BIBLIOTHÈQUES
DU MUSÉUM

BILAN DU PROJET DATAPOC.MNHN.FR (2018-2019)

RAPPEL DU CONTEXTE ET DE L'ORIGINE DU PROJET

Le projet « Datapoc.mnhn.fr » a été lauréat en 2019 de l'appel à projet 2018-2019 Collex-Persée.

Ce projet vise à évaluer la faisabilité et l'opportunité de construire un référentiel « personnes » commun à l'ensemble des services et types de collections et de données produits et gérés au Muséum national d'Histoire naturelle (MNHN). Il s'agit de consolider les données et les bases de données existantes, mais aussi de permettre aux chercheurs de croiser, lier et exploiter des données qu'il leur est difficile d'apparier aujourd'hui compte tenu de la dispersion des applications.

Le projet prend la forme d'une preuve de concept établie à partir d'un corpus limité de noms de personnes (environ 500). Il consiste en :

- Tester la performance de technologies et de méthodes de traitement et de rapprochement de données à partir des données disponibles,
- Réaliser un prototype d'interface orienté chercheur pour la visualisation humaine et la réutilisation par des machines des résultats et des données exposées,
- Émettre des préconisations pour un passage à l'échelle et une industrialisation (à la fois en volume et en type de données).

Un travail en amont, démarré en 2016, a permis de constituer un corpus d'environ 500 noms de naturalistes. Un ensemble de taxonomistes dits « historiques » ainsi que des chercheurs contemporains composent cette liste. D'abord construite à l'aide d'une monographie : Jaussaud, P., Brygoo, É.-R., Du jardin au Muséum en 516 biographies. Publications scientifiques du Muséum, Paris, 2004. 632p. cette liste a ensuite été complétée en équilibrant les différentes disciplines. Une fois le corpus constitué des noms, prénoms, dates de naissance et de mort (lorsqu'elles existent) et les principales disciplines, les identifiants IdRef (identifiants et référentiels de l'Enseignement supérieurs) ou numéros PPN correspondants ont été ajoutés.

Un chantier de vérification de ces identifiants a été mené en interne à partir d'une extraction fournie par l'Abes pour constituer le fichier étalon sur lequel le projet s'est basé.

La subvention de ce projet a permis le lancement d'une procédure de marché public afin d'engager une société de développement informatique pour développer la preuve de concept. Le budget prévisionnel pour cette PoC (Proof of Concept ou démonstrateur de faisabilité) se situait dans une tranche comprise entre 50 000 et 60 000 €.

Procédure de marché public et choix du candidat

La publication du marché public de prestation informatique n°I8M53 « Visualisation des données associées aux principaux naturalistes du MNHN (projet datapoc) » a été lancée à partir du mercredi I2 décembre 20I8 jusqu'au mercredi I6 janvier 20I9. Trois candidats ont postulé à ce marché. Après dépouillement des offres, la société Logilab, spécialisée dans les traitements de données à l'aide des technologies du web sémantique, a été sélectionnée.

Il est à noter que chacun des candidats demandait un budget largement supérieur à celui proposé (les offres se situaient entre 82 000 et I20 000 €). La phase de négociation n'a pas eu lieu parce que la marge de négociation était faible et que l'équipe projet du MNHN a jugé peu pertinent de faire décaler le calendrier par rapport à ce qu'il pourrait obtenir en termes de prix. En outre, les ressources propres de la Direction des bibliothèques et de la documentation étaient suffisantes pour abonder le projet. Le calendrier prévisionnel a été maintenu et la réunion de lancement a eu lieu le II avril 20I9.

