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
0.1	Silvère Camponovo (CBNBP) Solène Robert (UMS Patrinat)	02/03/2021		
0.2	Silvère Camponovo (CBNBP) Solène Robert (UMS Patrinat)	04/06/2021	Judith Panijel (UMS Patrinat) Frédéric Vest (UMS Patrinat)	Corrections et apports suite réunion cercle 1
1	Silvère Camponovo (CBNBP) Solène Robert (UMS Patrinat)	09/07/2021	Julie Delauge (CEN PACA) Paul Fromage (FAUNA) Camille Monchicourt (PN Ecrins) Valérie Raevel (DREAL Haut-de-France) Paula Spinosi (CBNC) Manon Zeyer-Linden (DREAL Grand-Est)	Corrections et apports suite relecture cercle 1 et 2

1. Contexte

La majorité des plateformes du SINP sont installées et disposent maintenant de plusieurs années d'exercice de transmission de données entre acteurs et plateformes.

Le groupe de travail "Gestion des mises à jour des données et des métadonnées du SINP" doit permettre de répondre aux enjeux de mise en œuvre opérationnelle du système d'information du SINP. Il s'inscrit dans le cadre du groupe de travail "Architecture".

Le présent document a pour objectif de synthétiser et mettre à profit les retours d'expérience de l'ensemble des acteurs afin de proposer un référentiel de bonnes pratiques et une méthodologie* commune pour assurer la transmission des données et leur qualité.

Le groupe de travail pourra se saisir de ces propositions pour établir les principes d'une organisation commune des mises à jour de données et métadonnées au sein du SINP.

* la notion de méthodologie est employée dans la suite de ce document au sens de l'ensemble des méthodes et techniques mises en œuvre dans un domaine particulier.

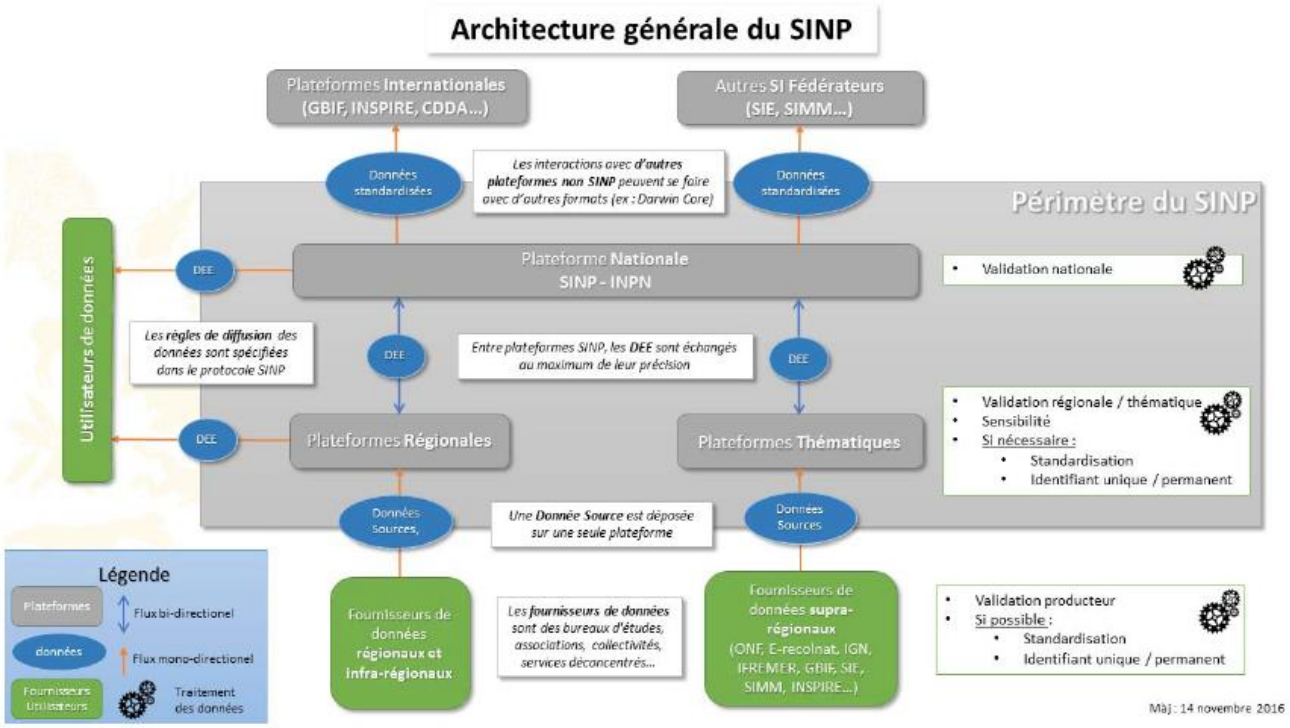
Pour mémoire

Mandat

Les objectifs fixés au groupe de travail “Gestion des mises à jour des données et des métadonnées du SINP” sont de :

- Clarifier l’organisation des échanges de données dans le cadre des mises à jour, corrections, restructurations des données ou métadonnées ;
- Clarifier l’organisation dans le cas d’évolution des référentiels ou standards d’échange ;
- Définir une organisation cohérente et opérationnelle permettant aux différentes plateformes du SINP de disposer d’une méthodologie commune de gestion des mises à jour.

Schéma d’architecture du SINP – version 2017



Source : « Protocole du Système d’Information sur la Nature et les Paysages » - 28/09/2017 - NOR : TREL1704934N - annexe G p.41

Certaines notions présentées ci-dessus ne sont plus en vigueur ou doivent être revues (DEE, plateformes thématiques par exemple). Toutefois ce schéma permet de rappeler l’organisation générale des échanges au sein du SINP aujourd’hui.

2. Cadrage

1. Consultations

Le recensement des pratiques et des travaux existants relatifs à la mise à jour des données et métadonnées au sein du SINP s'est déroulé en deux temps : un appel à contribution suivi d'une série d'entretien auprès d'acteurs du SINP.

