L'objectif de BaOIA est de développer, à partir de corpus numérisés de La contemporaine et de la BnF, des outils d'exploration et d'analyse de corpus massifs, de les documenter et de les mettre à disposition des chercheurs pour une réutilisation ultérieure.

BaOIA : *une « boîte à outils» pour explorer les corpus numérisés*

Le projet de recherche Boîte à Outils d'Intelligence Artificielle (BaOIA), financé par CollEx-Persée, est né de la rencontre entre les intérêts d'un chercheur pour les humanités numériques et ceux d'un établissement documentaire disposant de nombreux corpus numérisés. Quand Julien Schuh du Centre des sciences des littératures en langue française (CSLF) de l'université Paris Nanterre lui a soumis son projet de développement d'outils d'analyse de données massives, La contemporaine a saisi l'opportunité de valoriser ses collections numérisées et de développer un axe de sa propre politique de recherche.

La contemporaine poursuit de très longue date une mission d'appui à la recherche. La mise à disposition de collections spécialisées permettant d'écrire l'histoire des relations internationales et l'histoire politique et sociale depuis le début du XXe siècle en constitue la clé de voûte : monographies, revues, presse généraliste ou à diffusion restreinte, affiches, brochures et tracts, photographies, archives privées et audiovisuelles, dessins ou peintures, sont ainsi depuis 1917 collectés et signalés, sans hiérarchie, en tant que *matériaux* pour l'écriture de l'histoire. La volonté de mettre ces sources à disposition du plus grand nombre et de faciliter leur confrontation ont présidé à la mise en ligne d'une bibliothèque numérique, l'Argonnaute[1], devenue l'une des plus importantes de l'Enseignement supérieur. Ce même enjeu a guidé le projet de construction d'un nouveau bâtiment qui a ouvert ses portes en octobre 2021 : toutes les collections sont désormais consultables dans une même salle de lecture, quel que soit leur support. La contemporaine continue de placer la notion de *matériaux pour l'histoire* au centre de sa réflexion sur l'adaptation de sa politique documentaire aux évolutions induites par le numérique natif, et au centre de sa politique de recherche. Elle participe à des programmes de recherche, en suscite parfois, autour de ses collections et de ses expositions. En tant que membre fondateur du Labex « Les passés dans le présent », porté par l'université Paris Nanterre, elle participe à des projets de recherche dans le champ des humanités numériques : le projet ModOAP (Modèles et Outils d'Apprentissage Profond) poursuit des objectifs voisins de ceux de BaOIA et associe l'équipe de recherche du CSLF et la BnF, également partenaires du projet BaOIA.

RENOUVELER ET DIVERSIFIER L'EXPLOITATION DES COLLECTIONS NUMÉRISÉES

Le projet de recherche Boîte à Outils d'Intelligence Artificielle [2] a été soumis à CollEx-Persée début 2020. Il s'agit de développer, à partir de corpus numérisés de La contemporaine et de la BnF, des outils d'exploration et d'analyse de corpus massifs, avec une focale particulière sur l'exploitation de l'image, de les documenter et de les mettre à disposition des chercheurs pour une réutilisation ultérieure sur d'autres ensembles de données. Le projet offre l'opportunité de diversifier et de renouveler l'exploitation des collections numérisées, au-delà de leur mise à disposition du public dans les bibliothèques numériques et à travers des modalités de valorisation comme les expositions virtuelles, dans une dimension de recherche active. Les corpus utilisés sont de nature différente afin d'exploiter textes et images : des guides touristiques de la BnF ; un fonds d'affiches françaises et internationales de La contemporaine ; des dépêches d'agences de presse soviétiques provenant des fonds de La contemporaine ; un corpus de presse illustrée et d'estampes satiriques (BnF, La contemporaine, bibliothèque de Heidelberg).

À partir de ces corpus ont été développés des outils correspondant à plusieurs objectifs : constitution des corpus, étude du contenu des documents, enrichissement et visualisation des données. Les outils se présentent sous la forme de *notebooks* Python et s'utilisent *via* l'environnement Google Colab. Actuellement, les outils sont hébergés sur GitHub, et sont accessibles depuis le site du projet BaOIA, où l'on trouve des tutoriels qui précisent leur utilisation et des exemples de réalisations tirés des tests effectués sur les corpus.

PLUSIEURS SÉRIES D'OUTILS COMPLÉMENTAIRES

La première série d'outils est consacrée à l'extraction des données et à leur transformation pour les rendre exploitables. Des scripts permettent ainsi d'extraire

[1] https://argonnaute.parisnanterre.fr

[2] https://baoia.huma-num.fr

depuis Gallica des documents au format texte (s'il s'agit d'un document océrisé) ou image (JPEG ou TIFF). Un *scraper* développé pour la bibliothèque numérique d'Heidelberg est réutilisé pour récupérer le texte brut océrisé, mais aussi les illustrations et les métadonnées. Un autre outil d'océrisation de JPEG permet d'en extraire du texte brut. Grâce à ces différents outils d'extraction ou de transformation vers un autre format, développés ou adaptés pour le projet, les chercheurs pourront constituer des corpus de textes ou d'images exploitables par une deuxième série d'outils, développés quant à eux pour étudier le contenu d'un document et enrichir ses métadonnées. Un outil de reconnaissance des entités nommées permet ainsi, à partir d'un fichier texte, de repérer en plusieurs langues des personnes, lieux, organisations, évènements ou œuvres d'art, et d'effectuer des calculs statistiques sur les résultats obtenus (totaux, calcul de proportions). L'enrichissement des métadonnées est rendu possible à l'aide d'un outil qui récupère des informations sur ces entités par une requête *via* Wikidata. Enfin, la dernière série d'outils permet une visualisation des données en les cartographiant. La première cartographie s'effectue par requête Wikidata pour trouver des coordonnées géographiques à partir d'une liste de lieux. Les outils permettent également de créer des cartes interactives à partir du type de lieu (monument, ville, etc.) et de tracer des parcours entre les différents lieux. Ces outils sont utilisables indépendamment les uns des autres, selon les objectifs et besoins des chercheurs.

