Projet BaOIA « Boîte à Outils d'Intelligence Artificielle »

Bilan scientifique

Durée : 23 mois (novembre 2020-septembre 2022)

Etablissement porteur : La contemporaine – Université Paris Nanterre

Coordination scientifique : Cécile Tardy (La contemporaine) et Julien Schuh (Université Paris Nanterre – Centre des Sciences des Littératures en langue Française - CSLF)

Budget total : 187 100 € dont 80 400 € de subvention.

Rappel des objectifs

Le projet vise à créer un ensemble d'outils dédiés à l'apprentissage profond pour les SHS, comprenant des modèles pouvant être aisément adaptés à différents usages (classification d'images, reconnaissance de structures, détection de similarités…), réutilisables par les projets de recherche concernés et, au-delà, à l'ensemble de la communauté des chercheurs et des professionnels de la documentation.
Ces outils seront accessibles à tous depuis un site web créé dans le cadre du projet.

Documentation et outils techniques

- **Outils**

Les outils développés dans le cadre du projet prennent la forme de notebooks Google Colaboratory. Ce service basé sur Jupyter Notebook permet l'exécution de notebooks (interfaces de programmation permettant à un utilisateur d'exécuter un script de manière simplifiée avec la présence de documentation) sans qu'il soit nécessaire d'installer un logiciel (fichiers ipynb). Une aide technique a été rédigée sur le site du projet afin de faciliter leur utilisation[1]. Google Colab a été choisi pour le lien avec un drive où stocker les fichiers de travail plus facilement (Google Drive), et pour la mise à disposition d'un processeur graphique qui réduit le temps d'exécution de certaines tâches.

- **Site internet**

Le site internet[2] du projet est structuré de la manière suivante : une page d'accueil présente le projet, et cinq onglets donnent accès respectivement à des informations pratiques sur le projet, aux descriptions des corpus mobilisés, aux réalisations effectuées, aux tutoriels de certains outils et, enfin, aux actualités du projet sous la forme d'articles de blog (ateliers, présentations).
Il garantit la description et le suivi du projet ainsi que la prise en main des outils.

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

- **Dépôt GitHub**

Les outils développés ont été versés dans un dépôt GitHub[3] dédié au projet. Un tutoriel complet de l'utilisation des outils développés pour l'étude du corpus de guides de voyage est également disponible.
La structure des dépôts GitHub permet pour chaque dossier et sous-dossier de rajouter une notice descriptive de leur contenu. L'utilisateur peut donc facilement se déplacer à travers les différents *repositories* GitHub et trouver ce qu'il recherche.
Ces outils peuvent être réutilisés dans le cadre d'autres projets.

Les outils du projet BaOIA

Quatre corpus avaient été sélectionnés pour le projet BaOIA, pour permettre des recherches exploratoires différentes. Tous ont été abordés. Cependant, deux ont permis des expérimentations et développements d'outils plus poussés, et les deux sont restés à des phases plus exploratoires.

1. Guides de voyages numérisés par la BnF

Pour ce corpus numérisé issu de Gallica, différents types d'outils ont été créés qui peuvent être réutilisés sur tout autre document disponible sur Gallica dont la structure est proche de celle des guides de voyage.

- **Extraction de données**

Les outils développés assurent l'extraction des pages d'un document sur Gallica en format jpg, la récupération du texte océrisé du document (si le document est océrisé) et enfin, l'extraction des pages du document en format IIIF. Toutes ces informations sont récupérées via des requêtes à l'API de Gallica.
Un outil a également été développé afin d'océriser des fichiers jpg grâce au logiciel de reconnaissance optique de caractères, Tesseract. Par exemple, dans le cas de guides de voyage non océrisés par Gallica, il sera possible de récupérer ses pages en format jpg puis d'utiliser ce dernier script pour récupérer le contenu textuel des documents.

- **Étude de contenu d'un document et enrichissement de données**

Si le texte d'un document est récupérable dans un fichier txt, un outil permet ensuite de repérer les différentes entités nommées (personnes, lieux, organisations, évènements, œuvres d'art) présentes dans ce texte et d'effectuer des calculs statistiques sur les résultats trouvés (totaux, calcul de proportion). Un des outils de reconnaissance des entités nommées utilise la bibliothèque Python de traitement automatique des langues, SpaCy.
Un autre outil récupère des informations sur les lieux ou personnes grâce à des requêtes via la base de connaissance Wikidata.

- **Visualisations cartographiques**

Un outil permet la création d'une carte à partir d'une liste de lieux dont les coordonnées géographiques sont récupérées via des requêtes à Wikidata. Un autre outil génère une carte

[3] https://github.com/baoia.

interactive où l'utilisateur peut choisir de faire apparaître tel ou tel lieu suivant son type (monument, église, ville, etc.). Enfin, deux derniers outils sont développés afin de faire émerger un ou deux parcours entre différents lieux identifiés sur une carte.

