# BnF DataLab : appel à projets 2021

**Présentation et objet de l'appel**

*Le BnF DataLab*

Depuis une vingtaine d'années, la BnF conduit une politique en matière de numérisation de masse, d'ouverture des licences sur les métadonnées, de mise en place d'outils de consultation, d'analyse et d'extraction, d'accueil et d'accompagnement d'équipes de recherche expérimentant des traitements de corpus et données numériques (archives du web, fouille de textes ou de logs, reconnaissance d'images…). Les opportunités offertes aujourd'hui par les usages computationnels des documents numériques et la maturité des outils d'analyse changent les paradigmes de la recherche et permettent de lui ouvrir de nouveaux champs.

Le BnF DataLab est un service d'assistance et d'accompagnement à la recherche mis en place par la BnF en partenariat avec la TGIR Huma-Num, pour l'accueil de chercheurs qui souhaitent travailler sur les collections numériques de la BnF. Ces collections représentent en effet une masse considérable de documents, d'une grande diversité, tant par leur forme que par leur contenu : documents numérisés dans Gallica et Gallica intramuros, archives du web, métadonnées bibliographiques, jeux vidéo, documents audiovisuels (DVD, vidéos), etc.

Cette variété rend parfois délicate l'étude de corpus et demande un accompagnement à plusieurs niveaux d'expertise : expertise des fonds, des formats bibliographiques mais aussi expertise technique.

A travers des parcours d'accueil, le BnF DataLab permet aux chercheurs de bénéficier de l'ensemble de l'expertise de la BnF sur ces collections : aide à la constitution du corpus, extraction de données et métadonnées, accès à une infrastructure serveur dédiée, suivi de projet. De nouveaux espaces de travail spécifiques, installés dans une des salles de recherche du site François-Mitterrand de la BnF, permettent d'accueillir des chercheurs souhaitant travailler sur des documents numériques, y compris ceux sous droits (archives de l'internet, documents de Gallica intramuros), d'utiliser les infrastructures informatiques dédiées ainsi que la boîte à outils du BnF DataLab, mais aussi de bénéficier de l'assistance d'experts sur les collections et les traitements.

La mobilisation des différents niveaux d'expertise permet donc d'accueillir un chercheur voulant faire de l'analyse sémantique des romans du XIX$^{e}$ siècle ou un laboratoire d'intelligence artificielle qui voudrait entraîner ses algorithmes sur des corpus massifs et hétérogènes.

*Missions du BnF DataLab*

A travers l'ouverture du BnF DataLab, la BnF entend :

- Favoriser l'accès aux collections numériques à travers la mise en place d'environnements informatiques mais aussi d'un accompagnement par un large panel d'experts
- Mettre à disposition un environnement informatique propice à l'analyse, à la connaissance et à la valorisation des collections en articulation avec l'infrastructure Huma-Num.
- Structurer un laboratoire d'expérimentation en favorisant les échanges entre la recherche académique et les bibliothèques en s'appuyant notamment sur les communautés constituées autour de l'infrastructure Huma-Num et ses partenaires.

Les enjeux de cet appel à projet sont donc doubles : fournir un accès facilité et un accompagnement de grande qualité aux équipes dans leurs travaux de recherche ; permettre à la BnF de mieux connaître les pratiques autour de ces collections et approfondir son expertise pour développer ou parfaire des outils internes.

*Objectifs de l'appel*

Dans la continuité de ses missions de diffusion et de valorisation des collections, la BnF souhaite par le biais de cet appel à projet accueillir des équipes de recherche, afin d'accroître la connaissance mais aussi les possibilités d'analyse et de traitement de ses collections numériques. Les projets proposés devront faire appel à des méthodes et outils de traitement numérique.

L'objectif du présent appel est de permettre à des équipes de recherche de proposer des projets susceptibles d'être soutenus par le BnF DataLab, par le moyen d'un financement et d'un accompagnement spécifique, que ces projets soient nouveaux ou déjà en cours.

