

Modèle d'auto-évaluation pour les institutions détentrices de données

Introduction

Cet outil d'auto-évaluation permet aux institutions détentrices de données de réaliser un examen de haut niveau de leur propension et progression à fournir une information digitale sur les ressources qu'elles possèdent. Il est organisé suivant un modèle à cinq composants couvrant la planification et la mise en oeuvre d'une stratégie de mobilisation de contenu durable. Cet outil est destiné à aider les gestionnaires de collections pour mieux comprendre les étapes de partage des données mais aussi afin d'identifier les points où focaliser leur attention pour permettre un suivi à long terme de ces tâches.

Dans le même temps, GBIF utilisera le modèle présenté pour planifier et organiser la formation et le matériel de support adéquat pour les différents composants. Le but est de garantir que les collections biologiques de toutes sortes soient bien positionnées pour s'inscrire dans une information globale interconnectée en support à une taxonomie du 21^{ème} siècle et une utilisation durable de biodiversité. Même des étapes fort simples, comme le partage des métadonnées de base d'une collection peuvent déjà contribuer à ces objectifs.

Cette auto-évaluation vise à établir un statut et une capacité de l'institution dans cinq domaines clés:

1. **La Stratégie:** Planifier comment exécuter une mobilisation de données. Les considérations clés incluent l'objectif de la mobilisation, la priorisation du matériel à numériser, les tâches/ l'expertise, l'équipement, le standard de données, le choix d'un système de base de données, les droits de propriété intellectuelle et les problèmes de licences.
2. **La Numérisation:** La conversion et la capture des données associées aux spécimens physiques et objets vers un format électronique et une base de données. Ce processus peut inclure de l'imagerie.
3. **La Publication:** La mise en ligne, en accès publique des données numérisées (incluant les images et multimédias associés) soit via des sites web institutionnels ou communautaires soit via des grands dépôts internet.
4. **La Conservation et la Maintenance:** L'exécution répétitive de tâches incluant le nettoyage, la correction et la mise à jour partiellement basée sur le retour des utilisateurs.
5. **La Préservation et l'Archivage:** La conservation à long terme du contenu numérisé et des bases de données afin d'assurer leur intégrité et persistance.

Bien que ces domaines sont très clairement liés, les détenteurs de données ne vont pas toujours s'y atteler dans l'ordre chronologique présenté. Dans certains cas, l'institution sera plus avancée dans l'un de ces domaines et moins dans d'autres. Un avancement dans un domaine peut également en renforcer d'autres.

Implémentation de l'auto-évaluation

Cette auto-évaluation est destinée aux institutions détentrices de données désireuses de développer leur capacité à mobiliser, gérer et disséminer leurs données. Une évaluation devrait mesurer le statut dans les cinq domaines principaux de capacité en utilisant une échelle à trois points pour chaque aspect mesuré. Les scores globaux aideront à mesurer la capacité institutionnelle au moment de l'évaluation mais aussi de donner des suggestions sur les possibilités d'amélioration. La section de notes est destinée à accueillir commentaires, liens, informations de contact, références, rappels, etc afin d'aider l'évaluateur à enregistrer les détails d'une section et d'en faire le suivi. Prendre des notes peut aussi aider l'institution à comprendre le contexte dans lequel l'évaluation a été réalisée, plus spécialement si les évaluations futures ne sont pas effectuées par la même personne.



Notes explicatives

Les institutions détentrices de données en biodiversité sont des entités ou organisations qui hébergent, ou sont en charge, de données de biodiversité sous différentes formes incluant des données numérisées ou non. Quelques exemples parmi d'autres : les muséums d'histoire naturelle, les herbiers, les collections culturelles, les jardins botaniques, les arborétums, les zoos, les organisations de science citoyenne et les bibliothèques.

La mobilisation de données en biodiversité est le processus de capture de données et d'information sur les organismes vivants, principalement les spécimens et les autres objets liés à la biodiversité, au format numérique, et la publication des données résultantes afin de les rendre disponibles à la découverte, l'utilisation et la réutilisation. La plupart des données détenues par les institutions détentrices de données en biodiversité dans le monde restent électroniquement inaccessible. L'information est parfois difficile à utiliser, et dans d'autres cas elle n'est pas utilisable du tout. Les institutions détentrices de données avec des données non mobilisées ont des difficultés à évaluer l'ensemble de leurs avoirs.

