Webinaire La BnF de chez vous #22 :
BnF Datalab

Mardi 2 juin, la 22e séance du webinaire "La BnF de chez vous", inscrite dans la thématique "Si j'avais un marteau", a accueilli Emmanuelle Bermès (DSR), Marie Carlin (ORB), Arnaud Laborderie (DCP), Alexandre Faye (DLN) et Olivier Jacquot (DSG). Ils ont présenté le Data Lab, nouveau service mis en place par la BnF pour développer la recherche sur les collections numériques (Gallica, Gallica intra muros, archives de l'internet...). Ce lieu conçu pour le travail collaboratif ouvrira ses portes en 2021 en salle X.

* * *

BnF Datalab est issu du projet Corpus qui a démarré en 2016 suite à des demandes de chercheurs. Ceux-ci souhaitaient pouvoir accéder à la consultation de corpus numériques massifs et appliquer à ces collections de nouvelles méthodes d'analyse des données adaptées aux besoins de leurs recherches : fouille de données, machine learning, etc. L'expérimentation du projet s'est déroulée sur trois ans, utilisant ces nouvelles méthodes pour analyser les données issues des archives du web, de Gallica, les métadonnées du catalogue etc., afin de répondre à des projets de recherche concrets. La dernière année a permis de faire le bilan de l'expérimentation et de définir les contours du service BnF Datalab qui sera proposé en 2021.

BnF Datalab est un projet transverse qui n'est pas rattaché à une direction et nécessite l'expertise de nombreux collègues de toutes directions et tous départements confondus.
Trois axes caractérisent BnF Datalab :

- c'est un lieu de travail, d'échange, de résidence pour les chercheurs
- c'est un catalogue de services pour accompagner les usagers
- c'est un laboratoire, favorisant l'expérimentation par le biais de partenariats

1/ L'implantation du Datalab

Initialement le Datalab n'avait pas un ancrage physique. Cette décision d'installer le Lab en salle X (arrière-banque et mezzanine) répond à une demande des chercheurs d'avoir un lieu de travail, de rencontre entre pairs ou entre chercheurs et experts BnF, afin de leur offrir une infrastructure numérique facilitant les opérations de fouille de données.
Les travaux ont débuté le 9 mars, ont dû cesser avec le confinement, puis ont repris le 25 mai.
L'implantation d'un espace physique pour le Datalab a également permis de résoudre la question des corpus sous droits qui ne peuvent pas être consultés à l'extérieur de la BnF.

2/ Un catalogue de services

Plusieurs catégories de services seront proposés aux chercheurs :

Les services relatifs aux collections.

Un groupe de travail d'agents de l'ORB a établi une liste d'une vingtaine de services à mettre en place pour accueillir et accompagner les chercheurs dans le Lab. En premier lieu, les chercheurs pourront prendre rdv avec un binôme d'agents de l'ORB, ou d'autres experts, pour les aider à définir leurs corpus. En effet, il est indispensable que les usagers du Lab sachent quelles collections sont disponibles, quelles sont celles numérisées ou qui peuvent l'être. Des parcours ont été définis pour déterminer les services, mais aussi les formations à proposer à chaque étape du projet de recherche.

Les services techniques

Une dizaine de services et de prestations seront proposés par les départements de la DSR.
Les besoins des chercheurs concernent essentiellement la fourniture de données des collections patrimoniales : images, textes issus de l'OCR, métadonnées. Il leur sera proposé des outils de requêtage des données (API) ainsi qu'un accompagnement à la prise en main de ces outils, mais aussi des services complémentaires comme l'extraction en masse de données ou un service de numérisation à la demande. De nouveaux environnements de travail seront mis à leur disposition : machine virtuelle, boîte à outils logiciels, etc. De nouveaux outils allant au-delà de la fouille de texte pourront être développés (segmentation des données, etc.) et leur mutualisation permettra d'enrichir Gallica, servant ainsi tous les usagers, au-delà de la communauté des scientifiques.

Les services du Datalab relatifs aux archives de l'internet

Les collections les plus anciennes datent de 1996 et progressivement, des outils de recherche sont venus enrichir l'interface d'interrogation des archives de l'internet : recherche plein texte dans trois corpus, parcours guidés. L'indexation en plein texte de toutes les archives n'est pas possible mais les chercheurs doivent pouvoir interroger et manipuler les données grâce à des scripts et constituer leurs propres corpus. Le Datalab permettra de proposer des services et outils de traitement spécifiques aux archives de l'internet : collectes web à la demande, extraction des données archivées,... Le Datalab sera un lieu de travail et d'échanges permettant de coconstruire le service avec les utilisateurs. En ce sens, l'organisation d'un Datathon est prévue en 2021, en collaboration avec d'autres bibliothèques nationales.

3/ La coopération scientifique autour du projet

Le développement du Datalab a été pensé dès l'origine pour travailler en partenariat avec les chercheurs de différents profils et différentes institutions scientifiques. Parmi les partenaires sollicités, le CNRS a répondu favorablement et le projet a abouti à la signature d'une convention avec Huma-Num. Un autre partenariat a également été noué avec l'OBVIL pour accueillir dans le Datalab des chercheurs sur une longue durée. La conclusion de ces partenariats permet à la BnF de bénéficier de l'expertise de ces institutions, de leurs moyens humains, de leurs outils de travail dans le champs des humanités numériques. C'est enfin une opportunité pour valoriser les actions qui seront menées dans le Datalab : manifestations, restitutions des travaux des équipes de recherche, publications, etc. Les futurs appels à projet de la BnF seront l'occasion de mettre en valeur le Datalab et de proposer des projets de recherche communs Huma-Num/BnF.