Les humanités numériques à la Bibliothèque nationale de France

Thierry Pardé, Délégué à la stratégie et à la recherche, Bibliothèque nationale de France
Olivier Jacquot, Coordonnateur de la recherche, Bibliothèque nationale de France

Trois domaines d'innovations sont explorés à la Bibliothèque nationale de France, dans le cadre de plusieurs projets de recherche nationaux ou européens : l'amélioration de l'accès aux données numériques, l'exploitation de ces données massives, et l'étude des usages numériques. Ces recherches soulèvent des questions techniques mais aussi juridiques.

La recherche occupe à la BnF une place de premier plan tant par le nombre des programmes conduits (plus de 150 recensés en 2015) que par les financements en jeu et la qualité des partenariats engagés[1]. En raison de la richesse de ses collections patrimoniales, elle couvre un vaste champ de supports et de disciplines : la science des bibliothèques, la bibliographie, l'histoire du livre, de l'édition et des médias, l'iconographie, la numismatique, la musique, la cartographie, la conservation, la numérisation, la préservation des données numériques, etc.

A côté des programmes traditionnels sur le patrimoine, l'augmentation massive des collections numériques ouvre de nouvelles pistes de recherche, soulève des problématiques nouvelles d'exploitation des collections par les chercheurs et fait surgir de nouveaux usages. Les données produites et conservées, qu'il s'agisse des notices bibliographiques et d'autorité placées sous licence ouverte de l'Etat, des corpus numérisés et océrisés de Gallica ou des archives de l'internet constituées au titre du dépôt légal, offrent à cet égard d'immenses possibilités.

Pour répondre à ces enjeux, la BnF a élargi ses collaborations scientifiques aux humanités numériques en s'engageant dans des projets expérimentaux visant à exploiter ses corpus de données en leur appliquant des traitements informatisés. Cet engagement porte tout autant sur l'étude et la connaissance des usages en ligne du patrimoine numérique que sur la mise au point d'outils des sciences de la donnée (*big data* et *text and data mining* - *TDM*) destinés à l'exploitation intelligente de masses de données et permettant de faire émerger de nouvelles informations, ou cartographies d'informations, à partir des corpus numériques constitués ou collectés. Ces nouveaux enjeux s'inscrivent pleinement dans les défis sociétaux retenus dans la Stratégie nationale de recherche 2020.

Trois principaux domaines d'innovation sont aujourd'hui explorés à la BnF : l'amélioration de l'accès aux données numériques (correction, enrichissement, interopérabilité), l'exploitation de ces données massives (fouille, analyse, indexation) et enfin, l'étude des usages numériques.

Amélioration des données et de leur interopérabilité

Depuis plusieurs années, la BnF a investi le champ de la recherche appliquée à la reconnaissance optique de caractères (OCR). Il s'agit pour l'institution de communiquer à ses usagers des documents numérisés en mode texte de la meilleure qualité possible, une mission que peut seule satisfaire l'acquisition de savoir-faire et d'outils adéquats, conçus dans le cadre de travaux de R&D.

Ainsi, le projet européen Europeana Newspapers (2012-2014) destiné à faciliter l'accès du public aux articles de la presse quotidienne européenne a-t-il permis de traiter des questions aussi diverses et complexes que la correction des contenus océrisés, la reconnaissance optique des articles de presse et leur structure logique (rubriquage, titrage, découpage de l'article), l'enrichissement sémantique des données et métadonnées relatives aux contenus (contenus éditoriaux, changement du nom du journal ou d'éditeurs, présence de photographies, de dessins, de graphiques…) ainsi que les possibilités de recherche en texte intégral. Avec le concours du Laboratoire d'informatique de Paris 6 (LIP6), la BnF a adopté des outils de traitement automatisé de la langue permettant de reconnaître des entités nommées singularisant dans un texte les noms de lieux, de personnes et d'institutions.

[1] La BnF est membre de 5 Labex : ARTS-H2H (Arts et médiations humaines), CAP (Création, arts et patrimoines), PATRIMA (Patrimoines matériels, savoirs, patrimonialisation, médiation), PASP (Les passés dans le présent), OBVIL (Observatoire de la vie littéraire) et de 3 Equipex : BIBLISSIMA (Bibliotheca bibliothecarum novissima), PATRIMEX (Patrimoines matériels, réseau d'instrumentation multisite expérimental), ORTOLANG (Outils et ressources pour un traitement optimisé de la langue). Elle exerce la co-tutelle de l'UMR IReMUS (Institut de recherche en Musicologie) dont elle héberge une partie des équipes de recherche.