Quatre sujets sont proposés par la BnF mais les candidats peuvent également proposer spontanément leurs propres sujets de recherche, s'ils entrent dans le cadre du BnF DataLab. Les projets retenus pourront bénéficier de l'aide et l'assistance des experts de la BnF, tant d'un point de vue scientifique que technique.

*Description des services du BnF DataLab dont pourront bénéficier les projets retenus*

L'utilisation des services se fera dans les limites de la capacité du BnF DataLab et sera défini entre l'équipe du BnF DataLab et le responsable de chaque projet retenu.

<u>Accueil dans les espaces du BnF DataLab:</u>

- Possibilité de réserver une salle de groupe de 4 places, accès à des box individuels de travail équipés d'un poste informatique et de deux écrans ;
- Accès à un espace serveur et une machine virtuelle depuis les infrastructures dédiées de la BnF et de la TGIR Huma-Num ;
- Accès à la boîte à outils du BnF DataLab : un environnement de travail Linux (Ubuntu), un environnement de programmation Anaconda, un IDE (Pycharm) et un client FTP (FileZilla). Cet environnement pourra être complété par le chercheur selon ses besoins (sous réserve de la validation de la BnF).

<u>Accompagnement et suivi de projet :</u>

- Aide à la constitution de corpus : assistance bibliographique, aide pour l'extraction de corpus web et/ou pour l'extraction de données et/ou de documents ;
- Formations : utilisation des API BnF, utilisation des formats bibliographiques, suivi de projet ;

- Assistance technique : installation d'outils, entretiens avec des experts BnF, accompagnement technique et opérationnel avec les ingénieurs du BnF DataLab ;
- Accès à la grille de services de la TGIR Huma-Num en accord avec l'équipe technique : notamment le dépôt accompagné dans l'entrepôt Nakala, accès à des GPU pour traitement de corpus, accompagnement technique et opérationnel pour la diffusion et l'exposition des données (Nakala Press, site web), référencement dans Isidore.science.

*Résultats et reversement des outils*

Dans le cadre de la politique pour la science ouverte formalisée dans le Plan national pour la Science Ouverte, les réalisations produites dans le cadre de cet appel à projet pourront être valorisées et mises à disposition des communautés de chercheurs souhaitant réutiliser les applications, les scripts et les corpus utilisés. Ces réalisations pourront rejoindre, sous réserve de validation, la boîte à outils du BnF DataLab.

*Types de projets visés par le présent appel*

- Durée max. 1 an / budget max. 25 000 €
- Les projets proposés peuvent permettre d'amorcer une nouvelle recherche ou compléter un programme déjà existant.
- Lien fort avec le programme et les missions du BnF DataLab.
- Le projet doit présenter un clair enjeu scientifique pour le traitement et l'analyse des collections numériques de la BnF. Il reviendra à l'équipe de recherche de mettre en œuvre ces méthodes et outils de traitement.

*Préconisations méthodologiques*

Tout en précisant les enjeux scientifiques, les projets soumis à l'appel devront respecter les obligations suivantes :

- porter obligatoirement sur les collections numériques de la BnF ;
- présenter des étapes de traitement numérique des collections (à décrire dans la proposition) ;
- proposer des corpus, des méthodes ou des outils d'analyse originaux présentant un intérêt pour une communauté bien identifiée et en lien avec les problématiques du BnF DataLab, et livrer des résultats et outils d'aide à la recherche utiles et accessibles à la communauté scientifique ;
- s'engager sur un résultat final sur la base d'une description précise du livrable attendu et d'un planning détaillé ;
- s'engager à faire un retour d'expérience sur sa recherche dans le cadre d'ateliers du BnF DataLab.
- prévoir des actions de valorisation de la recherche (carnets de recherche, publications, communications lors de colloques…) ;
- détailler les modalités d'accès aux sources, aux corpus et données et leurs modes de constitution, de traitement et de conservation ;
- être attentifs aux questions de la pérennisation des données et des résultats de la recherche ;
- prendre en compte les questions de protection des données personnelles et les droits de propriété intellectuelle sur certains types de données ou corpus.

