Note de synthèse des besoins pour la mise à jour des données et métadonnées du SINP

Groupe de travail "Gestion des mises à jour des données et des métadonnées du SINP" - sous-groupe du GT Architecture

Historique du document

Version	Auteurs	Date	Relecteurs	Sections modifiées
1.0	Silvère Camponovo (CBNBP) Solène Robert (PatriNat)	22/06/2022		

Objectifs du GT

- Répondre aux enjeux de mise en œuvre opérationnelle du SINP.
- Définir une organisation cohérente et opérationnelle permettant aux différentes plateformes du SINP de disposer d'une méthodologie commune de gestion des mises à jour et d'intégration des évolutions de référentiels et standards.

Attentes fortes

- Progressivité des solutions pour permettre à toutes les plateformes et producteurs de suivre les évolutions à leur rythme et sans bloquer les plus avancés.
- Simplification des échanges pour l'ensemble des acteurs par la mise à disposition de services partagés : rendre plus automatique les flux de données pour limiter les actions chronophages à faible valeur ajoutée (mises au(x) format(s), animation des allers-retours avec les producteurs et plateformes, adapter, nettoyer, etc.).

Méthodes

La majorité des échanges de données entre producteurs et plateformes s'effectue aujourd'hui sous la forme de processus « annule et remplace » de l'intégralité des données. La volumétrie croissante des paquets de données transmis commence à poser problème, tout comme la difficulté à conserver les identifiants uniques et à toujours bien identifier les doublons.

Certaines plateformes régionales ont commencé à travailler sur des échanges différentiels, c'est à dire contenant uniquement les occurrences ayant été modifiées, corrigées ou supprimées. Ce processus permet d'échanger moins de données (volumétrie), d'optimiser ainsi les traitements nécessaires lors des intégrations (contrôles, normalisation, ...) et de garantir une meilleure traçabilité. Il demande en revanche plus d'ingénierie pour être mis en œuvre.

Propositions

- → Accompagner progressivement les producteurs et plateformes à échanger des données de manière différentielle, en tenant compte des possibilités de chacun.
- → Produire une note de cadrage définissant précisément les modalités des échanges différentiels.

Unicité

La garantie d'unicité de la donnée réside dans la capacité du système à y apposer un identifiant national unique au plus près de sa production et à le conserver tout au long de la vie de cette occurrence. Si cette cible est partagée par l'ensemble de la communauté, elle n'est pas toujours réalisable : certains producteurs, dits « accompagnés », ne disposant pas des moyens nécessaires pour produire et surtout maintenir dans le temps cet identifiant unique accolé à leur observation taxonomique.

Propositions

- → Accompagner les producteurs qui le peuvent à intégrer un identifiant unique dès la première numérisation de la donnée.
- → Lorsque les données ne présentent pas d'identifiant unique, les intégrer via des jeux de données temporels, basés sur une période de date, permettant ainsi de mieux identifier les potentiels impacts des nouvelles entrées à chaque mise à jour et de limiter les doublons.
- → Conserver et partager l'information des identifiants uniques supprimés pour assurer que chaque producteur ou plateforme puisse nettoyer les occurrences concernées dans leurs systèmes.

Gestion des doublons

Corollaire du point précédent, des doublons d'une occurrence au sein des plateformes du SINP existent et d'autres cas seront difficilement évitables. Il faut bien-sûr agir pour limiter les éventualités, il est néanmoins impossible de les empêcher totalement tant la diversité des sources et des acteurs au sein du SINP rend l'exercice complexe, d'autant plus que le dispositif ne pourra jamais contrôler la vie d'une donnée en dehors du SINP : une donnée qui en est extraite peut bien souvent y être réinjectée, parfois avec une forme différente.

Propositions

- → Appliquer les règles de gestion de l'unicité pour limiter les cas.
- → Sensibiliser les acteurs au sein et au-delà de l'écosystème du SINP.
- → Analyser les données avant intégration aux plateformes (cf. expériences et outils existants en France comme à l'international).
- → Réduire les délais d'échanges entre plateformes pour synchroniser les données (et donc pouvoir mieux prévenir les doublons potentiels).

Référentiels et standards

L'évolution des standards (OccTax, OccHab, MTD) et des référentiels qu'ils utilisent (TaxREF, HabREF, Campanule, Organismes, référentiels administratifs) demande aujourd'hui un travail conséquent et coûteux à tous les acteurs pour intégrer ces nouvelles versions et effectuer les traitements de transcription des données qu'elles requièrent. Ces opérations s'effectuent alors bien souvent au détriment d'autres actions (collecte de données, animation de réseaux, validation scientifique des données, etc.)

Propositions

- → Proscrire les suppressions, mettre en place un principe de dépréciation des clés ou entrées.
- → Garantir la compatibilité ascendante entre les versions de référentiels et standards : c'est à dire que chaque mise à jour peut être mise en œuvre aux seins des différents systèmes d'information sans avoir à gérer des problèmes d'intégrité ou des incohérences référentielles. Les évolutions apportées par chaque mise à jour peuvent ainsi être assimilées progressivement par les différents acteurs.
- → Prévoir de stocker (voire de diffuser) la valeur initiale fournie en complément de son expression dans la nouvelle version du référentiel ou de la nomenclatures utilisée.
- → Ne pas modifier la valeur d'origine d'un référentiel au sein de l'occurrence lors d'une montée en version, mais l'enrichir d'une transcription des valeurs référentielles dans la nouvelle version.