Equipes

GROUPE DATA (MNHN)	PRESTATAIRE INFORMATIQUE (LOGILAB)
Chloé Besombes (chef de projet, DBD) Cécile Callou (UMS 3466 BBEES) / Simon Chagnoux (DSI) / Emmanuel Côtez (Publications scientifiques) / Pierre-Yves Gagnier (DGD-C) / Gildas Illien (DBD) / Cindy Lim (DBD) / Thomas Milon (UMS PatriNat) / Marc Pignal (RecolNat)	Pierre Choffé (chef de projet) Olivier Cayrol (directeur général adjoint Logilab) Fabien Amarger / Simon Chabot (développeurs) Elouan Martinet (stagiaire)

- Consultation d'experts des bases de données du MNHN, lors de la phase de développement
- Sollicitation de collègues gestionnaires de bases de données taxonomiques pour les tests utilisateurs : agents du MNHN et des institutions membres du réseau RecolNat.

Méthodologie adoptée

La méthodologie adoptée pour mener le projet était apparentée à une méthode de développement agile, alternant : - des discussions avec des utilisateurs métier (principalement gestionnaires de bases de données et chercheurs) afin de définir des cas d'usages, - des phases de développements et retours vers les utilisateurs sous forme de tests, - avant de lancer une nouvelle phase de développements. Des tests d'utilisateurs extérieurs au MNHN étaient également prévus en fin de projet, afin d'enrichir les développements. Le cahier des charges avait volontairement été rédigé de manière très ouverte, afin de permettre aux candidats de développer leur créativité. Cependant, étant donné le calibrage insuffisant du budget prévisionnel par rapport au temps de développement nécessaire, la phase de développement de la PoC a été écourtée par rapport au calendrier prévisionnel (9 mois au lieu de I2 à partir de la notification du marché). Seules des modifications ponctuelles et à la marge ont pu être apportées à la PoC suite aux tests utilisateurs.

Réajustement du périmètre des bases de données incluses dans le projet

La densité et la diversité des bases de données identifiées dans le projet de départ a contraint l'équipe de développement et l'équipe de suivi du projet à renoncer à l'intégration de certaines d'entre elles. Ces renoncements ont permis d'une part de tenir le calendrier et d'autre part de développer des fonctionnalités plus poussées pour les utilisateurs, notamment en termes de visualisation des données.

Ainsi, les bases effectivement intégrées sont les suivantes :

X BasExp: description des expéditions menées par le Muséum depuis les années 80.
- BHL: Biodiversity Heritage Library. Numérisations de publications
X Calames: catalogue d'archives, manuscrits, objets d'art…
- Collections numérisées: fonds numérisés de la Direction des bibliothèques et de la documentation.
- Bases de données des collections d'histoire naturelle: 22 bases de données taxonomiques du Muséum
- HAL – Hyper Articles en Ligne: archives ouvertes des publications
X InDoRES: Inventaire des Données de la Recherche en Environnement et Société.
- INPN: données d'observation d'espèces. Accessibles depuis l'API de GBIF.org
- Muscat: catalogues des imprimés
- Publications scientifiques: revues et monographies édités par le Muséum
X RefBiblio: contient des fiches bibliographiques d'articles relatifs aux expéditions ou aux spécimens du Muséum
- Wikidata : pour aider à la construction de la plateforme

Rôle et choix des identifiants pérennes

Le travail de consolidation des IdRef liés aux noms de personnes constituant le corpus a permis de d'intégrer cet identifiant à l'algorithme d'alignement. Cela a, entre autres, permis d'aligner le corpus des noms avec Wikidata, et de récupérer et d'afficher une sélection d'identifiants pérennes liés à la personne, afin de constituer sa « carte d'identité ». On trouve donc, associés à chaque personne du corpus, des identifiants liés au domaine bibliographique (BHL, BnF, IdRef, iSNi, ORCiD, ViAF), des identifiants liés au domaine taxonomique (Harvard Botanists Index, iPNi, Zoobank) ; ainsi que l'identifiant Wikidata.

Synthèse et calendrier des objectifs du projet et de leur réalisation

Les objectifs liés à la phase de développement et de déploiement ont été retranscrits dans le tableau suivant.

