 décembre 2017

Je pense que la fouille de données est appelée à se développer

\- +

Je préfère avoir accès

Sur place — A distance

Je rencontre des problèmes relationnels

\- +

Je rencontre des problèmes juridiques

\- +

Je fais de la fouille de données

Accompagné — Seul

Je travaille

Petit à petit — Script qui tourne des heures

Je rencontre des problèmes techniques et j'ai du mal à les résoudre seul

\- +

J'aurais besoin de développer des outils

\- +

Je peux éventuellement les développer

-  +

Les formats que je ne connais pas me posent problème

+  -

J'ai une formation dans le numérique

-  +

Je fréquente la BnF

OUI  NON

Je ne veux travailler que sur mon portable

-  +

Je trouve tout ce qu'il me faut en ligne

NON  OUI

Mon objectif est

Recherche en SHS  Aider la recherche en SHS

Je viendrais sur place pour une infrastructure performante

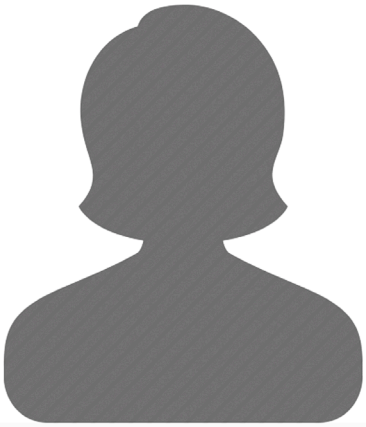
-  +

Je viendrais sur place pour rencontrer l'humain

- +

BnF

Atelier **CORPUS**
8 décembre 2017



CÉCILE

Âge : 50 ans

Métier : Professeure en histoire de l'art

Aisance numérique :

SHS :

« Moi je ne suis pas informaticienne, je ne suis pas née avec le numérique, ce n'est pas ma génération. »

« On est dans une complémentarité : quand j'écris un article, l'ingénieur rédige la partie technique et moi j'assume l'interprétation. Cela m'arrive d'avoir des problèmes avec les enseignants chercheurs en informatique parce que les visées d'un informaticien ne sont pas les mêmes que celles d'un historien. Le problème qui se pose dans toute interdisciplinarité c'est l'instrumentalisation de l'autre discipline. Le principal pour l'informaticien c'est qu'il teste l'efficacité de son logiciel. Or pour nous la perspective n'est pas là. »

« Il faut travailler en tryptique : le chercheur ou l'équipe de recherche, le conservateur et le soutien informatique. »

« Il ne faut pas penser que n'importe quel outil va répondre à nos besoins. »

Elle mène ses projets de recherche et elle guide aussi d'autres programmes de recherche dans son domaine.

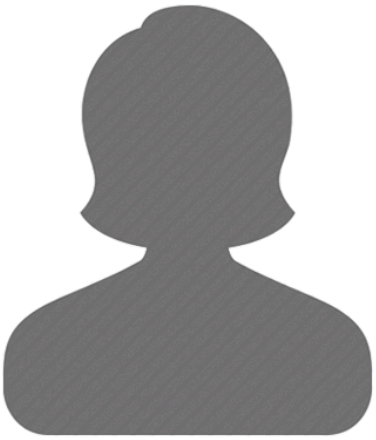
Elle souhaite faire développer des outils.

Buts clés

- OBTENIR DES FINANCEMENTS
- AVOIR DES LOCAUX
- ENSEIGNER ET FORMER SES ETUDIANTS
- TISSER DES LIENS INSTITUTIONNELS
- FAIRE DE LA RECHERCHE

Caractéristiques

- CRITIQUE
- SÛRE D'ELLE
- FORCE DE PROPOSITION ET POLITIQUE
- CAPABLE DE PLANIFIER LE TRAVAIL DE SES INGÉNIEURS DE RECHERCHE
- LE PROGRÈS DE SA RECHERCHE EST TRIBUTAIRE DE SES ÉQUIPES



SOFIA

Âge : 34

Métier : Ingénieur de recherche

Aisance numérique : ●●●○

SHS : ●●○○

« C'est une triade : il y a le système documentaire, le système informatique et le système éditorial. Le *do it yourself* s'arrête très vite. »

« Moi je n'ai pas de lieu pour pouvoir me réunir avec mes chercheurs. »

« Hors de question de prendre un ordinateur d'ici ou de n'importe où. Sur mon portable j'ai toutes mes archives, mes manips. »