2. Fonds d'affiches numérisées de La contemporaine

- **Métadonnées**

Le travail effectué a essentiellement porté sur les métadonnées des affiches numérisées qui représentent un corpus de 26 712 fichiers jpg. Dans un premier temps, les métadonnées issues des instruments de recherche Calames (description par lot et à la cote) et les métadonnées issues d'un fichier d'inventaire des affiches sur tableur (description par lot) ont été rassemblées. Ensuite, à l'aide des fichiers jpg, les affiches numérisées ont été repérées et isolées pour constituer le corpus de travail. Un travail de normalisation et d'harmonisation a été réalisé, en privilégiant tourjours les informations tirées des instruments de recherche Calames, plus fiables. Deux logiciels de gestion de données de masse ont été utilisés pour aider à l'harmonisation des termes, Dataiku et OpenRefine.
Ces métadonnées d'origine ont ensuite été enrichies. D'une part, l'océrisation des affiches via l'API Google Vision a permis de récupérer la langue du texte de l'affiche lorsqu'elle n'était pas identifiée ainsi que le contenu textuel de l'affiche. D'autre part, grâce à des bibliothèques Python spécialisées, les doublons ainsi que les images similaires à une image donnée ont pu être récupérés.
Un fichier final de métadonnées des affiches a ainsi pu être établi : identifiant Calames, cote, nom du fichier jpg issu de la numérisation, titre Calames, numéro de pochette/tube, date, date de début et date de fin lors de la présence d'un intervalle de date, pays/aire impliqués, langues, taux de fiabilité de la langue détectée lorsqu'elle est issue de l'océrisation, description des sujets des pochettes/tubes, thématiques, sujets, origine, description Calames, auteurs, commanditaires, producteurs/éditeurs, imprimeurs, lien du fichier dans l'Argonnaute, dimensions réduites et entières, texte océrisé, doublons et enfin, images similaires et distances calculées entre l'image de référence et l'image similaire repérée. Les fichier jpg du corpus d'affiches numérisées ont été vectorisés grâce à un script réalisé dans le cadre du projet ModOAP[4], porté par le Labex « Les passés dans le présent ». Ces informations ont été rassemblées dans plusieurs fichiers json qui renseignent pour chaque fichier jpg du corpus son vecteur.
Ce fichier de métadonnées a été versé dans la base de données du CSLF. A partir de cette base, une plateforme ouverte a été créée, rendant possible l'interrogation de l'ensemble des métadonnées des affiches numérisées à partir d'une seule interface et facilitant le travail de recherche sur le corpus[5].
Cette plateforme permet de rechercher des affiches de manière multi-critères, en utilisant les métadonnées produites par le projet (océrisation des textes, dates, lieux, auteurs, commanditaires…). Elle inclut également les données de similarités, permettant de parcourir les affiches selon des rapprochements graphiques inédits, et des outils de visualisation des données statistiques dans la base. Ce site est en cours de finalisation et sera accessible fin 2022.

- **Visualisations**

Plusieurs outils de visualisations ont été utilisés pour ce corpus notamment à l'occasion d'un atelier organisé autour des affiches numérisées et des usages potentiels des outils d'analyse. Le visualiseur

[4] https://github.com/MODOAP/detection-doublons-images/blob/main/detection_doublons.ipynb.
[5] https://baoia.sociodb.io/affiches (en accès restreint le temps de sa finalisation).

PixPlot[6] a été utilisé afin d'établir une pré-classification du fonds d'affiches numérisées. Il permet de représenter l'ensemble des images d'un corpus : plus deux images sont proches, plus elles sont similaires. De même, il génère des clusters, groupes thématiques cohérents d'images, dont il s'agit ensuite d'identifier la particularité commune. Cet outil a révélé l'importance des affiches textuelles dans l'ensemble du corpus et ainsi encouragé l'océrisation des affiches. Un lien est disponible sur le site du projet pour accéder à la visualisation PixPlot des affiches antérieures ou égales à l'an 1945 (celles libres de droit) : https://baoia.huma-num.fr/pixplot1945/output/index.html
D'autres visualisations ont été effectuées avec le logiciel gratuit Tableau Public. Ces visualisations, ayant été réalisées au début du processus de regroupement des métadonnées, ne portent pas sur l'ensemble du corpus d'affiches numérisées mais sur un échantillon d'environ 11 000 fichiers. Les liens donnant accès à ces visualisations en ligne sont renseignés sur le site du projet :

- Un premier tableau comportant des visualisations sur la date, la langue, le lieux de production et les pays concernés.
- Un second tableau concernant les auteurs, les commanditaires, les imprimeurs et les sujets abordés par les affiches.