Liste des acteurs consultés :

- La plateforme régionale PDL (05/10/20, contribution – Arnaud LE NEVE, DREAL PDL)
- La plateforme régionale Réunion (13/10/20, contribution – Valentin LE TELLIER, DEAL Réunion)
- La plateforme régionale NA (15/10/20, contribution – Paul FROMAGE et Nathan HAUQUIN, FAUNA)
- La plateforme régionale PACA (25/11/20, entretien – Julie DELAUGE et Géraldine KAPFER du CEN PACA, Jean-Michel GENIS, Lucile VAHE et Jean-Pascal MILCENT du CBNA)
- La plateforme régionale CVL (10/12/20, entretien – Mathieu WILLMES, DREAL CVL)
- Le Conservatoire du Littoral (30/11/20, entretien – Pascal CAVALLIN, Pascal BOULESTEIX et François CAVALLO)
- L'Office National des Forêts (01/12/20, entretien – Francesco ROMANO)
- L'équipe du GBIF France (09/12/20, entretien – Sophie PAMERLON et Sylvain MORIN)
- L'équipe GeoNature (10/12/20, entretien – Camille MONCHICOURT du PN Ecrins et Amandine SAHL du PN Cévennes)
- L'équipe données de l'UMS PatriNat (25/11/20, entretien – Frédéric VEST)

Le présent document n'a pas pour vocation de dresser un état des lieux exhaustif, ni de présenter une synthèse de ces différents échanges, mais de se concentrer sur les objectifs fixés par le mandat.

2. Périmètre

Les échanges se sont concentrés autour des modalités actuelles mises en place pour assurer les flux de données au sein du SINP, les difficultés rencontrées, les points de vigilance identifiés ainsi que les évolutions projetées et les attentes du réseau concernant les principes de transmission et de mises à jour des données et métadonnées.

L'ensemble de la chaîne de transmission est considéré :

- Flux entre producteurs et plateformes du SINP (régionales et nationale) ;
- Flux entre plateforme régionale et plateforme nationale ;
- Flux entre plateforme nationale et plateformes internationales (GBIF particulièrement).

Tous les flux ont été étudiés de manière bidirectionnelle.

3. Principes de réflexion

Les échanges menés auprès de l'ensemble des acteurs ont permis de dresser la situation suivante :

- Une réalité : les plateformes n'avancent pas au même rythme (acteurs multiples, moyens humains disponibles et sources de financement assez disparates) ;
- Un constat : l'ensemble des plateformes ont les mêmes problématiques, les mêmes enjeux, les mêmes besoins, mais des SI et des outils internes différents qui rendent les solutions mises en œuvre spécifiques du contexte de chacun ;
- Une volonté : tous les acteurs attendent que l'architecture du SINP intègre la diversité des

situations (cf. ci-dessus), afin qu'aucune plateforme ne soit bloquée en cas d'évolution des principes de transmission, tout en donnant dans le même temps aux acteurs la possibilité d'avancer au rythme de leurs capacités.

Ces éléments induisent une organisation agile et un équilibre complexe qui doivent s'intégrer dans un cadre permettant de définir les limites entre ce qui est nécessaire et obligatoire pour le bon fonctionnement des transmissions de données et métadonnées au sein du SINP et ce qui est possible ou souhaitable pour augmenter les services et fonctionnalités selon les besoins et capacités de chacun.

Ces réflexions fondent les enjeux et les principes retenus pour établir les propositions ci-après et l'attente du livrable de ce GT :

→ Production d'une méthodologie basée sur un socle commun appliqué par tous (le minimum viable), pouvant être augmentée par des mécanismes plus ambitieux.

4. Vocabulaire partagé

Donnée

Une donnée est une information structurée selon un standard d'échange permettant de décrire une observation ou un évènement au travers d'un ensemble d'attributs associés.

Métadonnée

Une métadonnée décrit de manière structurée une source et un contexte d'acquisition d'une donnée.

Cadre d'acquisition (CA)

Le cadre d'acquisition est une métadonnée qui décrit le contexte stratégique et opérationnel d'un programme ou d'un projet. Il conduit à la production d'un ou de plusieurs jeux de données.

Jeu de données (JDD)

Le jeu de données est un regroupement de données homogènes au regard de la compréhension et de l'usage qui peut être fait des données. Il permet par exemple de décrire la méthode de collecte ou le protocole mis en œuvre.

Occurrence

Niveau élémentaire d'information, structuré en plusieurs entrées, dont les valeurs décrivent une donnée (observation, évènement) ou une métadonnée (JDD, CA).

Mise à jour

Ensemble d'actions modifiant les valeurs d'une occurrence, pouvant être de natures différentes. On distingue principalement les actions de modification, correction, validation, et résultats de calculs.

Validation

Assignation d'un statut à l'occurrence qualifiant la fiabilité de la donnée, associé à des informations de contexte (validateur, méthode, date).

A noter : les standards d'occurrence du SINP reconnaissent 3 niveaux de validation : producteur, régional, national.

Modification

Mise à jour d'une ou plusieurs valeurs de l'occurrence impactant sa validité.

A noter : la grande majorité des personnes interrogées identifie 3 champs principaux dont la mise à jour doit remettre en cause systématiquement la validité d'une occurrence : date, localisation, espèce. Selon les groupes et le protocole de collecte suivi d'autres champs pourraient également être retenus.

Correction

Mise à jour d'une ou plusieurs valeurs de l'occurrence sans impact sur validité.

A noter : les corrections sont souvent apportées pour enrichir la donnée plus que pour mettre à jour une valeur existante.

Restructuration

Evolution du format d'expression de l'occurrence (changement du typage des entrées, modification de l'organisation, de la structuration des informations au sein de l'occurrence).

Flux de données ascendants et descendants

Un flux de données est un échange de données contractualisé entre deux acteurs qui définit les modalités d'exercice de cet échange.

Le SINP est un dispositif décentralisé constitué de plusieurs niveaux : plateformes nationale et régionale, producteurs. On entend par flux ascendant l'échange de données entre un producteur et une plateforme régionale, ou une plateforme régionale et la plateforme nationale. A l'inverse, un flux descendant décrira un échange de données entre une plateforme régionale et un producteur, ou la plateforme nationale et une plateforme régionale.

3. Problématiques

Le mandat expose une liste de questions (*reprises en italique ci-dessous*) permettant de cadrer et décrire les problématiques posées par les mises à jour de données et métadonnées. Elles sont reprises ci-dessous par thèmes et complétées de sujets ayant émergés au fil des entretiens.

Chaque thème développe les principes partagés, les complexités identifiées, les attentes ou contraintes et propose des pistes méthodologiques.