**Sujets proposés dans le cadre de l'appel**

Quatre sujets sont proposés en lien avec des collections ou des thématiques spécifiques à la BnF. Les candidats peuvent également proposer leur propre sujet de recherche s'il rentre dans le cadre des objectifs de l'appel et des enjeux du BnF DataLab tels que décrits plus haut.

Les équipes de recherche sont encouragées à contacter la BnF en amont du dépôt pour poser toute question sur leur projet. Merci de contacter datalab@bnf.fr

1) *Covid19 et Archives du web*

Les archives web de la Covid-19 en tant que collection sont définies comme l'ensemble des contenus collectés par la BnF entre février et fin juillet 2020 dans le cadre de ses collectes courantes. Les données sont conservées dans 15 504 fichiers WARC représentant 15To de données compressées dont 1To de vidéo. Les contenus collectés ont été indexés avec le moteur de recherche Solr.

Pistes de recherche :

- Information et vulgarisation scientifique en période de crise.
- Le vocabulaire de la Covid-19, les registres mobilisés et leur réception critique.
- Le web comme moyen de mobilisation et de témoignage
- Comparaison des politiques nationales et comparaison avec d'autres pandémies historiques

Ces pistes ne sont pas exhaustives et les sujets proposés peuvent concerner les aspects médicaux, sociaux, politiques, économiques et culturels.

2) *Archives de la jouabilité et sources de l'histoire des jeux vidéo*

La BnF détient une collection de jeux vidéo constituée à partir du milieu des années 1990, quand le dépôt légal fut étendu à ces supports. Elle s'enrichit chaque année avec des acquisitions de titres plus anciens visant à renseigner les débuts de l'édition vidéoludique. La bibliothèque collecte et conserve également de nombreux autres supports nécessaires aux chercheur(e)s en sciences du jeu : monographies et périodiques, sites web et vidéos en ligne dans les archives de l'internet, enregistrements sonores et vidéos, archives, etc. Ces documents en lien avec les jeux vidéo sont conservés dans des services et des lieux divers au sein de la BnF, et leur consultation ne peut se faire de façon simple et centralisée.

Le projet consisterait à travailler sur des corpus regroupant ces différents types de documents, utiles à l'histoire du jeu vidéo et déjà présents dans les collections de la BnF (notamment monographies, périodiques, vidéos et archives de l'internet). L'étude devra comporter une identification de ces différentes sortes de contenus, avec une attention particulière apportée aux enregistrements de parties et témoignages de joueurs relevant des archives de la jouabilité à proprement parler. À titre de test, on pourra se concentrer sur un panel limité de jeux, d'un type ou d'un éditeur particulier, ou sur une période donnée.

La réflexion pourra porter également sur les métadonnées qui doivent accompagner ces documents et la manière de les lier entre elles, en vue de faciliter leur signalement et au final leur visibilité par les chercheur(e)s travaillant sur ces questions. Il s'agira aussi d'envisager des moyens techniques permettant de consulter sur un même poste les jeux eux-mêmes et les matériaux de tous types qui peuvent les concerner.

En fonction du sujet proposé et des titres concernés, on travaillera enfin à la possibilité de produire des archives de la jouabilité inédites : réalisation de vidéos de parties, en lien avec les besoins des chercheur(e)s et avec l'aide et l'accompagnement éventuel d'éditeurs et de studios de développement.

(Voir : Hugo Montembeault et Simon Dor, « À quoi pensent les archives de la jouabilité? », *Conserveries mémorielles* [En ligne], #23 | 2018, mis en ligne le 10 octobre 2018, consulté le 12 avril 2021. URL : http://journals.openedition.org/cm/3171).

*3) Gallica et les collections numérisées*

Bibliothèque numérique de la BnF et de ses partenaires, Gallica comprend des documents de toute nature : monographies, périodiques, estampes et photographies, manuscrits, cartes et plans, partitions musicales, etc. Avec plus de 8 millions de documents, l'enjeu est d'outiller la bibliothèque numérique pour les chercheurs et de travailler la collection en tant que données. Les projets proposés pourront ainsi recourir aux techniques IA pour l'enrichissement des collections (production de métadonnées, extraction de contenu…), utiliser les API, etc.

