Système d'identification des données et métadonnées du SINP
-
Analyse de l'existant et nouvelles propositions

Note technique

Historique des versions du document

Version	Date	Commentaire
0.1	02/10/17	Version initiale
0.2	19/10/17	Version stabilisée suite à la réunion du GT Architecture du 10 octobre
1	27/11/17	Validée

Rédacteurs

Thomas MILON (UMS 2006 PatriNat)

Relecteurs

Cercle 1 du GT Architecture (présenté en séance lors de la réunion du 10 octobre)

Cercle 2 du GT Architecture (présenté par mail du 31 octobre au 27 novembre)

Table des matières

1 Contexte

Malgré la publication d'une note en avril 2014 sur la définition opérationnelle de l'identifiant permanent de la donnée élémentaire d'échange (DEE), la production, la gestion et le partage de cet identifiant semblent encore limités.

Pour assurer au SINP une réelle traçabilité[1] des données, limiter la problématique de doublon dans le système et répondre à l'entrée prochaine des données brutes de biodiversité, cette définition opérationnelle de l'identifiant permanent doit évoluer.

L'objectif de cette note est :

- d'analyser l'existant pour faire ressortir les points d'améliorations ;
- de proposer une nouvelle architecture cible à la vue de ces points d'améliorations ;

2 Analyse de l'existant

2.1 Éléments théoriques

Le document « Identifiant permanent de la DEE - Définition opérationnelle dans le cadre du SINP pour la thématique occurrence de taxon »[2] a pour objectif de présenter l'identifiant permanent de la donnée élémentaire d'échange (DEE) du SINP.

Attention : les chapitres 2.1.*x* reprennent de manière synthétique les **éléments théoriques** proposés dans le document « Identifiant permanent de la DEE - Définition opérationnelle dans le cadre du SINP pour la thématique occurrence de taxon ».

2.1.1 Objet de l'identifiant selon la définition opérationnelle

Le protocole du SINP indique que l'identifiant permanent du SINP porte sur la Donnée Elémentaire d'Echange (DEE) : « *les informations obligatoires pour des utilisations nationales des DEE (listées dans le protocole à l'article 10.3.6) sont notamment : un identifiant unique national de la donnée élémentaire d'échange, afin d'éviter les doublons et de gérer la traçabilité des données en particulier lors de leurs validations successives* ; ». L'identifiant permanent du SINP ne porte pas sur la donnée source ni sur tout autre objet (photo de l'observation, spécimen…).

2.1.2 Format de l'identifiant selon la définition opérationnelle

Le format de l'identifiant permanent suit le modèle suivant :

[1]Ensemble des informations nécessaires et suffisantes pour connaître soit les différents états successifs pris par une donnée dans les différents processus auxquelles elle est soumise (notion d'historique), soit la nature des processus ou des opérations qui l'ont affectée (création, modifications, suppression, par qui, quand, etc.).
Référence: *Ateliers du GT définition de l'architecture / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

[2]https://inpn.mnhn.fr/docs/standard/sinp_identifiantpermanent.pdf

- une architecture en URL permettant d'accéder potentiellement à la donnée. L'URL prend cette forme : http://*nomdomainedelaplateforme/thematique/UUID* avec
  - « nomdedomainedelaplateforme » un nom de domaine pérenne et défini par les plateformes ;
  - « thématique » la thématique de la donnée (ex : occtax pour les occurrences de taxon) ;
  - « UUID » une suite alphanumérique pseudo-aléatoire générée par des algorithmes assurant à très forte probabilité le caractère unique de l'identifiant dans le monde (ISO/IEC 9834-8:2008). Son format est le suivant : xxxxxxxx-yyyy-zzzz-aaaa-bbbbbbbbbbbb. Exemple : a0eebc99-9c0b-4ef8-bb6d-6bb9bd380a11.

La combinaison de ces éléments permet :
- un accès à la donnée rendu possible par l'URL ;
- une pérennité accrue de la donnée via la pérennité du nom de domaine ;
- une unicité assurée par la combinaison de l'URL et l'UUID ;
- un compromis entre opacité de l'identifiant et information sur l'objet identifié.

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).