1. Responsabilité

Qui est en droit de mettre à jour une donnée, à quel moment, dans quel outil, à quelle fin ?

Corrections et modifications

Pour assurer la pérennité des mises à jour d'une donnée comme leur transmission à l'ensemble des intéressés, il est convenu que les actions de modification ou correction doivent être réalisées sur les données sources et ainsi assurer la conservation d'une donnée dont le contenu est maîtrisé de bout en bout de la chaîne. Tant que possible les mises à jour doivent donc être effectuées par le « propriétaire » de la donnée.

Toutefois cette bonne pratique est concrètement difficile à mettre en œuvre. La réalité et la diversité des producteurs poussent à dissocier 2 cas :

- Le producteur « autonome » doté d'une base de données et disposant de moyens pour assurer son maintien et sa maîtrise à long terme, rendant possible la structuration de processus d'échanges avec la plateforme réceptrice.
- Le producteur « accompagné » qui n'est pas doté d'une base de données et/ou ne dispose pas de moyens pour assurer son maintien, nécessitant de fait une adaptation particulière et régulière de la plateforme réceptrice.

Quel que soit le cadre, les modifications et corrections de données ne doivent être effectuées que par des personnes ayant un rôle bien défini au sein du SINP, par ordre de priorité : le propriétaire de la donnée tant que possible, un administrateur de base de données du producteur, un administrateur ou un expert de la plateforme qui a procédé à l'intégration des données au sein du SINP, dans des cas extrêmes à limiter, un administrateur ou un expert de la plateforme nationale.

Lorsque les mises à jour ne peuvent être effectuées directement par le propriétaire, une chaîne de communication doit être mise en œuvre pour lui transmettre les raisons et informations modifiées ou corrigées (afin de lui permettre de conserver ces mises à jour pour de prochains usages de la donnée).

Propositions :

→ Cas producteur « autonome » :

- Sollicitation du producteur pour mise à jour de la donnée et si nécessaire, en attendant la modification ou correction, le statut de validité régionale est assigné en invalide ;
- Transmission de la mise à jour via les canaux classiques de transmission entre le producteur et la plateforme.

→ Cas producteur « accompagné » :

- Modification/correction de la donnée directement au sein de la plateforme qui a procédé à l'intégration des données au SINP par un administrateur ou un expert ;
- Transmission des informations mises à jour au producteur via la production d'un jeu de données retour par exemple.
- Transmission de la mise à jour par la plateforme intégratrice aux autres plateformes via les canaux classiques de transmission inter-plateformes.

Validation

La validation est un processus à forte valeur ajoutée, opéré par un nombre d'acteurs restreints répartis tout au long de la chaîne de transmission du SINP (producteur, régional, national). La diffusion de cette information et de sa mise à jour est de fait complexe à mettre en œuvre : c'est un processus essentiellement ascendant aujourd'hui.

Or pour garantir l'atomicité de l'occurrence (la maîtrise de son contenu) et valoriser les informations produites par le processus de validation, une transmission descendante facilitée est attendue (cf. « Transmission multidirectionnelle des mises à jour » ci-dessous).

2. Méthodes de mise à jour

Quelles sont les différentes méthodes de mise à jour possibles : par jeu de données, par lot ou de manière unitaire, en présentant le document complet ou uniquement les champs nécessitant mise à jour ?

Quels sont les différents canaux de mise à jour mis à disposition ?

Contexte

La majorité des plateformes et producteurs échangent aujourd'hui leurs données sous forme de lots souvent composés de plusieurs jeux de données, contenant l'ensemble des occurrences et leurs valeurs. Le processus de mise à jour appliqué consiste à remplacer les occurrences précédentes de chacun des jeux de données par les nouvelles. Cette méthode permet de couvrir à la fois la création, la mise à jour et la suppression des occurrences au sein d'un jeu de données. C'est une stratégie simple dont la mise en œuvre nécessite peu d'ingénierie, qui peut donc être assimilée et soutenue largement par l'ensemble des acteurs.

Elle pose toutefois certaines limites : volumétrie croissante des paquets de données échangés, conservation de l'identifiant unique et difficulté à identifier les doublons.

Techniquement, les mises à jour de jeux de données par lot selon la méthode « annule et remplace » sont assez lourdes en termes de temps de calcul et d'écriture côté machine. Or le contenu réellement mis à jour au sein de ces volumes d'informations est faible, aussi il serait donc préférable de réserver cette méthodologie à des cas particuliers où le contenu d'un jeu de données est largement revu.

Certaines plateformes régionales opèrent déjà des mises à jour incrémentales avec leurs principaux producteurs de données, c'est-à-dire qu'ils n'échangent que les occurrences ayant été ajoutées, modifiées, corrigées, supprimées ou validées. Toutefois ces traitements de données nécessitent la mise en place de stratégie plus fine, une bonne coordination entre les acteurs et bien-sûr des ressources pour développer et maintenir ces systèmes de transmission plus complexes.

Les propositions doivent de fait intégrer la diversité des acteurs et de leurs moyens, chaque maillon doit être analysé indépendamment (producteur / plateforme régionale, plateforme régionale / plateforme nationale, etc.)

Enfin, les mises à jour, comme la diffusion des données, sont aujourd'hui des processus essentiellement maîtrisés dans le sens ascendant. Si toutes les plateformes régionales, comme certains gros producteurs, sont intéressées, très peu sont outillées pour assurer des transmissions descendantes de données depuis la plateforme nationale (cf. « Transmission multidirectionnelle des mises à jour » ci-dessous).

Stratégies de transmission

Les solutions de transmission de données peuvent se décrire comme une combinaison de 3 facteurs :

- Le paquet, c'est-à-dire l'ensemble de données choisi :
  - Paquet complet (toutes les occurrences d'une banque/base de données) ;
  - Paquet par cadre d'acquisition ou jeu de données (au sens SINP) ;

- A l'occurrence (paquet ne contenant qu'une occurrence).

