Identifiant permanent de la DEE

Définition opérationnelle dans le cadre du SINP pour la thématique occurrence de taxon

Julie Chataigner (MNHN/SPN), Laurent Poncet (MNHN/SPN), Marie-Elise Lecoq (GBIF France), Simon Chagnoux (MNHN/DSI), Frédéric Vest(MNHN/SPN)
Relecture du GT Standard de données *Avril 2014*

1 Introduction

L'objectif de ce document est de présenter l'identifiant permanent de la donnée élémentaire d'échange (DEE) du SINP. Ce travail a fait l'objet d'un sous-groupe de travail rassemblant des structures ayant expérimentées la mise en place d'identifiants permanents : le SPN du Mnhn, le service des collections du Mnhn et le GBIF France.

Le protocole du SINP indique que l'identifiant permanent du SINP porte sur la Donnée Elémentaire d'Echange (DEE). En effet une observation peut être représentée par plusieurs « objets » cf Figure 1. Chacun de ces objets est donc susceptible d'être identifié, mais seule la DEE doit l'être au niveau national. Pour rappel, les DEE sont des données standardisées interopérables. Elles sont élaborées à partir des données-source (DS) selon un format standard national propre à chaque thématique du SINP (observations de biodiversité, paysages, espaces protégés, etc)

La photo de l'observation	Le specimen	…	La donnée source (DS)	La donnée élémentaire d'échange (DEE)
		…		

Figure 1. *Exemples de différents objets pouvant être identifiés à partir d'une seule observation*

2 Choix de l'identifiant permanent

2.1 Proposition du GBIF

Cette partie du document reprend la conclusion de la solution technique proposée par le GBIF et décrite dans le document analyse_identifiant_gbif_20131025.pdf[1]
L'analyse du GBIF France conclue sur l'utilisation d'une solution d'identifiant permanent clé en main pour les utilisateurs non avertis sans empêcher des plateformes ayant des ressources nécessaires de développer leurs propres algorithmes.

L'analyse propose de se baser sur :
- **une architecture permettant potentiellement d'accéder à la donnée** en URL. Le type d'URL (PURL, …) relève de l'architecture et n'est pas développé dans ce document.
- **un UUID** : Suite alphanumérique pseudo-aléatoire générée par des algorithmes assurant à très forte probabilité le caractère unique de l'identifiant dans le monde. L'UUID est normalisé par l'ISO/IEC 9834-8:2008. Son format est le suivant : XXXXXXXX-YYYY-ZZZZ-AAAA-BBBBBBBBBBBB. Exemple : a0eebc99-9c0b-4ef8-bb6d-6bb9bd380a11.

Ainsi la structure de l'identifiant proposée est la suivante : http://www.autorité/contexte/resourcePath/Objet

Avec
- **Autorité** : nom de la plateforme nécessitant un nom de domaine pérenne. Si la plateforme est capable de gérer son infrastructure, la création, la gestion et la maintenance des noms de domaine peuvent se faire par la plateforme, sinon elle peut être faite au niveau national. Au niveau national, seul l'annuaire des noms de domaine des plateformes régionales et thématiques devra être impérativement mis à jour et disponible.
- **Contexte** : le contexte représente les jeux de données ou les fournisseurs. Ils sont identifiés via un code. Deux solutions sont proposées : les Digital Object Identifier[2] (DOI) ou un code. S'il est souhaité de garder une homogénéité au niveau du territoire, le code devra être défini au niveau national sinon le choix de la codification peut être laissé libre aux plateformes tout en préconisant une organisation.
- **Resource Path** : (facultatif) définition du périmètre de l'URL : spécification d'une unique localisation pour les informations d'une même ressource. Même logique que de créer un répertoire pour y localiser les informations d'un même sujet. Ici, pour les données de jeux de données cela serait « data ».
- **Objet** : L'objet représente l'occurrence. C'est à ce niveau qu'est mis en place l'UUID. Aujourd'hui, la plupart des SGBD ont pris en compte un champ spécifique UUID. Attention, l'UUID ne remplacera pas la clé primaire. L'objet ne doit pas être lié au type de technologie utilisé.

