BIBLIOTHÈQUES
DU MUSÉUM

BILAN DU PROJET DATAPOC 2.0

1. RAPPEL DU CONTEXTE ET DE L'ORIGINE DU PROJET

Le projet « Datapoc 2.0 » a été lauréat en 2020 de l'appel à projets 2019-2020 Collex-Persée.

Ce projet fait suite au projet « Datapoc.mnhn.fr », lauréat de l'appel à projets 2018-2019 Collex-Persée. Ce premier projet, mené en 2019, a abouti à la construction d'une preuve de concept qui utilise les technologies du web sémantique pour explorer les possibilités d'alignement, de liage et de publication de données utiles à la recherche taxonomique (métadonnées, numérisations…), issues de plusieurs réservoirs de données du Muséum, et associées à un corpus d'environ 500 noms de savants et de collecteurs naturalistes. Le prototype https://datapoc.mnhn.fr permet d'évaluer la pertinence d'un nouveau mode de navigation dans ces données liées. Il offre aussi des possibilités de réutilisation, notamment aux fins d'enrichissement d'agrégateurs et de référentiels.

Le projet « Datapoc 2.0 » s'inscrit dans la continuité du projet « Datapoc.mnhn.fr », dont les résultats ont incité le Muséum national d'Histoire naturelle, porteur du projet, à poursuivre les réalisations et les réflexions dans trois directions :

(1) Dans la perspective d'un passage à l'échelle, l'extension du volume de données, à la fois du nombre de données Personnes mais également du nombre de sources de données mobilisées dans le cadre du projet (inventaires d'archives, données d'observation, publications déposées dans HAL…)

(2) Les traitements algorithmiques déployés ont révélé des problèmes de qualité et de structuration dans les données sources. Ce constat motive l'intention de doter le service d'un outil collaboratif de repérage et de correction des anomalies par les chercheurs et les collecteurs. Ces fonctionnalités seraient destinées tant aux scientifiques et gestionnaires du MNHN qu'aux amateurs, dans la dynamique des sciences participatives.

(3) Il s'agira enfin de tester plus avant les mécanismes de réutilisation et d'enrichissement croisé des données en coopération avec des partenaires extérieurs : Fichier national d'entités (Abes/BnF), bibliothèque numérique BHL (Biodiversity Heritage Library), infrastructure européenne DISSCO (Distributed System of Scientific Collections).

La construction d'une nouvelle PoC dans le cadre du projet « Datapoc 2.0 » est une nouvelle étape vers la construction d'un référentiel consolidé et transverse de noms de personnes dotés d'identifiants pérennes que le Muséum souhaite intégrer dans son système d'information afin de faciliter et normaliser la production de métadonnées dans les différents univers professionnels de l'établissement.

A cette fin, le Muséum souhaite que les solutions logicielles développées en Open Source par le prestataire soient intégrées dans l'infrastructure informatique de l'établissement afin que sa DSI soit en capacité de s'approprier ces solutions et d'avoir la possibilité de les intégrer dans le futur dans le cadre d'un projet d'évolution des bases de données documentaires et naturalistes dont les contours seront définis en 2021.

2. DEROULEMENT DU PROJET

Mise en place du projet

1) Première phase de travail (juillet 2020-juillet 2021)

Le démonstrateur développé au cours du premier projet a donné des résultats de mise en rapport et de visualisation des données satisfaisants, et a suscité l'enthousiasme du groupe test d'utilisateurs. Cependant, les choix technologiques adoptés par la société n'étaient pas satisfaisants sur plusieurs points : la technologie utilisée nécessitait une montée en compétences forte des équipes du service informatique de l'établissement (DSI) ; la mise à jour des données était très laborieuse et ne pouvait être effectuée aussi souvent que nécessaire.

Pour réaliser le projet Datapoc 2.0, le choix a été fait de lancer une procédure de marché public sur la base d'un nouveau cahier des charges, afin de bénéficier de la procédure de mise en concurrence. La crise sanitaire a ralenti l'ensemble du démarrage de ce projet transverse au sein de l'établissement. Le marché a été publié le 14 février 2021 et la consultation a pris fin le 2 avril 2021. Trois candidatures complètes ont été reçues et examinées. La proposition de l'une d'entre elles était d'une excellente qualité et répondait parfaitement aux exigences du cahier des charges, mais dépassait largement le seuil budgétaire alloué à ce projet. Les autres propositions étaient plus raisonnables en termes de budget – bien que l'une d'entre elles dépassait toutefois également le budget prévu - mais ne répondaient pas totalement ni exactement aux besoins et aux exigences exprimés dans le cahier des charges.