« Je pense qu'on a tous notre environnement de travail, on s'est cassé le cou à se faire nos environnements de travail, avec son petit raccourci clavier qui va bien, avec son petit éditeur, moi je ne travaille que sur Atom...on a téléchargé plein de plug-in pour que cela fonctionne dans notre besoin. »

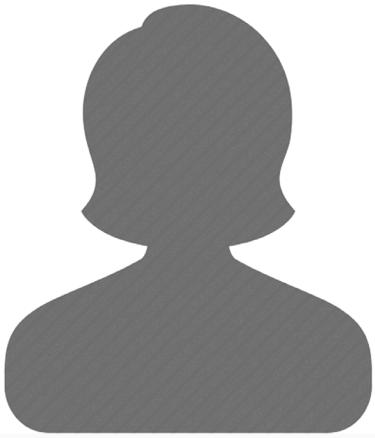
Sans mener un projet de recherche propre elle travaille dans un laboratoire de recherche en SHS. Elle aide les chercheurs sur les questions liées au numérique.

Buts clés

- ASSISTER ET FORMER LES CHERCHEURS
- PARTICIPER À UN OU PLUSIEURS PROJETS
- GÉRER ET MAINTENIR LES DONNÉES DE LA RECHERCHE
- DÉVELOPPER ÉVENTUELLEMENT DES OUTILS
- POURSUIVRE UNE CARRIÈRE DANS LE LABORATOIRE DANS LEQUEL ELLE TRAVAILLE

Caractéristiques

- OUVERTE
- PATIENTE
- CAPABLE DE TRAVAILLER EN ÉQUIPE
- INTÉRESSÉE PAR LA RECHERCHE EN SHS
- ELLE PRÉFÈRE TRAVAILLER SUR SA PROPRE MACHINE



LAURA

Âge : 27

Métier : Doctorante en histoire

Aisance numérique :

SHS :

« Je suis autodidacte. Je suis administratrice Wikipedia donc j'ai mis les pieds dans le monde des *geek*. »

« Je fais toujours des allées-retours entre les opportunités fournies par les outils et les questions sur comment faire émerger des choses. »

« Je ne travaille que sur mon portable. Tout mon travail est centralisé sur un seul poste.
C'est insupportable pour moi de devoir réinstaller le logiciel que j'ai déjà installé ailleurs. »

« Je travaille beaucoup chez moi mais je vois la limite. J'ai souvent des *crash*.
Il faut que l'infrastructure suive. »

« Je viens souvent à la BnF mais je préfère être chez moi quand je travaille à l'ordinateur. Souvent les meilleures idées je les ai à 1 heure du matin. »

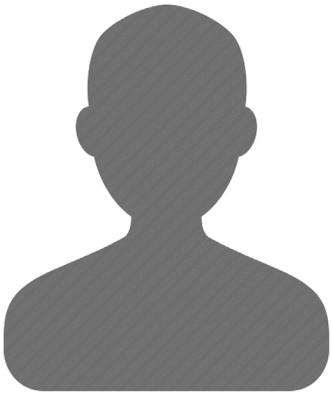
Elle débute son doctorat. Elle est en train de s'autoformer. Elle aime bien experimenter de nouvelles approches.

Buts clés

- COMMENCER SA THÈSE DE DOCTORAT
- EXPERIMENTER DE NOUVELLES METHODES DE RECHERCHE
- APPORTER DES CONNAISSANCES NOUVELLES DANS SON DOMAINE
- FAIRE DE LA RECHERCHE

Caractéristiques

- ELLE EST SEULE MAIS ELLE N'EST PAS SOLITAIRE
- ENTHOUSIASTE
- ENGAGÉE DANS L'OPEN SOURCE, L'OPEN DATA ET L'OPEN RESEARCH
- OUVERTE À L'INTERNATIONAL ET À L'ÉCHANGE
- ELLE EST TRIBUTAIRE DE SA PROPRE MACHINE MAIS ELLE N'A PAS DE RESOURCES POUR UNE INFRASTRUCTURE PERFORMANTE



ALEXIS

Âge : 28

Métier : Ingénieur spécialisé dans l'analyse de données

Aisance numérique : ●●●●

SHS : ●○○○

« J'ai une certaine ignorance de ce que peut mettre à disposition la BnF. »

« Mon script que j'ai lancé vendredi, j'ai calculé il va tourner une semaine sur une machine à 8 coeurs, donc il y a 4 processeurs qui tournent en même temps pendant une semaine. »

« Je n'ai pas besoin des données de la BnF sauf si je collabore avec un chercheur en SHS. »

« Les gens en SHS n'imaginent pas ce qu'on pourrait faire, c'est-à-dire qu'ils ne mesurent pas la difficulté de certaines choses qui sont évidentes pour l'oeil mais que nous on a beaucoup de mal à faire automatiquement et puis il y a des choses qui ne nous demandent pas trop de travail mais qu'ils n'imaginent pas. »

« La seule vraie raison de se déplacer aujourd'hui est d'aller voir l'humain. »

Employé ponctuellement, il aide des chercheurs ou des équipes de recherche en SHS.