D'où la proposition :
http://nom_plateforme.org/data/nom_jeu_donnee/UUID.

Précisions sur la proposition :
Une URL peut être utilisée comme identifiant, sans UUID pour identifier l'objet. Cependant dans ce cas, si l'autorité ou le contexte change alors on est susceptible de perdre l'unicité et la permanence de l'identifiant de la donnée. De plus, cela implique potentiellement de nombreuses variantes dans la création de l'identifiant de l'objet, rendant leur unicité incertaine au niveau national.

[1] Lecoq M-E., (2013) *Analyse et proposition d'identifiant permanent pour le SINP*, GBIF France, 23 pp
[2] Le système DOI fournit des outils pour gérer et identifier de façon pérenne les objets numériques. Les DOI sont standardisés et payants.

L'UUID utilisé sans URL est complètement opaque et permet de limiter les essais d'interprétation de la donnée à partir de son identifiant. C'est un identifiant permanent robuste mais il ne permet pas d'accéder directement à la donnée.
Ainsi, utiliser un identifiant combinant ces 2 types d'identifiant permet de cumuler leurs avantages sans créer de plus grands inconvénients :
- Accès à la donnée rendu possible par l'URL
- Pérennité accrue de la donnée via la pérennité du nom de domaine
- Unicité assurée par la combinaison de l'URL et l'UUID
- Compromis entre opacité de l'identifiant et information sur l'objet identifié. En effet, l'URL donne un minimum d'information sur la donnée échangée (nom de la plateforme, prise de connaissance que la donnée échangée concerne une donnée d'une plateforme régionale ou thématique, une URL pouvant identifier divers objets comme un document, un film, une page web …)

Pour plus d'informations techniques, se référer au document du GBIF.

2.2 Proposition finale au SINP

2.2.1 Cas général

Nous proposons de retenir la proposition du GBIF avec les quelques modifications et remarques suivantes :

- L'**autorité** : nom de domaine de la plateforme.
- Le **contexte** : comme l'identifiant est rendu unique par l'UUID, et que le contexte est seulement un élément du lien pour accéder aux données, **il est proposé de ne pas l'utiliser**. Ainsi, l'identifiant permanent gagne en concision et en robustesse.
- Le **Resource Path** : Cet élément permet de définir des périmètres suivant les thématiques du SINP. Pour la thématique du standard « occurrence de taxon », il est proposé de nommer le resourcePath : « occtax »
- L**'objet** correspondant à l'identifiant de l'occurrence est l'UUID.

D'où la proposition :
http://nomdomainedelaplateforme/thematique/UUID

Remarque sur la résolution de l'identifiant :
L'identifiant est une URL permettant l'accès à la donnée. Ainsi, le résultat qu'il donne peut être :
- une erreur (de type erreur 404)
- une ressource (page HTML ou RDF) décrivant les données de l'observation.
- un renvoi http vers une autre ressource utile à la personne ou à la machine s'intéressant à l'occurrence désignée par l'URI (approche du linked data)

Bien que potentiellement accessible via l'identifiant unique, la donnée sera aussi diffusée avec l'ensemble des attributs qui la compose par les plateformes du SINP.

2.2.2 Exemples

Cas de données transitant par la plateforme thématique « occurrence de taxon » : exemple des données de collection du Mnhn.
Les données de collection du Mnhn sont des données d'emprise nationale. Les DEEs sont donc créées par la plateforme thématique « occurrence de taxon ».