Objectifs fixés par le CCTP	Objectifs prévus par le calendrier prévisionnel	Dates de l'obtention des résultats	Remarques
Livraisons de logiciels. **Livrable I : jeux de données structurées &** **Livrable 2 : Site « People of Collections » et API**	Septembre 2019	Septembre 2019	Priorité aux données taxonomiques. Les données issues de BasesExp, InDoRES et de Calames n'ont pas été ajoutées. En revanche, les publications du Muséum disponibles sur BHL et les données ont été confortées avec les données de Wikidata.
Livrable 3 : Rapport sur les mécanismes de structuration	Août 2019	Première partie sur les alignements en Octobre 2019 et Décembre 2019	Une première partie de ce rapport sur les alignements avait été demandé pour la préparation de la communication au « Biodiversity Next » à Leiden (Pays-Bas).
Livrable 4 : Préconisations sur le passage à l'échelle	Janvier 2020	Décembre 2019	

Déploiement public et mise en production opérationnelle du logiciel	Octobre 2019	Février 2020	Retard justifié notamment par une période pendant laquelle plateforme n'était pas accessible : lors des sessions de tests utilisateurs) et un souci technique lié aux outils du prestataire. De plus, d'importants échanges ont eu lieu entre la DSI du Muséum et l'équipe de Logilab pour permettre le déploiement de l'outil dans les serveurs de l'établissement, justifiant ce retard de déploiement.
Tests utilisateurs : **I. Identification et mobiliser un réseau de testeurs** **2. Organisation des études d'usages**	Réseau de testeurs : Juillet 2019 Sessions de tests : Juillet-Décembre 2019	Identification et mobilisation d'un réseau de testeurs extérieurs au groupe Data : Septembre 2019 Sessions de tests : Octobre 2019 à Janvier 2020	Définition de trois vagues de tests : - Groupe Data ; - Experts des bases de données ; - Extérieurs : étudiants du MNHN et publics de la Bibliothèque centrale. La 3e vague n'a finalement pas été menée. Suite à la 2e vague, puisque la plateforme n'a pas subi de modifications majeures, organiser une 3e vague aurait pu apporter des remarques redondantes. Enfin, la période de mouvements sociaux en fin d'année 2019 et la période de confinement suite à la pandémie du Covid-19 en 2020 ont également empêché la réalisation de cette 3e vague. Une adresse générique a cependant été créé pour recueillir les avis des usagers.
Rapport sur les tests d'usages de l'outil	Février 2020	Février 2020	Document de synthèse sur les tests utilisateurs présent sur les serveurs en interne
Communication et valorisation du projet	Pas de dates spécifiques	Avril, Octobre et Décembre 2019	« CollExPro », Journées professionnelles Collex-Persée, Paris (4-5 avril 2019) « Biodiversity Next », Leiden (20-25 octobre 2019) « SemWebPro », Journée de présentations et de rencontres dédiées au web sémantique, Paris (3 décembre 2019)

Un accès SPARQL Endpoint a été développé, bien qu'il ne fût pas prévu dans les objectifs initiaux.

DATAPOC.MNHN.FR

Principales fonctionnalités

La plateforme datapoc.mnhn.fr présente quatre types de pages : une page d'accueil, une page « A propos », des pages « Naturaliste » et des pages « Taxonomie ».

Les pages « Naturaliste » permettent de visualiser une carte d'identité du naturaliste concerné : portrait, nom, prénom, identifiants pérennes, principaux champs d'études, bref texte de présentation (source : Wikipédia). La page présente ensuite, sous différentes rubriques, les résultats des activités du naturaliste, c'est-à-dire les données rattachées aux spécimens que celui-ci a identifiés et nommés (auteur de taxons), collectés, déterminés

ou observés et qui sont actuellement conservés dans les collections du MNHN. Viennent ensuite les éléments de bibliographie concernant le naturaliste (en tant que sujet ou auteur), ainsi que l'iconographie qui lui est attachée. Chaque item listé permet de rebondir vers la base de données source de la donnée. Dans chaque rubrique, une série de filtres permet de naviguer dans la liste de résultats afin d'affiner la recherche ; une visualisation géographique de l'activité du naturaliste est proposée en fin de page ainsi que, pour chaque rubrique, une mise en diagramme pour faciliter la visualisation des données.