Des travaux de R&D similaires s'appliquent à d'autres types de contenus, comme les partitions musicales ou les manuscrits, pour lesquels outils et méthodes de transcription font défaut. Le Fonds unique interministériel a financé le projet *OZALID* porté par Orange qui a été marqué par l'expérimentation d'une plateforme de participation collaborative, CORRECT[2], qui a permis de tester la correction et l'enrichissement collaboratifs de textes numérisés de Gallica. Cette expérimentation débouchera sur la réalisation d'une interface, dite « Gallica studio », distincte de la bibliothèque numérique : elle sera librement accessible aux internautes qui pourront importer des documents de Gallica afin de les corriger, les géolocaliser, les annoter, ou encore réaliser des projets créatifs. Les documents enrichis pourront ensuite être reversés dans Gallica.

Pour développer les processus d'ouverture, de partage et de réutilisation des données, la BnF est engagée dans un programme de recherche soutenu par l'ANR, *DOREMUS : DOnnées en REutilisation pour la Musique en fonction des Usages*[3], qui doit permettre aux institutions culturelles, aux éditeurs et distributeurs, ainsi qu'aux communautés de passionnés, de disposer de modèles de connaissance communs (ontologies), de référentiels partagés et multilingues ainsi que de méthodes pour publier, partager, connecter, contextualiser, enrichir les catalogues d'œuvres et d'événements musicaux dans le web de données.

Enfin, un projet autour de l'interopérabilité des entrepôts d'images adoptant le modèle de données informatique *Shared Canvas*[4] pour l'interopérabilité des manuscrits numériques et la spécification technique International Image Interoperability Framework (IIIF) a été développé en coopération avec l'Université de Stanford et l'Equipex BIBLISSIMA. Le projet permet l'interopérabilité des entrepôts d'images, c'est-à-dire la communication et l'échange via le Web d'images numériques entre différents entrepôts, quels que soient les types de documents concernés (livres, photographies, journaux, manuscrits, cartes…).

Les sciences de la donnée à la BnF

Trois types de corpus, de nature très différente, ouvrent des perspectives prometteuses dans le domaine de la fouille de données :
a) les corpus de métadonnées, accessibles librement via data.bnf.fr[5] et dont la mise à disposition encourage leur réutilisation pour créer de nouveaux services ;
b) les corpus de textes numérisés et océrisés qui peuvent faire l'objet de traitements avancés permettant de produire de la connaissance sur les collections à des fins diverses (analyse, valorisation, *crowdsourcing*) ;
c) les corpus d'archives du web qui peuvent donner lieu à de la fouille de métadonnées et de liens afin d'élaborer des cartographies du web ainsi qu'à de la fouille ciblée de textes et de médias.

Une partie des sources sur lesquelles se fondent ces recherches est agrégée en corpus exposés sous la forme de bases en ligne telles que « BnF - Archives et Manuscrits » pour les textes manuscrits et les fonds d'archives, « Reliures »[6] pour les reliures les plus remarquables de la Réserve des livres rares ou « BP16 »[7] pour la production imprimée parisienne du XVI^e^ siècle.

L'édition scientifique des textes passe par leur encodage en XML et TEI (technique et formats de conversion en mode texte) : elle est au cœur des projets menés en collaboration avec le Labex OBVIL qui procède à l'édition numérique annotée de corpus littéraires à partir de la fourniture de fichiers numériques par la BnF.

L'exploration des méga-données rencontre également des initiatives menées à l'échelle internationale : ainsi, l'extraction pour le Labex OBVIL d'un volumineux corpus de textes des XVIII^e^ et XIX^e^ siècles

[2] URL : <http://www.reseau-correct.fr/>.
[3] URL : <http://www.doremus.org/>.
[4] URL : <http://iiif.io/model/shared-canvas/1.0/index.html>.
[5] URL : <http://data.bnf.fr/>.
[6] URL : <http://reliures.bnf.fr/>.
[7] URL : <http://bp16.bnf.fr/>.

fera-t-il l'objet de recherches conduites en partenariat avec l'université de Chicago (projet ARTFL[8]) pour mettre au point des outils d'alignement de textes et d'analyse philologique.