Ainsi, dans le SINP, la DEE issue de cette DS de collection aura pour :

IdentifiantPermanent : http://nomplateformethematiqueoccurrencetaxon/occtax/a0eebc99-9c0b-4ef8-bb6d-6bb9bd380a11
IdentifiantOrigine : http://coldb.mnhn.fr/catalognumber/mnhn/ec/ec3060
L'identifiantOrigine étant l'identifiant des occurrences de specimen propres aux bases de collection du Mnhn. Comme l'identifiant est celui de l'objet physique (le specimen), la résolution de l'identifiant est une redirection vers une page web sur le specimen.

Cas de données transitant par une plateforme régionale
Dans le cas d'une DEE créée par une plateforme régionale :
identifiantPermanent : http://nomplateformeregionale/occtax/f0eevc75-9c0b-4ef8-bz7z-8zb9bz380g15
IdentifiantOrigine : identifiant du producteur

Remarque : le nom des autorités (domaine des plateformes) de ces exemples sont théoriques. Ils sont à définir par qui de droit.

3 Outils et « briques » nécessaires

3.1 Annuaire des autorités : nom de domaine décliné par plateforme

L'annuaire des autorités doit être partagé et géré au niveau national. Dans le cas du SINP, il liste le nom de chaque plateforme. La définition des règles de nommage et de gestion des domaines sont du périmètre de l'architecture du SINP et/ou des responsables des plateformes.
L'identifiant étant basé sur une URL dont l'intérêt est de pouvoir être résolu, il faut que les noms de domaine soient pérennes et que les responsables des URL aient les moyens techniques de rediriger les URLs.
Le format et les informations nécessaires dans l'annuaire dépendent des composants logiciels des plateformes et l'usage qu'ils en feront. Une ressource XML ou JSON statique donnant une liste de {nom de domaine, regexp de l'identifiant, email de contact} pourrait suffire.

3.2 Resource Path

Il est nécessaire de définir une liste de valeur disponible au niveau national. Cette liste peut être proposée par le GT Standard de données, au fur et à mesure de la définition des thématiques dans le SINP.

3.3 Objet de l'identifiant : algorithme

L'objet peut être directement mis en place et géré par les plateformes régionales ou thématiques. Des algorithmes pour générer les UUIDs sont disponibles dans des bibliothèques java, python, C++ etc, et dans les modules de base de données (PostgreSQL, Oracle etc).

3.4 Ressources humaines et matérielles

Les compétences suivantes sont requises pour créer et gérer l'identifiant permanent au niveau des plateformes régionales et thématiques et de la plateforme nationale :
- animateur national
- administrateur et gestionnaire « données » au niveau plateforme
- administrateur technique de la plateforme
- hébergeur de la plateforme

4 Processus

La mise en place d'un identifiant permanent de la DEE homogène et partagé par l'ensemble des plateformes du SINP proposé dans ce document n'empêche pas l'existence d'un identifiant régional sur la donnée dans le format régional.
La possibilité de véhiculer l'identifiant régional au niveau national en plus de l'identifiant permanent national et de l'identifiant origine de la donnée peut être discutée et pourra amener à l'ajout de cet attribut dans le standard DEE « occurrence de taxon ».

4.1 Processus d'attribution de l'identifiant à la DEE

Les attributions d'identifiant permanent national à une DEE sont faites par les plateformes régionales et thématiques. Cet identifiant est ensuite retourné au producteur et/ou fournisseur de la donnée source. Sa prise en compte dans leurs systèmes n'est pas obligatoire mais fortement conseillée.

4.2 Processus d'évolution de la DEE

Afin de rester robuste et simple dans la mise en œuvre, l'identifiant permanent ne gère pas le suivi des modifications ou de suppressions logiques d'une occurrence. Il faudra que cela soit porté dans les bases de données du SINP et dans le standard d'échange (exemple : date de dernière modification).

De plus, l'identifiant concerne l'ensemble des champs (attributs) de l'occurrence sous sa forme DEE. Par exemple, si des attributs facultatifs du standard sont ajoutés à l'occurrence, ce n'est pas considéré comme une nouvelle occurrence : l'identifiant de la DEE reste le même.