Parallèlement à l'avancée du projet, la DSI du Muséum a été profondément remaniée avec l'arrivée d'un nouveau directeur des systèmes d'information en janvier 2021. Celui-ci porte un important projet de refonte du système informatique (SI) du Muséum, incluant la refonte du SI des collections naturalistes et des bibliothèques. Les discussions avec la DINSI (nouvelle DSI, Direction de l'Innovation Numérique et des Systèmes d'Information) autour des offres reçues dans le cadre du marché Datapoc 2.0 ont conduit à faire émerger la nécessité de redéfinir le besoin dans le cadre de ce projet, dont la réalisation comporte un fort enjeu dans le cadre de la refonte du SI du Muséum. La décision a donc été prise de déclarer ce premier marché infructueux pour redéfinition du besoin. Cette décision a été actée par la Commission d'appels d'offres du Muséum le 18 juin 2021.

2) Deuxième phase de travail et procédure d'achat innovant (août 2021- mars 2022)

L'enjeu de cette deuxième phase de travail a résidé dans un redimensionnement du cahier des charges afin que la proposition financière des entreprises candidates puisse respecter le budget provisionné par l'établissement pour la réalisation de ce projet. En outre, les discussions avec la nouvelle équipe de la DSI et les divers projets de refonte du SI Collections recherche, qui comprend notamment la construction et la mise en place d'un datahub pour les données produites au Muséum, ont conduit les membres de l'équipe projet à privilégier le développement d'une PoC à usage plutôt interne, permettant avant tout d'atteindre le passage à l'échelle en incluant tous les noms de personnes figurant dans les bases de données des collections naturalistes et documentaires, mais également de proposer des fonctionnalités permettant de faciliter le repérage et la correction de formes de noms erronées dans les bases de données sources. Enfin, la PoC doit proposer, pour chaque nom de personne, un identifiant pérenne « Muséum », aligné, lorsque cela est possible, avec des identifiants dans des référentiels externes (IdRef, Wikidata, Orcid, ISNI, VIAF). Cette réorientation des besoins a eu pour principale conséquence d'éliminer du cahier des charges la partie de construction design de l'interface de consultation, pensée au départ pour être manipulable par un utilisateur de n'importe quel niveau de connaissances. L'outil imaginé dans le cadre de ce 2^{e} cahier des charges est à destination des chargés de collection et des chercheurs du Muséum ou travaillant sur les collections du Muséum. L'interface de consultation proposée ressemblera donc à l'interface de gestion en « Back office », et sera conçue pour être mise à disposition des administrateurs de l'application et des personnes habilitées à corriger les alignements et à travailler à leur structuration (validation/invalidation). La rédaction de ce nouveau cahier des charges a été achevée en octobre 2021.

Afin d'alléger et d'accélérer le processus de choix d'un prestataire, suite à cette redéfinition des besoins, c'est le dispositif de l'achat innovant qui a été retenu d'un point de vue administratif. Ainsi, les trois prestataires qui avaient répondu au premier appel d'offres ont été recontactés à l'issue de la rédaction de ce cahier des charges redimensionné. Ils ont tous les trois fait une nouvelle proposition chiffrée et ont été reçus par l'équipe projet afin que leur proposition soit examinée. A l'issue de ces entretiens, c'est l'équipe de Mnémotix, dont la proposition restait la meilleure sur les trois, qui a été retenue. L'achat innovant a été signé fin novembre 2021, et la réunion de lancement, marquant le début du projet, s'est tenue le jeudi 6 janvier 2022.

Il est à noter que la chef de projet a quitté l'établissement en mars 2022 suite à la réussite d'un concours de la Fonction publique. C'est donc le tandem formé par la coordinatrice scientifique chercheuse, et la coordinatrice scientifique IST, accompagnées par une équipe renouvelée, qui a repris le travail sur le projet pour accompagner l'équipe de Mnémotix tout au long de sa réalisation.

Méthodologie adoptée pendant la phase de développements

L'équipe de Mnémotix a proposé à l'équipe projet une méthodologie de travail qui combine des itérations courtes (2 à 3 semaines maximum), réunissant les deux chefs de projet et éventuellement des experts sur des points précis à résoudre, et des CopilTechs (2 sur l'ensemble du projet), réunissant l'ensemble des membres du comité de pilotage côté MNHN et l'équipe projet Mnémotix. Les Copiltechs ont pu se tenir à Paris au Muséum national d'Histoire naturelle, ce qui a véritablement été un atout pour l'équipe de Mnémotix, qui a pu mieux comprendre, à l'aide de présentation de collections et de rencontres avec les chercheurs, tout l'intérêt du travail autour de cette PoC, au-delà du pur travail sur les données.