La numérisation regroupe la conversion et la capture des données associées aux spécimens physiques et objets concernés (incluant les enregistrements papier) au format électronique et en base de données. Cela implique généralement la retranscription des étiquettes et l'imagerie d'objets ou de données existant sous la forme analogique comme du texte en format papier. La numérisation offre de nombreux bénéfices, comme la préservation digitale, l'amélioration de la gestion des collections et la dissémination via la publication des données.

La préservation numérique comprend les outils et les processus nécessaires afin d'assurer la persistance à long terme des collections numérisées, leur intégrité et leur durabilité. Ceci comprend la création d'une sauvegarde et/ou une réplique numérique de haute qualité de l'objet original. Partant du constat que la préservation est au coeur du mandat des institutions détentrices de données, la préservation numérique est une considération importante dans la planification des projets de numérisation. Un projet qui inclut de la préservation numérique sera d'autant plus exigeant dès lors que les attentes en termes de spécifications techniques sont plus élevées. Les formats des fichiers générés par la préservation digitale sont généralement des archives très lourdes qui ne sont pas destinées à la publication sur le Web ; ils doivent à cette fin être convertis dans d'autres formats avant d'être partagés via Internet.

Le système de gestion de base de données est un outil logiciel pour stocker et gérer des données de biodiversité et de l'information.



Figure 1 : Modèle d'auto-évaluation pour les institutions détentrices de données

Questionnaire d'auto-évaluation pour les détenteurs de données de biodiversité

Institution / Collection:

Section 1: La stratégie

	Non (0)	En cours/ partiellement (1)	Oui (2)
1. Votre collection a-t-elle une mission définie (documentée) déterminant son rôle et son objectif?			
2. L'objectif de l'activité de mobilisation a-t-il été défini ?			
3. La portée et l'ampleur des efforts de numérisation – leur étendue, les éléments de données, la préservation numérique, etc- ont-ils été définis?			
4. Les curateurs ont-ils identifiés et évalués les activités de pré-numérisation nécessaires ou souhaitables – le tri, la sélection d'échantillons, la révision taxonomique, la vérification des carnets de terrains, etc ?			
5. Est-ce que l'institution a une politique de données qui couvre des sujets tels que les droits de propriété intellectuelle, les responsabilités de garde, l'accès, les licences, la responsabilité, la vie privée et la sensibilité des données?			
6. Le système/logiciel de gestion de base de données a-t-il été choisi?			

Section 1: Score global

Remarques additionnelles sur la stratégie

Section 2: La numérisation

	Non (0)	En cours/ partiellement (1)	Oui (2)
1. L'espace de travail adapté à la numérisation a-t-il été identifié et installé ?			
2. La préparation de la numérisation a-t-elle été effectuée (le triage, la sélection de ce qui va être numérisé, la révision taxonomique et les étiquettes, condition et santé des spécimens, attribution d'identificateurs uniques comme des codes barre et suivi généralisé) ?			
3. Les processus et technologies de numérisation ont-elles été définies (encodage de texte, imagerie, multimédia, autre) ?			
4. Les éléments de données ont-ils été confrontés aux standards de données internationaux ?			
5. Les flux de travail, incluant la manutention d'objets physiques, ont-ils été bien définis ?			
6. Le personnel a-t-il été formé et équipé de manière adéquate et appropriée ?			
7. Les processus et standards de contrôle de qualité – comme la taxonomie, le géo-référencement, les erreurs typographiques, la confusion de champs, les spécifications techniques et les méta-données - ont-ils été clairement spécifiés ?			

Section 2: Score global

Remarques additionnelles sur la numérisation

Section 3: La Publication

	Non (0)	En cours/ partiellement (1)	Oui (2)
1. Les besoins des potentiels utilisateurs finaux et de la publication sur le Web ont-ils été complètement considérés lors du choix des formats de données ?			
2. Les licences choisies sont-elles lisibles par les machines et supportées par le GBIF ?			
3. L'outil de publication des données (sur le Web) a-t-il été choisi et installé ?			
4. L'institution a-t-elle accès à une infrastructure d'hébergement de données stable ?			

Section 3: Score global

Remarques additionnelles sur la publication

Section 4: La Conservation et la Maintenance

	Non (0)	En cours/ partiellement (1)	Oui (2)
1. Les processus de mise à jour suite au contrôle de qualité ou à de nouvelles informations (par exemple une révision taxonomique) ont-ils été définis ?			
2. L'institution a-t-elle les ressources et les processus afin de gérer les corrections et les suggestions des utilisateurs en ligne ?			
3. Les processus adéquats permettant d'actualiser les données numérisées par rapport aux informations de conservation du matériel original, et vice-versa, sont-ils en place ?			