- Le canal, c'est-à-dire le mode de transmission :
  - Mail : envoi de fichiers au travers d'une communication écrite, pouvant être expédiée à une boîte générique ou directement à la personne en charge de l'intégration ;
  - FTP ou cloud : espace de dépôt de fichiers, partagé entre les acteurs, assurant une transaction sécurisée ;
  - Guichet : interface permettant de déposer ses paquets de données en assurant à la fois une transaction sécurisée, une information automatisée aux intéressés, un traçage et suivi de sa prise en compte. Un guichet peut également permettre de structurer le dépôt en obligeant le déposant à définir l'objet de son dépôt (et donc de pré-alimenter, voire associer directement, les métadonnées aux données) ;
  - Interface web : application permettant à un producteur de données de saisir ses occurrences de données et métadonnées directement au sein d'une plateforme SINP.
  - Webservices : interface de programme sécurisée qui définit et structure des échanges d'informations entre applications, et permet à de multiples systèmes hétérogènes d'interagir, par exemple au travers de scripts ou d'un flux applicatif.
- Le véhicule, c'est-à-dire le vecteur de transmission de l'information :
  - Dump ;
  - Fichiers (.txt, .csv, .shp, .json, etc.) ;
  - Scripts (pour les échanges base à base ou les webservices) ;
  - Programme ou fonction intégrée à l'application d'une plateforme (pour l'interaction avec des webservices par exemple).

Stratégies d'intégration

Les méthodes identifiées pour intégrer les nouvelles données ou leurs mises à jour au sein des bases de données des plateformes (ou des producteurs) sont :

- Méthode intégrale : annule et remplace les occurrences que couvre le nouveau paquet de données : les occurrences sont d'abord supprimées de la base de données (filtre selon les critères définissant le paquet), puis les données contenues dans le nouveau paquet sont importées.
  - Intérêts : simple à mettre en œuvre, peu coûteux à maintenir ;
  - Vigilance : la couverture des données contenues dans le nouveau paquet doit être bien maîtrisée (la suppression des occurrences correspondantes existantes en base ne doit pas entraîner de perte de connaissance) ;
  - Limites : conservation de l'identifiant unique pas évidente (création d'un nouvel identifiant si non produit par le transmetteur et nécessité de conserver l'existence des identifiants précédents, pour les comparer et/ou diffuser les changements).
- Méthode différentielle : le paquet ne contient que les nouvelles occurrences ou leurs mises à jour (concepts du standard SINP), l'ensemble des informations de ces occurrences sont implémentées en base de données (mode upsert)
  - Intérêts : plus faible volumétrie d'occurrences échangées, traitement d'intégration plus rapide, meilleure traçabilité ;
  - Vigilance : processus plus complexe nécessitant des compétences informatiques pérennes au sein des structures concernées, en effet ce cas n'est envisageable que si les 2 parties peuvent à la fois générer un paquet ne contenant que les occurrences créées

ou mises à jour et intégrer un paquet de manière incrémentale ;

- Limites : si le producteur ou la plateforme ne peut générer un paquet restreint aux occurrences créées ou mises à jour depuis la dernière transmission, cette méthode n'apporte pas de gain substantiel. En effet, l'intégration par incrément d'un paquet contenant l'ensemble des données nécessite des traitements successifs avec des requêtes plus coûteuses que dans le cas annule et remplace (identification des écarts, puis applications). Par ailleurs la mise à jour de données par cette méthode n'est possible que si le producteur produit un identifiant unique sur lequel la plateforme peut s'appuyer pour opérer les traitements.

- **Méthode incrémentale :** le paquet ne contient que les nouvelles occurrences ou leurs mises à jour, sous forme de documents partiels ne contenant que les informations ayant subies un changement. Seules ces évolutions sont implémentées en base de données.
  - Le principe de mises à jour d'occurrences en ne présentant que les valeurs des champs modifiées n'a pas été retenu comme une solution possible à court ou moyen terme. Si elle présente l'avantage de la frugalité des volumes échangés et la précision des actions opérées sur les occurrences à reproduire tout au long de la chaîne de transmission, elle nécessite d'avoir intégré au sein des différents systèmes d'informations et de leurs bases de données un traçage fin des mises à jour, voire une journalisation de chaque occurrence. Ces stratégies sont à la fois assez lourdes à mettre en œuvre, génèrent des volumes importants d'informations à stocker en redondance, et/ou des calculs conséquents et couteux pour produire les journaux de mises à jour.

Critère de viabilité des propositions

Le choix d'une méthode d'intégration impacte fortement le contenu des paquets d'informations à transmettre. Si théoriquement toutes les associations de stratégies de transmission / intégration sont possibles, plus les méthodes d'intégration sont évoluées, plus les stratégies de transmission fines sont facilitées, et inversement.

Or, les solutions pouvant être mise en œuvre sont fortement induites par les conditions d'exercice et les moyens à disposition, en considérant bien les moyens de l'ensemble des parties concernées par le maillon de transmission étudié.

L'ensemble des acteurs s'accordent donc sur le fait qu'il est souhaitable que toutes les formes de paquets, de canaux, de véhicules et de méthodes restent possibles particulièrement entre les producteurs et plateformes régionales où c'est bien la plateforme qui doit s'adapter au producteur pour faciliter la vocation première du SINP : mutualiser et diffuser des données.

Les acteurs des plateformes s'entendent également sur le fait que les échanges inter-plateformes pourraient tendre vers une méthode différentielle rapidement, ce qui permettrait à la fois d'augmenter la fréquence de partage et de diffusion des données, tout autant que faciliter des flux descendant depuis la plateforme nationale.

Les propositions s'attacheront donc à mesurer leur faisabilité et leur maintenabilité au regard de la spécificité des différentes situations rappelées ci-dessous :

- Maillon de la chaîne de transmission :
  - Producteur autonome ou accompagné / plateforme régionale ;
  - Plateforme régionale / nationale ;
  - Plateforme nationale / supranationale ;
- Flux ascendant et descendant ;
- Données et métadonnées.

3. Unicité et identifiants uniques

Comment assurer dans chacun des cas l'unicité des données transmises et limiter les risques de doublons ?

Est-il possible d'assurer la pérennité des identifiants uniques dans ces contextes de mise à jour (en prenant soin d'analyser s'il faut gérer de façon différente les identifiants uniques de la donnée et de la métadonnée) ?

Contexte

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.

Toutefois cette bonne pratique ne peut pas aujourd'hui être mise en œuvre tout au long de la chaîne de transmission, pour les mêmes raisons que celles évoquées au chapitre « 1. Responsabilité ». De fait, si l'essentiel des acteurs peuvent assurer et ont déjà mis en œuvre l'identifiant unique national au sein de leurs systèmes d'information, il est nécessaire de prendre en compte les producteurs de données ne disposant pas de base de données ou de moyens pour les maintenir.