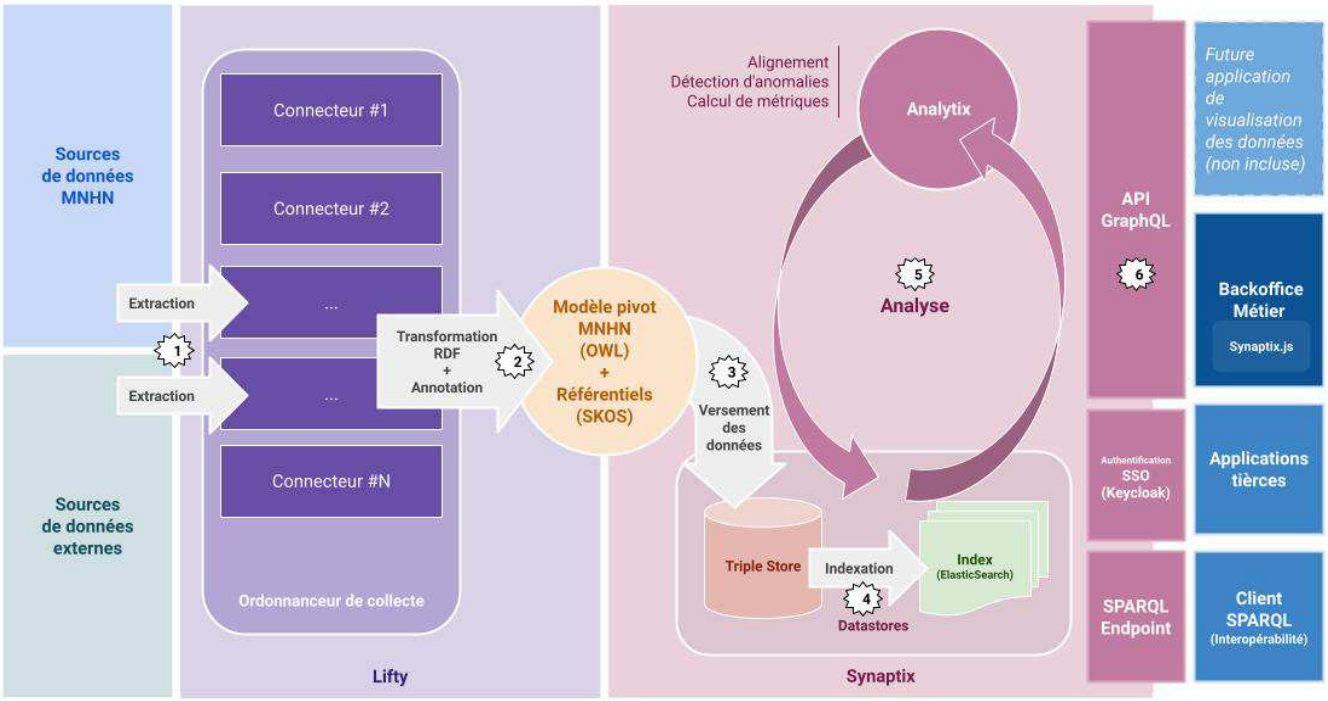
L'équipe a également mis en place une série d'outils de communication et dépôt de la documentation produite au cours du projet : Slack, Gitlab, Google Drive, email.

Objectifs fixés par le CCTP	Objectifs prévus par le calendrier prévisionnel	Dates de l'obtention des résultats	Remarques
Livrable I : Construction du référentiel Personnes -Livraison du modèle de données	7 février 2022	3 mars 2022	
Livrable I : Construction du référentiel Personnes -Livraison du hub de données	25 avril 2022	15 juin 2022	
Livrable I : Construction du référentiel Personnes -Livraison du référentiel personnes sous forme de graphe	30 juillet 2022	30 juillet 2022	
Livrable 2 : API et back office –	19 septembre 2022	18 octobre 2022	La version MOM a été envoyée pour test à l'ensemble de l'équipe projet, ce qui a