En matière d'iconographie, la BnF conduit avec le laboratoire ETIS de l'université de Cergy un projet soutenu par le Labex Patrima[9] qui porte sur la reconnaissance de formes, sur la recherche par similarité et sur la possibilité de procéder à une annotation semi-automatique d'images patrimoniales. À terme, de telles solutions pourraient venir enrichir les fonctionnalités de Gallica pour les quelque 1,25 million d'images aujourd'hui présentes.

L'étude des usages du numérique

De nouveaux objets de recherche sont apparus pour la BnF qui touchent aux usages du patrimoine numérique de Gallica et à la compréhension des types d'appropriation qu'il suscite, dans un contexte d'éclatement des pratiques numériques. À cette fin, la BnF a créé avec Télécom ParisTech le Laboratoire d'étude des usages du patrimoine numérique des bibliothèques (Bibli-Lab)[10].

Ses recherches s'appuient sur de nouvelles méthodes d'analyse, telles les techniques de fouille de texte ou de fouille de liens (*text mining* ou *link mining*) mais aussi l'ethnographie numérique. À titre expérimental, la BnF a piloté, dans le cadre du Labex « Les Passés dans le Présent »[11], en collaboration avec Télécom ParisTech et la Bibliothèque de documentation internationale contemporaine (BDIC), le projet de recherche « Le devenir en ligne du patrimoine numérisé : l'exemple de la Grande Guerre » (2013-2016) afin de mieux comprendre la manière dont un corpus numérisé – celui de la guerre 14-18 – se diffuse à travers le web, y suscitant des échanges et des appropriations multiples. L'une des phases de l'étude s'est appuyée sur un corpus issu du web collecté et archivé par la BnF. Des outils de TDM ont été utilisés, d'une part, pour cartographier les liens entre les sites, d'autre part, pour analyser les contenus de l'un des principaux forums en ligne consacrés à la Grande Guerre. Le programme a ainsi permis de situer les bibliothèques numériques patrimoniales à l'intérieur d'un réseau vaste et complexe de sites, blogs, forums, institutionnels ou individuels.

Conclusion

Ces nouveaux terrains de la recherche sur le numérique soulèvent des questions d'ordre technique mais également juridique. Régulièrement sollicitée pour fournir des corpus numériques, la BnF doit lever les freins qui rendent les recherches en humanités numériques aujourd'hui complexes et coûteuses.

Sur le plan juridique, la fouille de texte et de données n'est pas à l'heure actuelle inscrite dans un contexte législatif stabilisé, comme les débats autour du projet de loi pour une République numérique en témoignent. Jusqu'alors, les expérimentations conduites à la BnF ont porté sur des documents du domaine public ou ont été menées dans le cadre de l'exception dépôt légal qui en permet un usage dans des conditions définies par le Code du patrimoine.

Sur le plan technique, la création d'un environnement numérique adapté aux travaux des équipes de recherche requiert compétences et moyens spécifiques. Les chaînes d'extraction de données doivent être adaptées, ou parfois construites ex-nihilo, pour assurer la mise à disposition des corpus.

Aussi, la variété et la fertilité des coopérations aujourd'hui engagées avec des équipes de recherche - nationales comme internationales - à la pointe des humanités numériques sont-elles déterminantes pour inventer les outils, espaces et services numériques innovants que la BnF se doit d'apporter aux chercheurs pour qu'ils explorent le gigantesque réservoir des données numériques patrimoniales qu'elle préserve.

---

[8] L'American and French Research on the Treasury of the French Language (ARTFL) est un projet de recherche américain et français du Laboratoire ATILF (Analyse et Traitement Informatique de la Langue Française), du Centre National de la Recherche Scientifique (CNRS), la Division of the Humanities and Electronic Text Services (ETS) de l'Université de Chicago.

[9] URL : <http://actions-recherche.bnf.fr/BnF/anirw3.nsf/IX01/E2012000067_fondation-des-sciences-du-patrimoine>.

[10] URL : <http://www.bnf.fr/fr/la_bnf/pro_publics_sur_place_et_distance/a.bibli-lab.html>

[11] URL : <http://actions-recherche.bnf.fr/BnF/anirw3.nsf/IX01/E2013000132_labex-pasp>.