Or, c'est cette réalité qui rend extrêmement complexe la conservation des identifiants uniques et la gestion des doublons pour les occurrences transmises par ces producteurs.

→ Il serait intéressant de pouvoir quantifier les volumétries de données et métadonnées concernées par cette absence d'identifiants uniques par catégorie de producteur afin de prioriser et adapter les actions à mettre en œuvre au regard de leur impact potentiel sur la génération de doublons.

Stratégie de conservation de l'unicité

Cas d'un producteur autonome :

→ Inciter le producteur à assigner un identifiant unique dès la création d'une occurrence dans ses bases de données (poursuite des actions de sensibilisation des producteurs par les plateformes, accompagnement technique des producteurs) ;

→ Sinon la plateforme réceptrice doit assurer la création des identifiants uniques et gérer une correspondance entre les identifiants proposés par le producteur et ces identifiants uniques. Cette correspondance lui permettra lors des futurs échanges de paquets d'identifier les occurrences à mettre à jour de celles à créer ou supprimer.

Cas d'un producteur accompagné :

→ Flux ascendant : la plateforme produit un identifiant unique pour chaque occurrence et transmet en retour au producteur le paquet de données augmenté des identifiants assignés (*a minima* pour ses besoins d'échanges avec d'autres partenaires afin de repérer cette donnée lors de l'apport d'un autre acteur) ;

→ Méthodes : si le producteur ne peut stocker durablement et faire connaître ses mises à jour en réutilisant les identifiants uniques, seule la méthode de mise à jour par paquet complet et d'intégration en mode intégrale (cf. point 2 ci-dessus) permettent d'assurer qu'aucun doublon ne sera produit.

Stratégie de conservation de l'identifiant unique

Cas général :

→ Un identifiant unique est apposé sur chaque occurrence au plus près de sa production, il est de fait intégré aux flux de données du SINP et sa conservation est assurée tout au long de la chaîne de transmission (l'atomicité et la traçabilité de l'occurrence sont ainsi garantis dans le temps).

Cas des données d'un producteur accompagné :

- → Si un producteur ne peut stocker durablement et distinguer ses anciennes données des nouvelles à l'aide de cet identifiant unique lors de ses remontées d'informations successives, alors les identifiants uniques assignés aux occurrences de données de ce producteur ne peuvent être maintenus dans le temps lors des mises à jour.
- → Dans la cadre des envois de nouvelles données, il est conseillé de procéder par paquets de données correspondant à une nouvelle période d'observation. Cette bonne pratique si elle n'est pas infaillible évite la perte des identifiants uniques déjà produits pour les occurrences antérieures et la génération de doublons éventuels (par exemple le producteur accompagné envoie en début d'année n toutes les nouvelles observations effectuées l'année n-1).
- → Afin d'éviter toute confusion, la plateforme réceptrice de ces données doit conserver l'historique *a minima* des anciens identifiants afin que la trace de leur existence soit conservée. Un flux permettant de tracer ces identifiants dépréciés doit donc être mis en œuvre entre les plateformes du SINP.

Données et métadonnées

Le processus et les incidences des mises à jour sont en tout point identiques pour les données et les métadonnées.

La production de métadonnées est aujourd'hui réalisée en majeure partie par les acteurs (producteurs autonomes ou plateformes) disposant de bases de données et de moyens pour les maintenir. Aussi la conservation de l'unicité et des identifiants uniques est bien plus maîtrisée pour les métadonnées.

Par ailleurs, lorsque le producteur ne met pas à disposition de métadonnées avec son jeu de données, celui-ci est produit par la plateforme réceptrice des données. De fait, elle assure également le maintien de ce lien et de la cohérence entre les données et les métadonnées lors des mises à jour de données proposées par ce producteur.

Gestion des doublons

L'existence de doublons (voire de multiples plus importants) d'une occurrence au sein des plateformes du SINP doit être assumée :

- ♦ La conservation de l'unicité et celle de l'identifiant unique ne peuvent être garanties à 100% (cf. ci-dessus).
- ♦ L'organisation décentralisée du SINP permet techniquement la mise à disposition de données identiques sur différentes plateformes – même si ceci est tout à fait déconseillé (il arrive par exemple qu'un bureau d'étude partage à la fois ses données sur une plateforme régionale, les mette à disposition d'un producteur national « tête de réseau » et les dépose sur Depobio).
- ♦ Il est impossible de connaître la vie d'une donnée en dehors du SINP (par exemple une donnée extraite depuis une plateforme peut tout à fait être retraitée, voire agrégée, puis réinjectée plus tard au sein du dispositif SINP, potentiellement par un acteur différent). De fait, même si une occurrence est extraite avec son identifiant unique, nous ne pouvons être assurés qu'elle ne sera pas réimportée sans.

La grande diversité des sources et des acteurs rend l'identification de tous les cas potentiels assez complexe : les exemples ci-dessus sont de fait loin d'être exhaustifs.

Propositions :

Cette complexité rend la résolution de la problématique des doublons improbable (au moins à moyen terme) et appelle plutôt à mettre en place des stratégies de prévention.

Plusieurs axes de travail peuvent être menés pour limiter les risques et les doublons :

- Connaissance : documentation des expériences et cas de doublonnage d'occurrence connus (une meilleure appréhension permettant de mieux projeter et anticiper) ;
- Méthodes :
  - Rédaction de bonnes pratiques (cf. premières propositions ci-dessus relatives aux identifiants uniques, aux échanges entre producteurs et plateformes) ;
  - Processus d'analyses des données avant intégration (bonnes pratiques existantes à partager) ;
- Communication :
  - Sensibilisation des producteurs à cette problématique par l'ensemble des plateformes SINP ;
  - Déclinaison des bonnes pratiques et identification pour chaque producteur du tiers de confiance garant de l'unicité et de l'identifiant unique ;
  - Campagne de sensibilisation au-delà de l'écosystème SINP ;
- Outils : mise à disposition d'une méthodologie voire d'un algorithme permettant d'analyser les occurrences d'une plateforme pour identifier et traiter les doublons (cf. expérience et outil de détection de cluster du GBIF).