2.1.3 Les ressources nécessaires selon la définition opérationnelle

Les ressources nécessaires pour gérer ces identifiants permanents sont :
- un annuaire des autorités listant les noms de domaines
- la liste des valeurs pour les thématiques (ici : occtax)
- un générateur d'UUID
- des animateurs et administrateurs de données pour les plateformes avec un hébergement stable.

2.1.4 Préconisations selon la définition opérationnelle

« 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 afin d'améliorer la traçabilité des données ».

« Afin de rester robuste et simple dans la mise en oeuvre, 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). »

2.2 État des lieux au sein du SINP

Afin de constituer un état des lieux rapide de la production, la gestion et le partage des identifiants permanents au sein du SINP, une enquête a été envoyée aux membres du GT Architecture (Cf. 6.1 Questionnaire). Ce questionnaire a été ouvert du 2 octobre jusqu'au 9 octobre au cercle 1 et 2 du GT.

2.2.1 Bilan des réponses

15 réponses ont été récupérées. Ces réponses ont été discriminées à postériori selon le niveau de la plateforme dans le référentiel SINP :

- Niveau national : 2 réponses = Plateforme thématique « occurrences de taxon », Ecoscope ;
- Niveau régional : 10 réponses de plateforme régionale et pôle régionaux thématiques : Sigogne, SINP Occitanie, ODIN, OAFS, PIFH, Karunati, Cettia IDF, SILENE Faune, SILENE Flore, Borbonica ;
- Niveau Producteur : 3 réponses = GéoNature, SiCEN, BDN (ONF) ;

2.2.1.1 La plateforme produit et gère-t-elle des identifiants de jeu de données ?

- Production d'identifiant unique propre au système

Niveau national	GINCO	Niveau Régional	Producteur
PTF Th. OccTax = **oui** Ecoscope = **oui**	**Oui** : UUID de l'application nationale de métadonnées SINP	**6** ptf => **oui** **4** ptf => non	**1** prod. => **oui** **2** prod. => non

- Production d'identifiant unique vis-à-vis du SINP

Niveau national	GINCO	Niveau Régional	Producteur
PTF Th. OccTax = **oui** Ecoscope = non	**Oui** : UUID de l'application nationale de métadonnées SINP	**4** ptf => **oui** **4** ptf => non 2 ptf => récupération ptf nationale (**oui**)	**3** prod. => non (dont Géonature = développement en cours)

2.2.1.2 La plateforme produit et gère-t-elle des identifiants d'observation de taxon ?

- Production d'identifiant unique propre au système

Niveau National	GINCO	Niveau Régional	Producteur
ptf Th. OccTax = **oui** (*id propre*) Ecoscope = **non**	**Oui** (UUID)	**10** ptf => **oui**	**2** prod. => **oui** **1** prod. => ID unique d'observation (plusieurs taxons)

- Production d'identifiant unique vis-à-vis du SINP

Niveau National	GINCO	Niveau Régional	Producteur
ptf Th. OccTax = **oui** (*id permanent*) Ecoscope = **non**	**Oui** (UUID)	**7** ptf => **oui** **3** ptf => **non**	**2** prod. => **oui** **1** prod. => ID unique d'observation(plusieurs taxons)

2.2.1.3 Intégrez-vous dans votre plateforme des identifiants permanents venant de partenaires ?

Niveau National	GINCO	Niveau Régional	Producteur
ptf Th. OccTax = **oui** Ecoscope = **oui** (métadonnées)	**oui**	**6** ptf => **oui** 4 ptf => non	**2** prod. => **oui** 1 prod => selon structutre des jeux partenaires

2.2.1.4 Utilisez-vous des identifiants permanents pour...?

	Niveau National	GINCO	Niveau Régional	Producteur
Analyser les doublons	Ecoscope : **oui**	**Oui** : rejet des id permanent existant	**5** ptf	**2** prod.
Gérer l'unicité des jdd	ptf Th. OccTax : **oui** Ecoscope : **oui**	**Oui** : Se base sur la ptf nationale	**5** ptf	**2** prod.
Gérer l'unicité des observations	ptf Th. OccTax : **oui**	**Oui**	**8** ptf	**2** prod.
Pas d'utilisation			2 ptf	