Pistes de recherche :

- Éditorialisation et navigation dans des documents et corpus complexes (dictionnaires, CR de l'académie des sciences, etc.)
- Outils de visualisation de la collection
- Extraction des entités de lieux et géolocalisation des documents
- Anonymisation de l'indexation

*4) Vectorisation automatique de corpus cartographiques patrimoniaux*

Le patrimoine cartographique numérisé du département des Cartes et plans (près de 61 000 documents libres de droits et 10 000 documents sous droits) et de la Société de géographie (plus de 10 000 documents aux droits de diffusion négociés et 8000 documents sous droits) est accessible dans la bibliothèque numérique de la BnF, Gallica et Gallica Intramuros. S'y ajoutent les collections des partenaires de Gallica susceptibles de contenir des documents cartographiques (par exemple, la Bibliothèque historique de la Ville de Paris). Cette richesse documentaire peut se décliner en corpus pertinents pour des projets de recherche impliquant une géospatialisation de données historiques (géolocalisation d'objets et de documents, géo-référencement, géocodage à partir d'annuaires anciens) dans des domaines comme l'histoire des représentations cartographiques, l'histoire de l'urbanisme et de la géographie...

La BnF propose d'accompagner, dans ce contexte, des projets requérant la possibilité de vectoriser automatiquement des corpus de plans de villes numérisés, selon un prototype utilisant les techniques de l'intelligence artificielle (algorithme fondé sur un réseau de neurones spécifiquement entraîné) : voir la documentation technique disponible sur la page GitHub du projet : https://github.com/BnF-jadis. Cette application pourrait être réutilisée sur des corpus de plans de différentes villes, à partir des collections de la BnF et d'autres institutions patrimoniales, afin, par exemple, de réaligner ces cartes sur plan actuel ou d'en extraire les données.

*5) Sujet libre (à proposer par l'équipe de chercheurs)*

Les candidats peuvent proposer tout autre sujet de recherche qui rentre dans le cadre des objectifs de l'appel, des missions du BnF DataLab et des préconisations méthodologiques.

**Modalités de soumission**

Le responsable du projet doit remplir le formulaire joint à l'appel qui contient les éléments suivants :

- Description du projet : sujet, objectifs, livrables
- Détails de la mise en œuvre
- Présentation du lien avec les missions et les problématiques du BnF DataLab
- Pistes de valorisation
- Budget et le cas échéant les profils des postes à financer
- CV du responsable

Le dossier complet doit être envoyé sous forme d'un fichier PDF unique par courriel à l'adresse datalab@bnf.fr

Les dossiers doivent être reçus au plus tard **le 28 juillet 2021 (17h)**.

**Règlement**

En déposant un dossier, le candidat reconnaît avoir pris connaissance du présent règlement et déclare l'accepter.

*Recevabilité*

- Le dossier de soumission, sous forme électronique, doit être transmis dans les délais, au format demandé et être complet.
- Le dossier, dans le format fourni et y compris le CV du responsable, ne doit pas dépasser 10 pages.
- La durée maximum du projet présenté est de 12 mois ; le projet doit démarrer à partir du 1er octobre 2021 et au plus tard le 31 octobre 2021 et doit se terminer au plus tard le 31 octobre 2022.
- Le montant de financement demandé ne peut pas excéder 25 000 € TTC.

*Eligibilité*

- Le projet doit être porté ou soutenu par au moins un laboratoire ou une unité de recherche relevant d'un établissement public de recherche.
- Le responsable du projet doit être au niveau doctorant ou supérieur.