Traçabilité des suppressions

Une occurrence peut être supprimée pour de multiples raisons (erreur de transmission, résolution d'un doublon, cas des transmissions par méthode intégrale avec création d'identifiant unique par la plateforme nécessitant la suppression de toutes les occurrences d'un lot de données, etc.).

Si l'historisation des données portées par une occurrence supprimée n'apporte pas d'intérêt pour le SINP (cf. point 10 ci-dessous), la conservation de l'information de sa suppression est par contre cruciale : à la fois pour assurer la compatibilité des données diffusées entre les plateformes et pour vérifier le statut d'une occurrence avant de l'intégrer au sein de son SI.

Proposition :

La mise en place d'un service de dictionnaire des identifiants uniques supprimés (cf. proposition de journalisation des échanges au point 9 ci-dessous) permettrait à tous les acteurs du SINP de pouvoir à tout moment s'assurer de la bonne existence d'une donnée, tout autant que de pouvoir nettoyer leurs bases des occurrences qui auraient été supprimées.

4. Impact sur la validité

Quel est l'impact d'une mise à jour sur le niveau de validité (régional et national) ?

Comment gérer les évolutions des niveaux de validité ? À quelle fréquence faut-il actualiser le calcul du niveau de validité ?

Principes partagés

Les seules mises à jour pouvant induire une revue du niveau de validité d'une occurrence sont les modifications : toute modification d'une donnée devrait donc induire automatiquement la suppression de la valeur actuelle du statut de validation afin de faire repasser cette donnée dans le circuit de validation.

Les évolutions de validité d'une donnée ne sont pas considérées comme des modifications ou des corrections sur cette occurrence, mais comme une qualification de la fiabilité de ses informations.

L'historisation des statuts successifs pour chaque niveau de validité (producteur, régional, national) apporte des informations intéressantes d'autant plus si elles peuvent être associées à des commentaires. Si ces fonctionnalités sont aujourd'hui très peu mises en œuvre au sein des plateformes, il y a un intérêt

commun à développer ces réflexions.

Toutes ces informations ont vocation à être diffusées et transmises lors des échanges de données entre les acteurs (ascendant et descendant).

Cas des intégrations de données via la méthode intégrale

Le fait de nettoyer la base de données réceptrice ne permet pas d'assurer la maîtrise du contenu des données qui seront par la suite réimportées. De fait les statuts de validation des données supprimées ne peuvent être conservés et réattribués sans avoir recours à des traitements complexes.

La récupération des statuts de validation d'une occurrence mise à jour étant tout à fait souhaitable pour conserver la forte valeur ajoutée induite par ces informations, il est souhaitable de limiter le recours à la méthode intégrale pour limiter les coûts de traitement.

Calcul de validité automatique

Les calculs automatiques permettant d'assigner un statut de validation à une occurrence sont de plus en plus fréquents et forment un champ de recherche partagé par plusieurs acteurs du SINP. Leur intérêt est reconnu et leur apport est attendu pour concentrer le travail de validation manuelle des experts sur les données qui sortent du lot (espèce découverte sur un territoire, statut de protection, espèces difficilement identifiables, etc.)

Aujourd'hui, seuls quelques rares producteurs ou plateformes régionales et la plateforme nationale effectuent ce type de calculs avec des stratégies parfois différentes.

L'évolution de la connaissance a une incidence sur les résultats de validation produits par les algorithmes de validation automatique, aussi les rejouer périodiquement peut permettre de requalifier certaines données. Une réactualisation annuelle des statuts de validation par les algorithmes serait souhaitable.

5. Impact sur la sensibilité à la diffusion

Comment gérer les évolutions des niveaux de sensibilité ?

Aucun impact n'a été relevé concernant l'incidence des mises à jour de la sensibilité sur une donnée. En effet celle-ci est définie par l'appartenance d'une espèce au référentiel de sensibilité du SINP. Si une occurrence est concernée par un niveau de sensibilité, cela ne fait pas évoluer intrinsèquement les informations qu'elle contient.

Seule la modification d'une occurrence peut induire un traitement particulier si le niveau de sensibilité est stocké comme attribut de la donnée : une modification doit entraîner un recalcul du niveau de sensibilité pour mettre à jour le champ concerné lors de la propagation de la mise à jour en base de données.

6. Métadonnées

Comment gérer les mises à jour de métadonnées en cas de nouvel envoi ?

Les évolutions apportées aux métadonnées n'ont pas d'impact sur les données. En effet, les mises à jour sur les métadonnées ne remettent pas en cause le cadre contextuel qu'elles définissent, leur mise à jour n'induit a priori pas de traitement particulier.

A noter :

- Les mises à jour de métadonnées sont majoritairement des enrichissements, c'est-à-dire des compléments apportés aux attributs de l'occurrence, beaucoup plus rarement des corrections ;

- La suppression d'une occurrence de métadonnée est à proscrire, car elle engendre des données orphelines. Une métadonnée doit par défaut être inactivée ou dépréciée, et les données qui y sont associées rattachées (progressivement) à une métadonnée valide.

La mise à jour d'une occurrence peut à l'inverse avoir des impacts sur ses métadonnées : évolution de la période de collecte, du périmètre étudié, des groupes taxonomiques concernés, des organismes collecteurs, etc. Aussi la livraison d'un lot de données pour mise à jour devrait tant que possible être associée à ses métadonnées.

7. Référentiels

Comment gérer les changements de référentiels, notamment administratifs ou taxonomiques ?

Périmètre

Les données et métadonnées du SINP sont liées à plusieurs référentiels structurants :

- TaxREF
- HabREF
- COG (Code Officiel Géographique, i.e. référentiel administratif et communal, INSEE)
- Campanule
- Organismes

Impacts

Les mises à jour de ces référentiels ont des incidences variables sur les occurrences de données et métadonnées existantes :

- Ajout d'une nouvelle entrée dans le référentiel : sans impact direct, toutefois les occurrences non assignées à une valeur référentielle pourraient être réanalysées et mises à jour si cette nouvelle entrée le permet (par exemple, cas des nouveaux CD_NOM ajoutés dans TaxREF) ;
- Correction d'une entrée (sans changement de sens, par exemple erreur de typo sur un libellé) : sans impact ;
- Modification d'une entrée (changement du sens de la valeur, par ex. cas des taxons évoluant vers plusieurs entrées ou redécoupage de communes) : pour être exprimée dans le nouveau référentiel, la donnée doit être analysée et réattribuée par un expert ;
- Suppression d'une entrée : la référence disparaissant, la donnée doit être obligatoirement analysée et réattribuée à une valeur valide du nouveau référentiel ;

Les cas de modification et de suppression génèrent donc un travail conséquent et coûteux, particulièrement concernant TaxREF et COG dont l'usage concerne quasiment toutes les données (la problématique est similaire mais plus récente concernant HabREF).