2.2.1.5 Partagez-vous des identifiants permanents (produits ou reçus) avec des partenaires?

	Niveau National	GINCO	Niveau Régional	Producteur
Avec la ptf nationale	*Sans objet*	**oui**	**6** ptf	**2** prod.
Avec des ptf régionales		*Sans objet*	*Sans objet*	**1** prod.
Avec des producteurs de données	ptf Th. OccTax	**oui**	**5** ptf	*Sans objet*
Avec des réutilisateurs	Ecoscope	**oui**	**6** ptf	**3** prod.
Non			**1** ptf	

2.3 Conclusion

Malgré un a priori négatif sur les limites concernant la production, la gestion et le partage des identifiants permanents, les résultats du questionnaire montre que plusieurs acteurs du SINP ont intégré ces éléments dans leurs systèmes. Une amélioration du système semble cependant attendue et est discutée dans la suite de ce document.

3 Analyse des points d'amélioration

3.1 Objet de l'identifiant limité à la DEE

Selon la définition opérationnelle de l'identifiant permanent (2014) :« 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. ».

Un point positif de ce passage tient au fait que ce système n'impacte pas ou peu les systèmes producteur. Cependant il présente plusieurs problématiques.

Problématiques :

- La traçabilité liée à cet identifiant est interne au SINP et ne prend pas en compte ce qui se passe en dehors du SINP.
- Dans le cadre de mise à jour de données et métadonnées, si le producteur ne reprend pas les identifiants permanents dans leur système, une correspondance doit être gérée à chaque échange entre producteur et plateforme SINP (parfois complexe quand la standardisation ne prend pas en compte l'identification) ;
- Si le producteur modifie son système d'identification, il devient compliqué de gérer la traçabilité de cette donnée au niveau des plateformes régionales.

« *Le point important concernant l'identifiant unique est qu'il puisse suivre la donnée, où qu'elle soit. Cela implique qu'il faut qu'il existe dès la création de la donnée, donc qu'il soit produit par le producteur de la donnée lui-même. S'il est produit après, dans le circuit SINP, il y a un risque majeur que l'identifiant unique ne soit pas ou mal associé à la donnée source. L'enjeu est de réussir à garantir une unicité tout en ayant une production de ces identifiants totalement décentralisée* »[3].

Proposition : Promouvoir une production de l'identifiant au plus près du producteur

[3] Selon Camille Monchicourt (PN Écrins) – échange mail

3.2 Le format URI de l'identifiant

3.2.1 Légitimité du nom de domaine

Selon la définition opérationnelle de l'identifiant permanent (2014) : « le nom des autorités (domaine des plateformes) [...] sont à définir par qui de droit. »
Problématique :

- Le processus d'habilitation des plateformes n'est toujours pas acté.
- Toutes les régions n'ont pas une plateforme SINP.
- Toutes les régions ne gèrent pas des noms de domaines.
- La pérennité du nom de domaine n'est pas toujours assurée (ex : fusion des régions, changement d'outil vers GINCO). Cas de changement de nom de domaine = complexe à impacter sur le système

Selon la définition opérationnelle de l'identifiant permanent (2014) : « 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.»
Problématique :

- Un annuaire des autorités est un système supplémentaire à gérer (conception, mise en place, maintenance, animation)

3.2.2 Résolution de l'URI

Problématique :

- Même si cela n'est pas obligatoire si l'on regarde les préconisations décrites dans le document de définition opérationnelle de l'identifiant permanent, l'URI laisse à penser qu'il y a une ressource consultable.

Proposition : Ne pas baser l'identifiant sur un URL

4 Proposition d'évolutions du système d'identification

4.1 Objectifs d'amélioration

- simplifier la production et la manipulation de ces identifiants ;
- promouvoir la production au plus proche de l'observation pour faciliter une traçabilité hors SINP ;
- promouvoir l'utilisation et le partage de ces identifiants entre les plateformes du SINP.