Les pages « Taxonomie » se présentent de la même manière que les pages « Naturaliste » ; chacune d'entre elles s'intéresse à une famille de taxons, et permet de lister les naturalistes dont l'activité est liée à l'étude de cette famille, selon les mêmes catégories (auteurs de taxons, déterminateur, collecteur, etc…). Les mêmes types de filtres et fonctionnalités de visualisation des données sont également proposées.

Ainsi, datapoc.mnhn.fr permet par exemple à un utilisateur de générer un itinéraire et un calendrier des activités de collecte d'un naturaliste, de saisir en un clin d'œil quels sont les spécialistes d'un groupe taxonomique précis et d'accéder rapidement au texte intégral de publications scientifiques concernant un spécimen précis.

L'accès SPARQL Endpoint permet en outre d'explorer les données liées par le biais de la PoC.

Défauts et limites techniques de la PoC actuelle

Cette plateforme présente le principal défaut de ne pas pouvoir être générée de manière dynamique. Chacune des bases de données est sauvegardée localement de manière temporaire pour générer la plateforme, ce qui occasionne un délai de 6 à 8h. La mise à jour des données dans les bases de données sources met donc beaucoup de temps à être prise en compte, et la masse des données enregistrées est colossale, pour un corpus de personnes pour le moment limité à 500 noms. En outre, la plateforme telle qu'elle a été conçue ne permet pas d'en administrer facilement le contenu et la mise en page. Enfin, elle n'intègre pas l'ensemble des données initialement prévues.

CONCLUSION ET PRECONISATIONS POUR UNE EVENTUELLE SUITE DE CE PROJET

Le projet a suscité un vif intérêt auprès de la communauté comme en ont témoigné les séminaires ainsi que les tests utilisateurs. Il permet notamment de regarder autrement les données liées aux collections du MNHN, et a rapidement permis de repérer des erreurs dans les bases de données de collection, aussi bien bibliographiques que taxonomiques. Il représente donc un outil non négligeable pour la consolidation de ces bases, que ce soit par le biais de la construction d'un référentiel personnes partagé ou pour le repérage d'erreurs.

Il est également intéressant de noter que le développement de cette preuve de concept s'insère dans l'actualité des muséums européens, mobilisés pour développer des pratiques partagées autour de la description et de l'inventaire de leurs collections et des publications attachées, en envisageant notamment d'adopter des identifiants pérennes pour les personnes et les objets numériques.

Dans la perspective de donner une suite à ce projet, des pistes de développement et d'amélioration sont déjà dessinées. Une évolution souhaitable et fondamentale consisterait à revoir la construction de la plateforme afin qu'elle soit plus dynamique. Elle permettrait notamment une augmentation conséquente du volume de données, à la fois en ajoutant les bases de données précédemment écartées (BasesExp, Inventaire de données InDoRES, catalogue Calames et RefBiblio) mais en envisageant également d'augmenter significativement le corpus de noms de personnes afin de simuler un passage à l'échelle. Améliorer l'ergonomie du site pour les utilisateurs, en tenant compte des retours des tests utilisateurs, mais également pour les administrateurs, serait un point intéressant à intégrer dans un nouveau projet. La visualisation des données, la présentation des résultats méritent en effet d'être précisée ou rendue plus claire. Enfin, un nouveau projet pourrait permettre de développer et d'imaginer des fonctionnalités d'interactions des utilisateurs, qui pourraient signaler aux administrateurs des données les erreurs repérées lors de leur navigation sur l'outil.