Les mises à jour des référentiels Organismes et Campanule semblent à ce jour générer moins d'impacts.

Contexte

L'évolution des référentiels posent plusieurs questions tant opérationnelles que scientifiques :

- Adéquation entre la fréquence et les moyens : il n'est pas aisé aujourd'hui pour les acteurs du SINP de suivre l'ensemble de ces évolutions. La plupart des producteurs autonomes et certaines plateformes ne portent ces évolutions au sein de leur système que tous les 2 ans. Ce rythme semble plus en adéquation avec les moyens qu'ils peuvent mobiliser.

- Expertise concentrée sur peu d'acteurs : les réattributions sont le plus souvent gérées par les gros producteurs ou les plateformes. Les producteurs accompagnés ne sont jamais sollicités car ils manquent de moyens pour suivre les évolutions et réanalyser leurs données.
- Responsabilité : la mise à jour d'une valeur référentielle d'une occurrence, comme toute mise à jour, devrait être assurée par le producteur / collecteur de la donnée (cf. point 1 ci-dessus). En plus du problème de pérennité, la réattribution d'une nouvelle valeur de référentiel demande la plupart du temps une réinterprétation des informations de la donnée, induisant un risque de dénaturation la donnée.

Proposition de cadre

- Ne pas modifier l'occurrence d'origine, éventuellement l'enrichir avec une transcription des valeurs référentielles dans les nouvelles versions ;
- Garantir la compatibilité ascendante des référentiels afin de limiter la charge de travail induite pour les utilisateurs et d'éviter les potentielles régressions ou pertes de données ;
- Prévoir l'évolution des bases de données et standards pour intégrer systématiquement pour toutes les données référentielles un attribut précisant la version utilisée (en complément de la valeur référentielle).

Proposition de bonne pratique pour la gestion des bases de données des plateformes et producteurs :

→ Intégrer les nouveaux référentiels au fil de l'eau (la compatibilité ascendante de ceux-ci permettant de fait de simplifier cette étape) ;

→ Déprécier les entrées référentielles modifiées ou supprimées pour que les nouvelles saisies ou intégration de données puissent être toujours exprimées dans la dernière version (nécessite la mise en place d'une notion de valeur active/gelée ou non au sein des référentiels stockés) ;

→ Conserver l'occurrence et ses informations dans son état d'origine et proposer la transcription des valeurs référentielles dans les nouvelles versions si possible (prévoir éventuellement la possibilité de bancariser les données d'origine telles qu'elles avaient été diffusées par son producteur) ;

Facteur de facilitation

Sur les 5 référentiels cités ci-dessus, l'UMS PatriNat assure la coordination et la gestion de 4 d'entre eux. Le cadre proposé ci-dessus pourrait donc être discuté et élaboré conjointement au sein des équipes de PatriNat.

Par ailleurs ce travail pour limiter les cas critiques (modification ou suppression de valeurs référentielles) pourrait par la suite être prolongé par un service d'assistance (accompagnement pour assurer les réattributions, voire webservice permettant d'identifier les cas critiques à faire expertiser).

Il semble beaucoup plus difficile d'imaginer contractualiser (au moins à court terme) avec l'INSEE. Toutefois cet établissement délivre aujourd'hui à chaque changement de version la liste des évènements survenus au sein de son référentiel depuis son origine (1943). De fait, si leur référentiel ne répond pas au besoin de compatibilité ascendante, cet historique des évolutions permet de reconstituer les valeurs dépréciées et leurs relations.

Impact sur la validité ou la sensibilité

Si la mise à jour de l'occurrence entraîne une modification de la donnée (par ex. réattribution taxonomique ou changement de commune), le niveau de validité doit être réévalué et son niveau de sensibilité recalculé si besoin.

8. Standards d'échanges

Comment gérer les mises à jour des versions de standard d'échange ?

Impacts

L'évolution des standards d'échanges (occTax, occHab, MTD pour l'essentiel) peut induire une restructuration des données (cf. 2.4 Vocabulaire partagé).

Les restructurations peuvent engendrer deux types d'impacts :

- Evolution du format d'un attribut ou d'une nomenclature :
  - Les informations contenues au sein de l'occurrence ne sont ni corrigées, ni modifiées, elles sont simplement exposées de manière différente (d'une valeur unique vers un tableau, une surface exprimée de m^2 à ha, etc.) ;
  - Le traitement est sans incidence sur la donnée, l'évolution du standard offre une compatibilité ascendante.
- Evolution de la structure de l'information nécessitant un retraitement de celle-ci (dont la cardinalité) :
  - Les informations contenues au sein de l'occurrence sont corrigées, voire modifiées ;
  - Par exemple modification du type de donnée (texte en lien référentiel, cas de la description des protocoles mis en œuvre au sein d'un jeu de données dont l'information passe d'un texte libre à une relation vers une valeur de référentiel) ou répartition d'une information entre plusieurs champs cibles (cas d'un champ observateur transformé en nom et prénom de l'observateur) ;
  - Le traitement potentiellement appliqué à la donnée pour la transcrire dans la nouvelle version du standard transforme sensiblement les informations d'origine.

Le second cas engendre donc une charge de travail conséquente d'analyse mais aussi le besoin d'adaptation de l'ensemble des systèmes applicatifs de la chaîne d'information du SINP. Cette situation est donc à éviter autant que possible.

Proposition de cadre

Le contexte et les principes retenus étant identiques à ceux évoqués dans le point 7 ci-dessus, il semble souhaitable de :

- Garantir la compatibilité ascendante entre les versions (*a minima* sur 2 versions majeures, comme le pratique les éditeurs de langages de programmation ou de bases de données) ;
- Prévoir l'évolution des standards si nécessaire et des bases de données associées pour intégrer systématiquement la version d'expression utilisée pour chaque occurrence ;

9. Transmission multidirectionnelle des mises à jour

Comment informer les acteurs des modifications et répercuter les mises à jour dans l'ensemble du dispositif ou à l'extérieur du dispositif.