4.2 Description des propositions

4.2.1 Parler d'identifiant SINP

L'identifiant en question a pour objectif de gérer une traçabilité vis-à-vis du SINP. Le terme identifiant permanent faisant débat et identifiant unique étant très générique, il est proposé de **parler d'identifiant SINP.** Ainsi, dans le standard SINP, les champs concernant les identifiants permanents seraient rebaptisés en identifiants SINP. Il est à noter que ce principe a déjà été acté dans le standard « Occurrences d'habitat ».

4.2.2 Production au plus proche du producteur de données et chaîne de responsabilité

La production de l'identifiant au plus près du producteur de données facilite la traçabilité des données et limite l'intégration de doublons. Il n'est cependant pas possible, selon les principes non intrusifs du SINP, d'imposer la production de cet identifiant au producteur de données. Ainsi, une chaîne de responsabilité de production de l'identifiant SINP (cf. Illustration 1) est proposée. Elle suit la philosophie suivante : entre le producteur de données et la plateforme nationale du SINP est définie de manière implicite une chaîne d'information, où chaque plateforme constitue les maillons. À chaque maillon de la chaîne de l'information, on vérifie s'il y a un identifiant SINP.

- **Si c'est le cas,** l'identifiant SINP est diffusé vers le maillon suivant.
- **Si ce n'est pas le cas**, après avoir consulté le maillon précédent, l'acteur produit l'identifiant SINP et diffuse cet identifiant dans les deux sens :
  - *en amont*, vers le maillon précédent (l'acteur qui lui a fourni la donnée) pour qu'il le prenne en compte cet identifiant dans son propre système,
  - *en aval*, vers le SINP, pour assurer les fonctions de traçabilité attendues.

Ce système comporte cependant 2 limites :

- En mettant en place cette méthode, une « **interface de correspondance des identifiants** » est créée au niveau de l'acteur qui produit l'identifiant (correspondance entre l'identifiant SINP et l'identifiant d'origine). Si le fournisseur renvoie une mise à jour des données sans y joindre l'identifiant SINP, c'est à « l'interface de correspondance des identifiants » de gérer le rattachement à l'identifiant SINP. Ces interfaces, déjà existantes aujourd'hui, représentent un travail à ne pas négliger.
- Si le producteur de données ne communique pas bien avec les acteurs à qui il fournit la donnée, il y a un risque que plusieurs maillons de la chaîne de l'information produisent parallèlement des identifiants SINP pour les mêmes données. Ce cas devra être identifié dès que possible et la responsabilité de production des identifiants SINP devra être discutée pour éviter cela.

D'un point de vue pragmatique, les 3 cas présentés dans l'Illustration 1 se produiront simultanément sur un même territoire avec des acteurs différents. L'objectif est de promouvoir le cas 1, cas où le producteur de données crée lui même l'identifiant SINP, tout en permettant le cas 2 et 3 de se réaliser.