Livraison du Back office (MOM)			permis d'effectuer des tests et de rédiger des tickets au sujet des bugs et des améliorations possibles de navigation dans l'application.
Livrable 2 : API et back office – Livraison Back office (VA)	14 novembre 2022	17 novembre 2022	Une deuxième salve de tests et de tickets a été menée, mais uniquement par le binôme chef de projet métier/responsable scientifique.
Livrable 2 : API et back office – Recette	19 décembre 2022	20 janvier 2022	D'un commun accord avec Mnémotix, le développement d'une interface de gestion du chargement des données pour les administrateurs de l'application n'a pas été effectué car le processus de chargement des données de collections naturalistes (à partir du GBIF) et des données bibliographiques (à partir du Sudoc), fonctionnant par téléchargement de fichier, était trop complexe à organiser. Certaines petites améliorations concernant la navigation n'ont pas pu être menées (notamment le fait que les liens de résultats ne pointent pas toujours vers la page Internet sur laquelle est hébergée le résultat mais vers la fiche OAI-PMH du résultat).
Livrable 3 : Rapports et préconisations	19 décembre 2022	6 et 23 janvier 2022	Les rapports et préconisations pour la suite ont été accompagnés d'une formation effectuée par l'équipe Mnémotix à destination de l'équipe de la DINSI responsable du projet de développement du datahub du Muséum. Cette formation a permis de montrer comment les technologies utilisées par Mnémotix pourraient être employées dans le cadre du développement du datahub.

3. RESULTATS DU PROJET : HTTPS://DEV-DATAPOC2-BACKOFFICE.MNEMOTIX.COM/

Le développement d'une nouvelle application dans le cadre du projet Datapoc 2.0 a permis de créer un outil sur des technologies robustes et supportant le passage à l'échelle. Tous les outils développés et intégrés par Mnémotix (Lifty, Synaptix, Analytix) sont libres et régis par la licence Apache 2.0, qui est une licence permissive et non virale, à l'exception notable de la base de données GraphDB-EE, qui est un logiciel propriétaire édité par la société Ontotext.



A ce jour l'application intègre l'ensemble des noms de personnes présents dans les bases de données sources suivantes : le catalogue des imprimés de la bibliothèque (données Sudoc), l'instance Muséum de Calames, HAL-Muséum, le site sciencepress.mnhn.fr, qui recense l'ensemble des publications scientifiques du Muséum et en propose le texte intégral, les données de collections naturalistes disponibles dans le site science.mnhn.fr (à partir des données du Gbif). L'application ne constitue pas encore un référentiel personnes du Muséum à proprement parler, mais elle en pose fortement les bases. Ainsi, chaque nom de personne trouvé dans les bases sources possède sa propre « notice personne ». Pour chaque notice personne, un identifiant pérenne est proposé, ainsi qu'ensemble de fonctionnalités destinées à améliorer facilement la qualité des données dans les bases de données sources ainsi qu'à consolider les données associées aux personnes dans le datapoc. L'algorithme d'analyse et de rapprochement des noms de personnes propose des alignements entre différentes formes du nom trouvées dans les bases sources. Ces propositions apparaissent dans chaque fiche personne. Au sein d'un même groupe de notices personnes, une notice peut être considérée comme « agrégeante », c'est-à-dire que c'est la notice de référence pour la personne, qui contient la forme du nom retenu de la personne, et c'est autour d'elle que toutes les autres notices sont alignées. Cette notice agrégeante est la notice de référence pour une personne. C'est la forme du nom qu'elle propose qui est retenue et qui permet la correction des autres formes du nom dans les bases. C'est l'identifiant pérenne lié à cette notice agrégeante qui doit être retenu pour telle ou telle personne et être éventuellement intégré dans les bases de données sources. Par défaut, toutes les notices générées à partir des noms de personnes issus des catalogues de la bibliothèque et comportant un identifiant IdRef, ont été considérées comme des notices agrégeantes.

L'interface est moins graphique et moins orientée « tout public » que pour la première version issue du projet datapoc.mnhn.fr. C'est en partie dû à l'usage auquel on destine ce nouvel outil, moins orienté grand public et davantage destiné à une communauté de recherche, mais cela tient également au fait que la partie design de l'interface a été retirée du périmètre du projet pour des raisons budgétaires, comme c'est expliqué plus haut.