UNE OUVERTURE VERS DE NOMBREUSES DIMENSIONS EXPLORATOIRES

À l'heure actuelle, les outils sont développés mais leur exploitation sur les corpus et la documentation sur leur utilisation ne sont pas achevées : il s'agit de la deuxième phase du projet, toujours en cours. Toutefois, le site du projet donne déjà un aperçu des possibilités de leur utilisation combinée[3] : à partir du corpus des guides touristiques, il présente les différentes opérations réalisables, de l'extraction des documents à la visualisation des données par une cartographie. Les outils créés dans le cadre de BaOIA ont aussi été appliqués à un corpus de romans scolaires de la BnF, exploité par le programme ModOAP[4]. Ils ont permis de réaliser une cartographie des lieux évoqués dans ces manuels. Un autre exemple d'application est visible dans l'exposition *Elie Kagan, photographe indépendant,* en cours à La contemporaine : les outils d'analyse ont permis d'explorer la publication des clichés du photographe dans la presse, leurs occurrences et leurs réutilisations, et de proposer aux visiteurs un outil de visualisation des résultats (cartographie des sujets photographiés, exploration des thèmes récurrents).

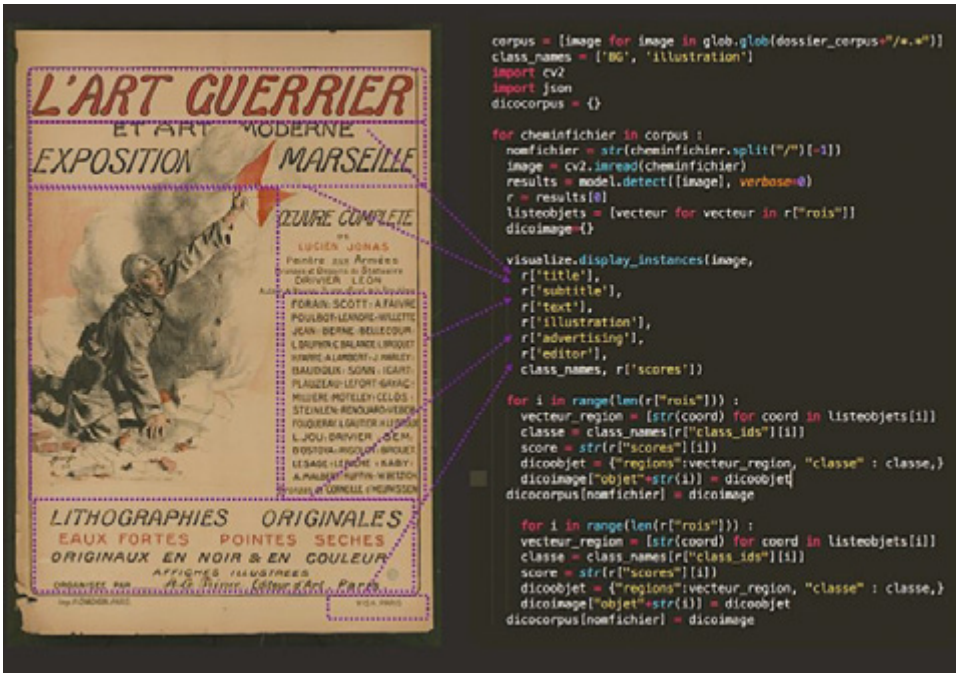
S'il est trop tôt pour dresser le bilan d'un projet en cours et de mesurer pleinement les capacités des outils développés et leurs apports à la recherche, BaOIA confirme une dimension exploratoire riche d'enseignements : exploration de corpus numérisés dont on perçoit des dimensions encore inexploitées ; exploration de dynamiques de travail renouvelées qui mettent à profit les compétences croisées de chercheurs, d'ingénieurs et de bibliothécaires autour de projets expérimentaux, dans des ateliers où les objectifs des uns et des autres sont discutés et mis en œuvre, les outils testés collectivement, adaptés et expliqués pour devenir aussi accessibles que possible ; exploration, enfin, de nouvelles modalités d'exposition et d'exploitation des données, pour la recherche et pour le grand public à travers des initiatives comme celle de la borne interactive qui clôt l'exposition Elie Kagan. Les potentialités d'exploitation par les humanités numériques des données conservées par les bibliothèques sont vastes, et nous n'en sommes sans doute encore qu'à une exploration de surface.

Cécile Tardy
Directrice-adjointe de La contemporaine
cecile.tardy@lacontemporaine.fr

Chloé Jean
Responsable de l'informatique documentaire
chloe.jean@lacontemporaine.fr

[3] https://baoia.huma-num.fr/contact/tutoriel-complet-de-lextraction-documentaire-a-la-cartographie

[4] https://modoap.huma-num.fr



La Contemporaine - Analyse d'une Une de presse