E-Col+

ENRICHIR ET VALORISER LE CORPUS DE SPECIMENS NUMERISES DU MUSEUM NATIONAL D'HISTOIRE NATURELLE, RECHERCHE & NAVIGATION VIA L'ANNOTATION D'IMAGES PAR FOUILLE DE TEXTES ET ANALYSE D'IMAGES

Informations clés sur le projet

Financement : ANR PIA3 Equipex ; Institution porteuse : MNHN
Coordinateur général : Pierre-Yves Gagnier, MNHN, 57, rue Cuvier, CP 43 - 75005, Paris, pierre-yves.gagnier@mnhn.fr
Coordinateur WP4 : Eric Chenin, IRD / UMMISCO, 32 av. Henri Varagnat, 93140 Bondy, France, eric.chenin@ird.fr
Durée du projet : 8 ans : démarrage Juin 2021
Budget IRD : 545 K€ ; Budget total : 4.850 K€

Institutions partenaires

IRD / UMMISCO ; MNHN ; Sorbonne Université ; CINES ; Université de Bourgogne ; CNAM

Contexte

Le Muséum National d'Histoire Naturelle a entrepris depuis une quinzaine d'années de numériser ses collections de spécimens, jouant un rôle pionnier dans ce domaine au niveau mondial. L'Herbier est entièrement numérisé en haute définition, ainsi qu'une partie des collections de zoologie. L'objectif général du projet E-Col+ est double :

1. enrichir le corpus de spécimens numérisés, particulièrement avec des images 3D surfaciques et tomographiques, pour aboutir à environ 10 Millions d'images 2D et 3D pour un volume de 2,3 Po ;
2. valoriser le corpus en le dotant d'une interface de recherche et navigation efficace appuyée sur une annotation des images produite par des techniques d'IA en fouille de textes, en analyse d'images et en représentation des connaissances.

C'est un projet financé par l'ANR dans le cadre du PIA3 Equipex, pour 4,85 M€ au total, dont 545 K€ pour UMMISCO. Sur 8 ans (4 ans de développement, plus 4 ans de mise en exploitation).

UMMISCO est en charge du WP4, qui regroupe toutes les applications des techniques d'IA et le développement de l'interface de recherche et navigation dans le corpus.

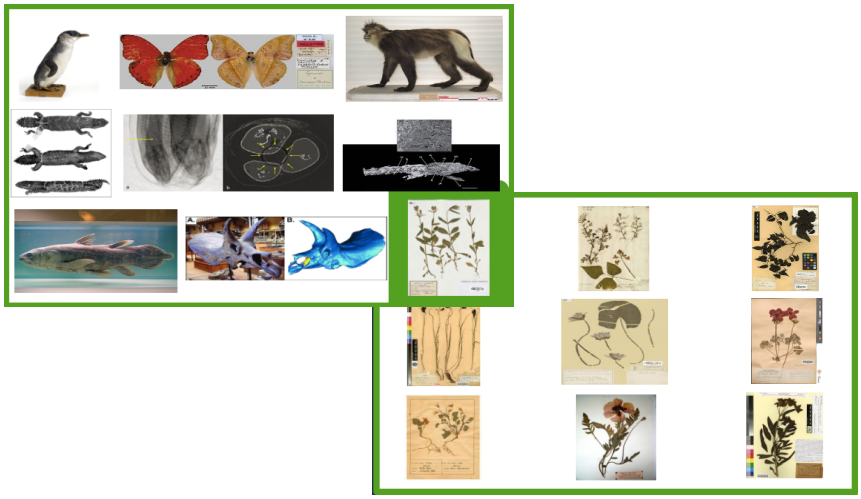

Figure 1 : Interface de recherche-navigation intuitive : recherche de proche en proche & exploration du corpus, utilisant l'annotation des images et l'indexation du corpus, sans champ de saisie, ni menu déroulant

Objectifs

Les techniques d'IA sont mises en oeuvre dans E-Col+ avec plusieurs objectifs spécifiques.

Pour les images 3D tomographiques, ces techniques permettent d'exploiter au mieux les captures scanner, en améliorant la définition à l'aide de modèles, ou en exploitant les tranches scannées pour calculer de nouvelles tranches, intermédiaires ou sous d'autres angles.

Pour l'annotation des images et l'indexation du corpus, l'IA est utilisée en fouille de textes naturalistes pour extraire des caractères descriptifs et des relations entre espèces et caractères, et elle est utilisée ensuite en analyse d'images pour extraire les caractères visuellement reconnaissables.