L'utilisateur peut rechercher un nom (nom de famille, nom de famille + prénom…) à partir de la barre de recherche. Une autocomplétion permet de lui proposer plusieurs résultats dès la recherche, mais il peut également appuyer sur la touche « Entrée » pour obtenir la liste des résultats liés au nom recherché. Chaque résultat est cliquable et permet d'accéder à la notice personne liée à la forme du nom. La notice personne d'abord une sorte de carte d'identité de la personne : nom, prénom, forme du nom dans la base, identifiant MNHN (généré par l'application datapoc), identifiants associés (IdRef, ISNI, data Bnf, Orcid, Wikidata), notes biographiques. Ensuite vient le statut de la notice (agrégeante ou non) et les alignements proposés par l'algorithme avec cette forme du nom, et qui prennent d'autres formes. Ensuite, sur la page, viennent différents onglets à déplier et qui proposent de visionner : les collaborateurs de la personne, les références bibliographiques dont elle est auteur, celles dont elle est illustrateur, contributeur, éditeur, ainsi que les spécimens de collections naturalistes dont la personne est le collecteur ou le déterminateur. Sous chacun des onglets les résultats se présentent sous la forme d'un tableau exportable, sauf celui des alignements proposés. Chaque colonne des tableaux de résultats possède une fonction de tri et pour chaque tableau on peut également filtrer les résultats par source de données.

Le nouveau datapoc est également un outil de construction et de consolidation d'un référentiel personnes pour le Muséum national d'Histoire naturelle, il propose aux utilisateurs ayant le statut d'opérateur ou d'administrateur des fonctionnalités de validation ou d'invalidation des alignements proposés, ainsi que des fonctionnalités de choix de conférer à la notice la plus complète sur une personne, avec une forme du nom correcte et complète, le statut de notice agrégeante. Le statut de notice agrégeante permet ensuite de déterminer la « forme retenue » du nom, et la notice devenue agrégeante devient ensuite la forme sur laquelle s'alignent les différentes autres formes du même nom de personne trouvées dans les bases de données sources. L'identifiant pérenne de cette notice agrégeante doit être, à terme, intégré dans les bases de données sources. Ainsi les données seront consolidées dans les bases de données sources et viendront ensuite consolider encore davantage le référentiel datapoc au moment du prochain rechargement de données.

Outre les fonctionnalités de validation ou d'invalidation des alignements et de création de notices agrégeantes, les opérateurs et administrateurs peuvent également modifier une notice, l'enrichir avec des informations complémentaires, des identifiants complémentaires, proposer des alignements et déterminer des liens de collaboration ou de parenté entre différentes personnes. Il est également possible de créer ou de supprimer une fiche personne.

4. CONCLUSION

Le retard pris au lancement du projet n'a pas permis de mettre en place comme prévu des tests utilisateurs pendant la phase de déroulement du projet. Seuls les membres de l'équipe projet ont eu la possibilité de tester l'application. Les retours de leur part sont positifs mais il paraît nécessaire d'imaginer d'autres séries de tests avec d'autres chercheurs et chargés de collection au MNHN, afin de confirmer ces premiers enthousiasmes ainsi que de détecter de nouveaux bugs et de rédiger des propositions pour améliorer les fonctionnalités liées au repérage des erreurs dans les bases de données sources.

La preuve majeure de la réussite de ce projet du point de vue de l'établissement réside dans la décision de la DINSI d'installer la PoC réalisée par Mnémotix au Muséum afin d'en assurer la continuité, mais également d'utiliser les technologies et l'architecture déployées par Mnémotix dans la construction du hub des données du Muséum. Cette installation au sein de l'établissement constitue un véritablement engagement du point de vue de la DINSI car elle implique notamment l'achat d'une licence GraphDB-EE à renouveler chaque année, ainsi que la formation des agents de la DINSI à ces technologies.

Cette pérennisation de l'application au sein de l'établissement permet d'envisager que l'application puisse évoluer vers la constitution d'un véritable référentiel pour le Muséum, ouvert et lié à d'autres référentiels nationaux et internationaux. Pour atteindre cet objectif, plusieurs pistes de travail sont envisagées : autour de l'amélioration des fonctionnalités de l'outil afin de permettre la mise en qualité des données dans les bases de données sources ; dans l'intégration d'un champ permettant la saisie d'un identifiant pérenne lié dans les bases de données de collections naturalistes ; dans l'actualisation des pratiques de saisie des objets de collection afin d'intégrer la donnée de référentiel dans les bases. Les projets actuels de la DINSI, notamment ceux concernant la gouvernance des données de l'établissement et la refonte du SI Collections-Recherche, sont de bon augure pour continuer à avancer sur ce projet au-delà du financement obtenu dans le cadre de Collex-Persée. Concernant la partie amélioration de la qualité des données, un partenariat avec l'Abes pourrait également être envisagé, en lien avec l'amélioration des données dans IdRef.