Flux multidirectionnels

Les plateformes régionales ou nationale comme certains producteurs souhaiteraient pouvoir réintégrer plus simplement les données partagées au sein de leurs systèmes d'information. L'organisation stratifiée et la lourdeur des méthodes actuelles (préparation des lots de données et métadonnées, envoi, éventuels allers-retours qui nécessitent une intervention humaine) freinent ce processus, un cycle moyen du producteur à la plateforme nationale est actuellement de 1 an. De fait la fraîcheur des informations véhiculées au sein des différentes strates du SINP est très disparate.

Propositions

- → Automatiser les échanges de données, tant ascendants que descendants.
- → Journaliser les flux afin de permettre une interaction plus directe entre les différents acteurs.
- → Faciliter l'accès aux données dès qu'elles intègrent tout point du dispositif SINP.

Scénarios

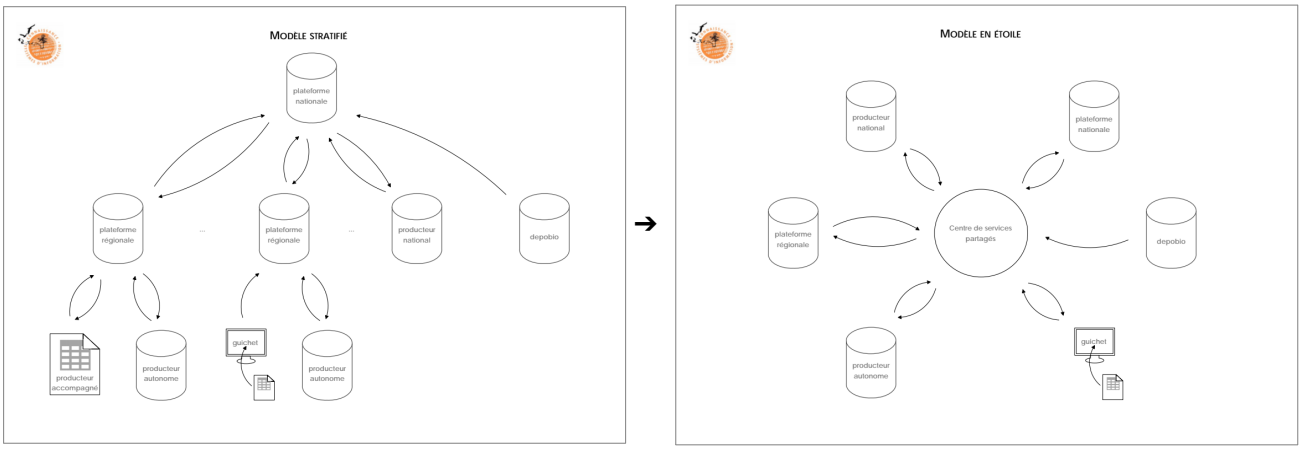
Perspective

L'ensemble des plateformes sont tenues aux mêmes engagements et rencontrent les mêmes problématiques opérationnelles, aussi une mutualisation des outils et des processus permettrait de répondre au besoin de simplification, d'utiliser au mieux les deniers publics et d'accompagner tous les acteurs quelque soit leurs moyens.

Propositions

3 scénarios ont été retenus par le GT permettant de tracer une méthodologie autour d'un socle commun appliqué par tous (scenario 1, dit « plancher », le minimum viable), pouvant être enrichie par des mécanismes plus ambitieux (scenario 2, dit « augmenté », offrant des évolutions à portée ; scenario 3, dit « prospectif », répondant aux enjeux à plus moyen terme).

Ces propositions peuvent se résumer ainsi :



A court terme, solidifier le modèle stratifié actuel pour garantir l'atteinte des exigences du SINP (respect des standards, identifiant unique au plus près de la production de donnée, flux ascendants de données réguliers entre les producteurs et plateformes, mise en place de flux descendants plus fréquents).

A moyen terme, installer un centre de services partagés mutualisé entre les producteurs et plateformes, permettant de réduire le nombre de flux et de les rendre plus directs, d'accélérer l'information de disponibilité des données et donc de prise en compte au sein des différents systèmes, d'avancer vers l'automatisation des intégrations, de journaliser les dépôts et statuts des identifiants uniques pour mieux maîtriser le circuit d'information et disposer en tout point d'une donnée à jour.

Les explications détaillées de ces scénarios et leurs visuels sont consultables dans le rapport Synthèse des besoins pour la mise à jour des données et métadonnées du SINP – 09/07/2021 – S. Camponovo, S. Robert et al. (https://inpn.mnhn.fr/docs-web/docs/download/385349) et son annexe 1 (https://inpn.mnhn.fr/docs-web/docs/download/406069).