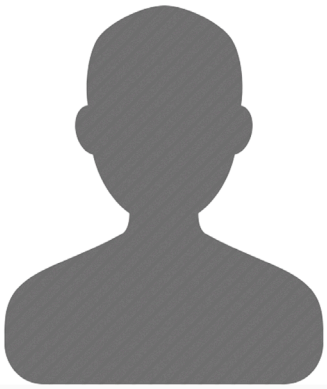
Il ne se limite pas à un seul domaine de recherche.

Buts clés

- DÉVELOPPER DES ALGORITHMES
- LES TESTER SUR DES MASSES DE DONNÉES
- LES VALIDER
- APPORTER UNE ASSISTANCE INFORMATIQUE POUR D'AUTRES
- AIDER LA RECHERCHE

Caractéristiques

- IL EST PARFOIS SOLITAIRE
- IL EST PARFOIS PROVOCATEUR
- IL RENCONTRE SOUVENT DE LA DIFFICULTÉ À DIALOGUER AVEC LES CHERCHEURS EN SHS
- IL EST CONVAINCU DE TROUVER TOUT CE QU'IL LUI FAUT SUR INTERNET
- IL MAÎTRISE SON ENVIRONNEMENT QUEL QU'IL SOIT



PIERRE

Âge : 36 ans

Métier : Chercheur en littérature

Aisance numérique :

SHS :

« J'ai eu du mal à ouvrir les fichiers de données dans leur base car je ne savais pas sur quelle application les ouvrir, je ne connaissais pas le format tout simplement. »

« On retient que ce qu'on a envie de retenir et ce qui nous intéresse. Ce qui veut dire que faire une formation une seule fois permet d'avoir seulement une partie des informations. Ce n'est pas comme un contact récurrent avec quelqu'un qui peut nous fournir des réponses. Parce que ce qui nous n'a pas intéressé la première fois devient un problème la fois suivante. »

« On doit pouvoir travailler sur nos ordinateurs portables. »

Il mène un projet de recherche propre dans un domaine spécifique. Il aimerait bien experimenter de nouvelles approches mais il n'a pas les capacités.

Buts clés

- POURSUIVRE UNE CARRIERE ACADÉMIQUE
- MONTRER UN INTÉRÊT POUR DE NOUVELLES APPROCHES AFIN D'ÊTRE PLUS VISIBLE
- APPORTER DES CONNAISSANCES NOUVELLES DANS SON DOMAINE
- BÉNÉFICIER D'UN ACCOMPAGNEMENT INFORMATIQUE
- FAIRE DE LA RECHERCHE

Caractéristiques

- CRITIQUE
- CRÉATIF
- FORCE DE PROPOSITION ET CAPABLE DE TRAVAILLER EN ÉQUIPE
- EN DEMANDE D'AIDE ET D'ACCOMPAGNEMENT
- IL EST TRIBUTAIRE DE SA PROPRE MACHINE OU DE L'INGÉNIEUR DE RECHERCHE AVEC LEQUEL IL TRAVAILLE

BnF | Bibliothèque nationale de France

François-Mitterrand
Quai François Mauriac, Paris 13e | 33 (0)1 53 79 53 79

Richelieu
5, rue Vivienne et 2, rue Louvois, Paris 2e | 33 (0)1 53 79 53 79

Bibliothèque de l'Arsenal
1, rue de Sully, Paris 4e | 33 (0)1 53 79 39 39

Bibliothèque-musée de l'Opéra
Place de l'Opéra, Paris 9e | 33 (0)1 53 79 37 40

Maison Jean Vilar
8, rue de Mons, 84000 Avignon | 33 (0)4 90 86 59 64

Centre technique (CTBnF)
14, avenue Gutenberg, 77607 Bussy-Saint-Georges | 33 (0)1 53 79 38 44

Centre de conservation Joël Le Theule
Le Château, 72300 Sablé-sur-Sarthe | 33 (0)2 43 95 19 92

www.bnf.fr