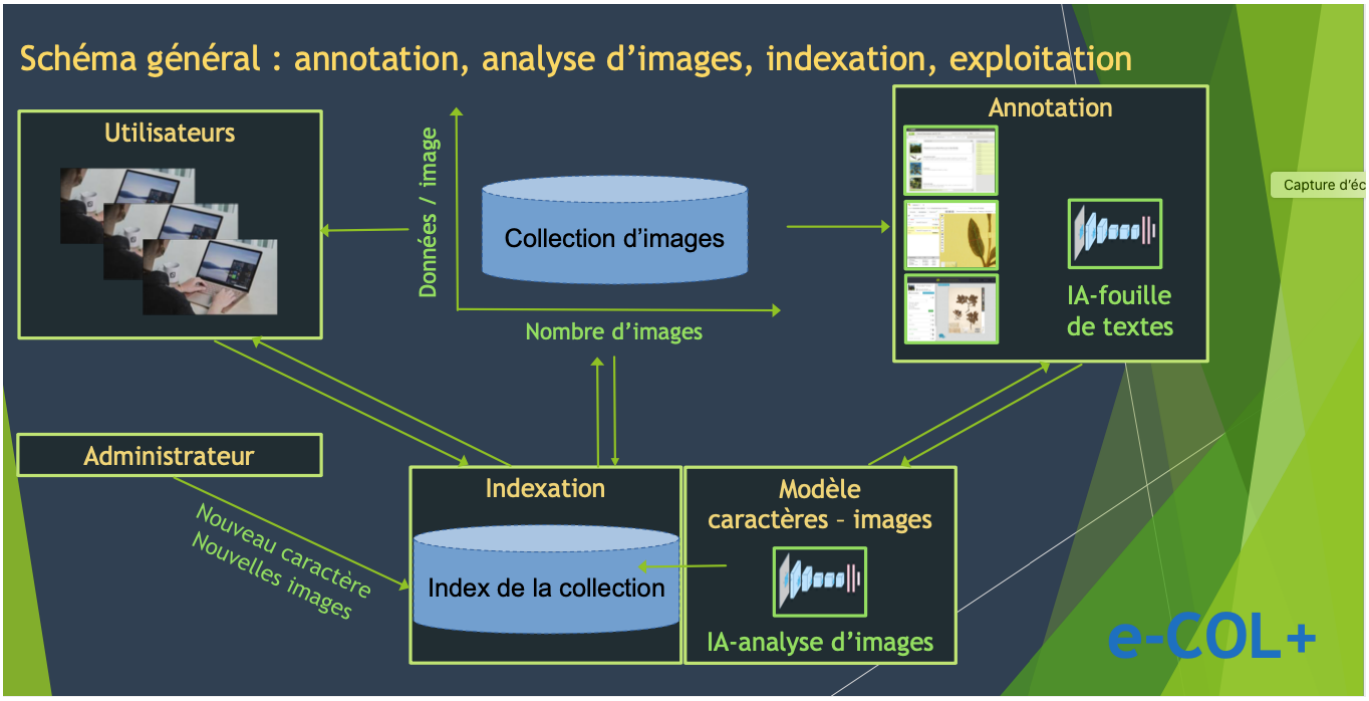

Figure 2 : Schéma de fonctionnement du système global, qui prévoit non seulement l'annotation et l'indexation du corpus d'images ainsi que la recherche et la navigation en son sein, mais aussi l'évolution du corpus indexé et de l'interface, pour intégrer au fil du temps de nouvelles images et de nouveaux caractères.

Organisation et résultats attendus

L'annotation des images de spécimens s'effectue de façon manuelle, à l'aide de l'outil Annotate et du dispositif de science citoyenne Les Herbonautes, elle est enrichie par les relations entre espèces et caractères extraites par fouille de textes, et elle est complétée en effectuant un balayage complet du corpus avec le réseau de neurones entraîné à la reconnaissance de caractères visuels.

L'annotation des images est utilisée pour indexer la totalité du corpus, et l'interface de recherche et navigation s'appuie sur cette indexation. Une approche intuitive sera privilégiée pour l'interface : une sélection aléatoire d'images est présentée à l'utilisateur, qui peut désigner l'image la plus proche de ce qu'il cherche, ou demander une nouvelle sélection aléatoire. L'indexation du corpus permet ainsi de rechercher de proche en proche, ou d'explorer le corpus, voire d'alterner entre recherche et exploration.

Le développement du dispositif sera effectué de manière itérative, en commençant avec les images, les textes et les annotations déjà disponibles - l'herbier, plusieurs Flores et les caractères déjà renseignés avec Annotate - ; puis en augmentant progressivement l'ensemble d'images, de textes, et d'annotations manuelles à mesure qu'ils deviennent disponibles. Cette approche itérative dépasse la seule phase de développement : le système global sera construit de telle sorte que durant son exploitation, il sera à tout moment possible d'intégrer de nouvelles images ou de nouveaux caractères, qui enrichiront le corpus et affineront l'interface de recherche et navigation.