*Calendrier*

- Lancement de l'appel à projet : juin 2021
- Date limite de dépôt des dossiers : 28 juillet 2021
- Jury de sélection : semaine du 13 septembre 2021
- Notification des résultats : semaine du 20 septembre 2021
- Démarrage des projets : entre le 1er et le 31 octobre 2021

*Critères et processus de sélection*

- Les projets seront évalués à partir des critères suivants :
    - Recevabilité : dépôt avant date limite du dossier complet et respect du budget maximum ;
    - Adéquation du projet proposé avec les missions du BnF DataLab et la valorisation des collections numériques de la BnF ;
    - En particulier, le projet doit présenter une problématique de recherche portant obligatoirement sur les collections numériques de la BnF et doit obligatoirement

présenter des étapes de traitement numérique des collections (à décrire dans le dossier) ;
    - Les projets portant sur des corpus déjà disponibles (déjà numérisés / déjà collectés) seront privilégiés pour permettre une mise en œuvre rapide du projet ;
    - Qualité du dossier : originalité, gestion du projet, compétences mobilisées, faisabilité scientifique et budgétaire ;
    - Qualité du livrable ; s'il s'agit d'un logiciel, une préférence sera donnée aux réalisations sous licence libre ;
    - Prise en compte des problématiques de gestion et de pérennisation des données ; la présence d'un plan de gestion des données le cas échéant sera un atout.
- Tout projet déposé fera l'objet d'une évaluation préalable par le BnF DataLab et les départements de la BnF concernés. Le choix des projets retenus se fera sous réserve de la faisabilité en ce qui concerne la disponibilité des ressources nécessaires au sein de la BnF. De même, l'utilisation des services par les projets retenus se fera dans les limites de la capacité du BnF DataLab et des autres départements de la BnF concernés.
- Les projets retenus seront choisis par un jury composé des membres de la BnF et de la TGIR Huma-Num.

La non sélection de projets ne peut faire l'objet d'une contestation.

*Dépenses*
- Toutes dépenses confondues, le montant de financement demandé ne peut excéder 25 000 € (TTC).
- Dépenses éligibles :
    - recrutement IGR / IGE ; stages ;
    - missions (le montant maximum des missions ne pourra pas dépasser 10% de la subvention demandée) ;
    - prestations techniques de fouilles de données ou de développement d'outils liés au traitement numérique des collections, ainsi que les licences logiciels nécessaires.
- Dépenses non éligibles :
    - frais RH de personnel permanent ;
    - frais de développement ou de maintenance de type site web, stockage des données ;
    - frais de publication.
- Le responsable s'engage à fournir un tableau de dépenses certifié et toutes les pièces justificatives.

*Livrables / utilisation des résultats*
- Le responsable doit décrire dans sa proposition les livrables du projet.
- Il est attendu que le livrable prenne la forme soit d'un outil, soit d'une démonstration d'une technique d'analyse, susceptibles d'être valorisés dans le BnF DataLab.
- Le responsable d'un projet retenu s'engage à participer à un atelier de partage d'expériences sur sa recherche organisé par le BnF DataLab.
- Les résultats peuvent être utilisés (à des fins commerciales ou non-commerciales) par la BnF et par l'établissement porteur. En particulier, les logiciels, scripts etc. ont vocation à rejoindre la boîte à outils du BnF DataLab (cf. supra) et être proposés sur la grille de service de Huma-Num.
- Pour les logiciels développés dans le cadre du projet, l'utilisation d'une licence libre est préconisée.
- La BnF pourra communiquer sur les projets retenus ; les chercheurs s'engagent à respecter les mentions obligatoires dans toute communication.

*Convention*

- Une convention sera signée entre la BnF et l'établissement porteur du projet pour encadrer le versement de la subvention, son utilisation, l'utilisation des résultats et les conditions d'accueil.
- L'équipe de recherche d'un projet retenu s'engage à respecter le règlement des salles de lecture de la BnF et, de manière générale, toute charte ou réglementation applicable relatives notamment à la sécurité des collections, la circulation dans les espaces de la BnF et la reproduction des documents issus des collections de la BnF.

*Modalités de versement*

- Le financement accordé à tout projet retenu sera versé en deux temps, au début des travaux et à la réalisation des livrables.
- Les modalités de versement seront précisées dans la convention.