3. Dépêches d'agences de presse soviétiques

Deux titres de périodiques ont été explorés, *l'Annuaire URSS* et le *Digest Spoutnik*, tous deux numérisés. Comme pour les affiches, un travail de récupération des métadonnées disponibles sur ces fonds a été réalisé à partir du catalogue de La contemporaine. Ces documents étant océrisés au format pdf, un notebook a été créé afin de récupérer le texte issu de l'océrisation dans un fichier de métadonnées qui contiendrait le texte océrisé ainsi que la page jpg correspondant au texte. Ce notebook a permis la création d'une multitude de fichiers de métadonnées, un par volume, avec le numéro de page, le fichier jpg correspondant à la page et le contenu textuel de la page.
Ces données ont été intégrées dans un visualiseur hébergé par Numapresse, partenaire du projet, permettant de chercher dans les revues en plein texte :
http://www.numapresse.org/exploration/presse_russe/query_calendar.php
et
http://www.numapresse.org/exploration/presse_russe/query_text.php

4. Presse illustrée et estampes satiriques

Dans le cadre de ce corpus, un outil a été créé permettant l'extraction de périodiques de la bibliothèque numérique de l'université d'Heidelberg. Pour chaque volume de périodique est récupéré : le texte brut océrisé, des fichiers xml-alto de métadonnées, les illustrations en format jpg et les métadonnées des images IIIF[7].

Participation à des projets voisins

Comme annoncé dans le dossier d'appel à projet, certains programmes du Labex « Les passés dans le présent » présentaient de nombreux aspects complémentaires du projet BaOIA, ce qui a permis la réutilisation et le développement conjoint d'outils, en particulier avec le projet ModOAP.

[6] https://dhlab.yale.edu/projects/pixplot/.
[7] https://baoia.huma-num.fr/outil-dextraction-de-documents-de-la-bibliotheque-numerique-heidelberg/

- **Projet ModOAP**

L'équipe de BaOIA a ainsi participé à l'élaboration de plusieurs notebooks pour le projet ModOAP[8] dédié aux romans scolaires. Un premier outil est utile pour extraire la table des matières, quand elle existe, de documents numérisés sur Gallica. Deux autres permettent de récupérer toutes les informations disponibles sur Wikidata à propos d'un mot clé pour l'un et d'un lieu pour l'autre. Ce dernier récupère les métadonnées nécessaires pour d'autres outils de création de cartes de chaleur statistiques par région et par département.

- **Projet MonumenTAL**

En appui au programme MonumenTAL[9], également soutenu par le Labex « Les passés dans le présent » et portant sur l'évolution des titres et appellations des monuments antiques, deux notebooks ont été développés. Le premier permet, à partir d'un terme de recherche, de récupérer des données textuelles et iconographiques dans les bases de données Joconde, Europeana et Gallica. Par exemple, pour les termes de recherche « Vénus de Milo », les résultats de recherche sont toutes les métadonnées des œuvres présentes sur ces bases de données comportant ces deux termes. Le notebook télécharge aussi les images reliées à ces termes de recherche et crée une cartographie où sont représentés les lieux de conservation des œuvres récupérées suite à la requête initiale. Le second outil récupère des informations sur des artistes à partir d'un fichier Excel qui les recense. Cet outil utilise la base de connaissance Wikidata pour récupérer des informations comme les dates et lieux de naissance, de décès ainsi que le genre.

Evénements autour du projet

Les comptes-rendus de ces événements sont disponibles sur le site internet du projet via l'onglet « Actualités du projet ».

1. **Séminaires et atelier**

La crise sanitaire a compliqué la tenue des ateliers initialement prévus. Trois séances d'atelier se sont tenues :

- Atelier le 2 décembre 2020 autour du corpus des guides de voyage du projet BaOIA et des manuels scolaires du projet ModOAP. Pendant cette séance, les différents corpus ainsi que des exemples d'outils adaptés à leurs problématiques ont été présentés.
- Atelier le 11 février 2021 de présentation et de discussion autour des outils élaborés depuis la séance précédente.
- Atelier le 31 mars 2022 à La Contemporaine autour du corpus d'affiches : définition des objectifs de valorisation et d'exploitation du fonds des affiches numérisées.

[8] https://modoap.huma-num.fr/.
[9] http://passes-present.eu/fr/monumental-monuments-antiques-et-traitement-automatique-de-la-langue-44335.

2. Présentations du projet

Le 13 juin 2022 a eu lieu une présentation du projet devant le comité de direction du Centre national d'arts plastiques (Cnap) à La Défense. Le but de cette dernière était de présenter le projet afin d'initier des pistes de réflexions sur les possibilités d'exploitation des corpus du Cnap.

Le 28 juin 2022 s'est déroulée une table-ronde intitulée "IA et bibliothèques" au Pixel (SCD Université Paris Nanterre), durant laquelle le projet BaOIA a été présenté.

Publication : C. Jean et C. Tardy, « BaOIA : une « boîte à outils » pour explorer les corpus numérisés », in *Arabesques* n°105, 2022, *Humanités numériques. Une renaissance 3.0* : https://publications-prairial.fr/arabesques/index.php?id=2859