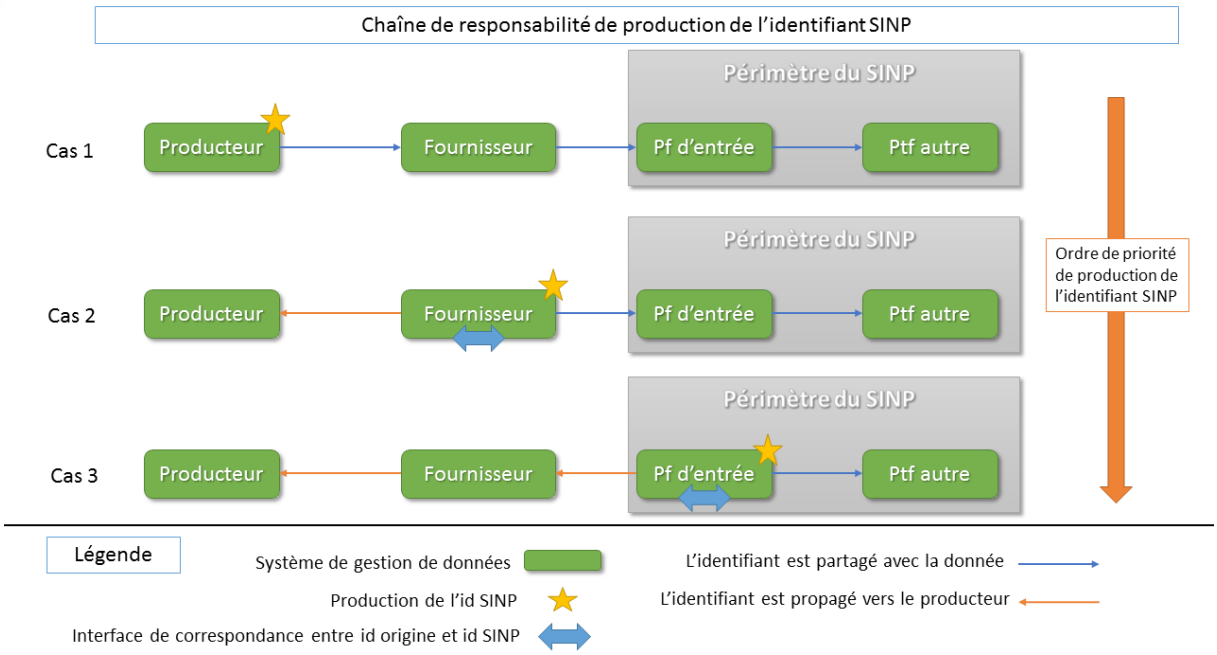

Illustration 1: Chaîne de responsabilité de production de l'identifiant SINP

4.2.3 L'identifiant SINP est uniquement un UUID

Il est proposé que l'identifiant SINP pour les jeux de données et les occurrences soit, dans le standard SINP, un UUID.
Il serait ainsi plus simple à produire (service en ligne de production de UUID, algorithme natif de production de UUID dans les SGBD standards) et ne pose plus la question du choix du nom de domaine ni de la mise en place d'un annuaire des autorités.

Ainsi, les identifiants permanents déjà produits pourront être simplifiés à leur simple UUID.

4.2.4 Généraliser la réflexion à d'autres objets

Il est proposé de généraliser l'utilisation d'UUID et la dénomination d'identifiant SINP pour d'autres objets présents dans les standards SINP comme les regroupements et les cadres d'acquisitions.

5 Conclusion

Ces propositions ont été présentées lors de la réunion du GT Architecture du 10 octobre 2017 et ont été validées sur le principe. Dans le cas d'une validation plus large du GT, comme le prévoit la feuille de route du GT, une prochaine note proposera les spécifications et impacts techniques autour de ce nouveau format d'identifiant.

6 Annexe

6.1 Questionnaire

URL questionnaire : https://docs.google.com/forms/d/1vNksYlrdepIBcxG4QBGD2udao0T_t8K7w69tI8BkTR8 (aujourd'hui fermé)

Dans ce questionnaire, quelle plateforme allez-vous décrire (nom de la plateforme et/ou organismes porteurs)?
1) La plateforme produit et gère des identifiants de jeu de données ....
a) ... unique pour votre système (identifiant du jeu de données)?
□ Oui
□ Non
b) ... unique vis-à-vis du SINP (identifiant permanent du jeu de données)?
□ Oui
□ Non
2) La plateforme produit et gère des identifiants d'observation de taxon ....
a) unique pour votre système (identifiant de l'occurrence de taxon)?
□ Oui
□ Non
b) unique vis-à-vis du SINP (identifiant permanent de l'occurrence de taxon)?
□ Oui
□ Non
3) Intégrez-vous dans votre plateforme des identifiants permanents venant de partenaires (producteurs, plateforme nationale…)
□ Oui
□ Non
4) Utilisez des identifiants permanents ?
□ Oui, pour analyser les doublons à l'entrée de votre plateforme
□ Oui, pour gérer l'unicité des jeux de données
□ Oui, pour gérer l'unicité des observations de taxons
□ Non
5) Partagez-vous des identifiants permanents (produits ou reçus) avec des partenaires?
□ Oui, avec la plateforme nationale
□ Oui, avec des plateformes régionales
□ Oui, avec des producteurs de données
□ Oui, avec des réutilisateurs de données (lors de demande d'extraction, d'accès...)
□ Non

6.2 Acronymes et abréviations

Acronyme / abréviation	Description
Ptf	Plateforme