Section 4: Score global

Remarques additionnelles sur la conservation et la maintenance

Section 5: La Préservation et l'Archivage

	Non (0)	En cours/ partiellement (1)	Oui (2)
1. L'institution a-t-elle accès à un dépôt d'archivage de données à long terme ?			
2. Cette solution garantit-elle adéquatement la sécurité des données et la sauvegarde contre l'obsolescence des formats de données et d'applications ?			

Section 4: Score global

Remarques additionnelles sur la préservation et l'archivage

Documentation de référence

Initiating a Collection Digitisation Project (<https://www.gbif.org/document/80574/initiating-a-collection-digitisation-project>)

Ce guide explique comment planifier et exécuter une tâche de numérisation. Bien que centré sur les collections d'histoire naturelle, il peut s'appliquer à d'autres situations.

Consortium of Northeastern Herbaria (http://neherbaria.org/digit_resource)

Une liste annotée de ressources (et liens) utiles pour la numérisation de collections biologiques. Ce document est destiné à ceux qui entament un effort de numérisation de collections biologiques. Les ressources sont groupées en plusieurs thèmes : vue d'ensemble, base de données (numérisation de texte), géo-référencement, imagerie, mobilisation, standards et outils.

iDigBio Digitization Resources (https://www.idigbio.org/wiki/index.php/Digitization_Resources)

Cette page web offre des ressources et des informations sur une série d'ateliers de numérisation conduits par iDigBio ainsi qu'une pléthore d'informations et de ressources. Sont inclus de nombreux liens vers des documents, des sites internet, des vidéos, des présentations et d'autres informations importantes relatives à la numérisation de collection biologique. iDigBio propose aussi fréquemment des webinars gratuits et ouverts à tous.

The Atlas of Living Australia (Atlas) Digitisation Guidance (<http://www.ala.org.au/about-the-atlas/digitisation-guidance/>)

Guidance sur les différents aspects de la numérisation, développé par les partenaires de l'Atlas. Cette guidance explore différents thèmes : Qu'est-ce que la numérisation ? L'imagerie et la numérisation faisant appel à des volontaires. Cette page web fournit des liens vers des documents détaillés sur les trois thèmes cités. Ce matériel de guidance peut être réutiliser tel quel ou adapté pour d'autres projets que l'Atlas.

Biodiversity Data Standards (<http://www.tdwg.org/>)

Des recommandations de base et des accords documentés sur les représentations, les formats et les définitions pour les données primaires de biodiversité. Biodiversity Information Standards, l'organisme responsable des standards est communément connu sous son acronyme original TDWG.

BioSharing.org (<https://biosharing.org>)

Ce portail de recherche sur les standards interdépendants de données, les bases de données, les politiques liées aux sciences de la vie, environnementales et bio-médicales.

Towards demand driven publishing: Approaches to the prioritisation of digitization of natural history collections data (<https://journals.ku.edu/index.php/jbi/article/view/3990>)

Cet article rassemble des informations de haut niveau (métadonnées) sur les collections d'histoire naturelle susceptible d'être numérisées et leur offre une visibilité (au niveau des objets/spécimens concernés) avant même d'être complètement numérisées. Il défend une approche à la demande de la numérisation et offre des outils de mesure qui permettent de prioriser la numérisation des collections lorsque des ressources ne permettent pas de couvrir l'ensemble de la collection.

Data publishing guidance from GBIF and benefits (<https://www.gbif.org/publishing-data>)

GBIF.org propose des lignes directrices et des processus sur la publication de données via le réseau GBIF incluant les avantages de la publication des données, l'approbation de l'éditeur de données par le point nodal et d'autres liens vers des ressources additionnelles. Pour les avantages et bénéfices de la publication en libre accès des données de biodiversité, veuillez suivre les liens suivants : <https://www.gbif.org/article/11g6LFdAOylaUOM0EuCqso/quick-guide-to-publishing-data-through-gbiforg#incentives>; <http://bioscience.oxfordjournals.org/content/59/5/418.full>

Creative Commons licensing (<https://www.gbif.org/terms>)

Cette ressource décrit le processus, les résultats et la raison d'être de l'adoption par GBIF de l'une des trois licences lisibles par des machines pour les jeux de données d'occurrences d'espèces publiées sur GBIF.org à savoir : CC0, CC-BY et CC-BY-NC.