Enjeux

Une donnée dont les valeurs sont peut-être corrigées et sûrement enrichies par des statuts de validité et sensibilité au fil de sa transmission ascendante au sein des plateformes du SINP, n'est plus tout à fait la même donnée, au même titre que lorsqu'elle est modifiée par son producteur. L'ensemble des plateformes, et leurs producteurs si possible, doivent donc pouvoir partager entre eux les mises à jour de cette donnée afin de pouvoir tous bénéficier d'un contenu de la donnée identique, afin de garantir sa

pérennité, son unicité et sa conservation dans le temps et assurer aux utilisateurs des différents systèmes de profiter d'une donnée qualifiée et de qualité.

De fait, au-delà de la bonne gestion des mises à jour au sein de chaque plateforme, l'information et la transmission régulière des mises à jour entre les plateformes est également un enjeu important pour assurer la cohérence et la robustesse des informations partagées au sein du SINP : plus la chaîne sera fluide et l'information diffusée rapidement plus le dispositif sera résilient.

Flux identifiés

Les transmissions de données recensées concernent *a priori* tous les échanges du SINP :

- Flux intra-dispositif SINP
  - Producteur accompagné ou autonome => plateforme SINP (régionale ou nationale)
  - Plateforme régionale => plateforme nationale
  - Plateforme nationale => plateforme régionale
    - Partage des statuts de validité nationaux
    - Partage des données diffusées par la plateforme nationale au sein d'une plateforme régionale et inversement.
  - Plateforme régionale ou nationale=> producteur autonome (*a minima*)
    - Partage des statuts de validité nationaux et régionaux
- Flux extra-dispositif SINP
  - GBIF.org
  - Depobio
  - Autres outils/SI métiers du SIE, SIMM, SIB

Comme rappelés précédemment (cf. point 2) les flux descendants ne sont aujourd'hui quasiment pas mis en œuvre (une expérience seulement sur le territoire). Tous les acteurs relèvent l'intérêt de ces transmissions d'informations retour, mais peu arrivent à dégager des moyens pour assurer les traitements de données induits. Par ailleurs le fait qu'il n'y ait pas de méthodologie partagée semble freiner les démarches.

Proposition

Afin de faciliter l'information et fluidifier les transmissions au sein du dispositif SINP, deux axes de travail peuvent être envisagés :

- Communication autour des flux et de leur structuration :
  - Production d'un schéma national des flux entre plateformes régionales et nationale et des méthodes employées ;
  - Production d'un schéma par plateforme portant à connaissance les acteurs, leurs canaux de transmissions et les modes d'intégration des données.

  → Documenter pour mieux maîtriser et faire savoir
- Journalisation des échanges :
  - Conception d'un cadre commun de gestion de la traçabilité des échanges entre acteurs.
  - Voire proposition d'un (web)service partagé de journalisation

  → Informer les acteurs intéressés de l'existence de données mises à jour ou nouvellement apportées sur leur territoire

→ Inciter la diffusion multidirectionnelle

→ Faciliter la mise en œuvre de stratégies d'intégration de données avancées (méthode différentielle).

- Automatisation des flux :
  - Proposition d'une méthode de gestion de flux automatisée
  - Voire mise à disposition d'un webservice global de gestion des flux intra-dispositif SINP

  → Assurer des échanges réguliers (passer progressivement d'un horizon annuel à une vision journalière)

  → Abaisser le coût de transmission en limitant les interventions humaines.

10. Historisation

Les mises à jour et évolutions successives de la donnée doivent-elles être historisées ? Si oui comment et à quel niveau au sein du SINP ?

Une majorité d'acteurs interrogés s'accordent sur l'absence d'intérêt et de besoin à historiser l'ensemble des mises à jour successives que connaît une occurrence tout au long de sa vie :

- L'évolution des valeurs d'une donnée ou métadonnée n'apporte pas d'information en soi : chaque étape devrait être documentée pour apporter un matériau exploitable. Or la fourniture de ces informations explicatives de l'évolution d'une donnée n'est pas la raison d'être ni la mission du SINP, à la différence de l'intérêt qu'une telle fonction peut avoir dans un outil naturaliste producteur de données, ou pour certains programmes de recherche.
- Par ailleurs l'historisation exhaustive demande la mise en œuvre de processus applicatifs pouvant être gourmands en ressources et surtout générer des volumes d'informations conséquents à stocker en bases de données dont la valeur ajoutée est faible.

Toutefois une attention particulière doit être apportée aux occurrences supprimées :

→ S'assurer que les occurrences soient bien supprimées de l'ensemble des plateformes et bases des acteurs

→ Garantir que les identifiants uniques de ces occurrences ne soient pas réintégrés par erreur ou par un tiers suite à un circuit de transmission non maîtrisé.

→ Prévoir la conservation et la réintégration des statuts de validation manuelle apposés par les experts sur des données dont l'identifiant unique ne peut être conservé dans le temps (cas fréquent pour les jeux de données de producteurs accompagnés).

De fait la mise en œuvre d'une historisation des identifiants uniques supprimés serait souhaitée, ainsi que leur diffusion entre les plateformes.

4. Synthèse des propositions

3 scenarios progressifs :

- → Plancher : le minimum viable, impact faible voire nul sur l'existant ;
- → Augmenté : les évolutions possibles, selon les intérêts et moyens des plateformes ;
- → Prospectif : modèle cible potentiel pour répondre aux enjeux à long terme.

Chacun doit pouvoir apporter une réponse maîtrisée aux enjeux identifiés précédemment.

À noter :

Autant le scénario « plancher » décrit les services et niveaux d'échanges minimums attendus entre les acteurs du SINP, autant les scénarios « augmenté » et « prospectif » ne s'entendent pas comme des cibles à atteindre en bloc : les évolutions proposées peuvent être mises en œuvre progressivement, sur une partie des flux seulement, entre quelques producteurs/plateformes uniquement, dans un premier temps sur les échanges de métadonnées (moins complexes *a priori* que les données), etc. Ces 2 scénarios proposent des pistes de travail et d'améliorations dont la mise en œuvre nécessite aussi de s'adapter aux contextes de chacun et dont la pertinence mérite d'être validée graduellement au cas par cas.

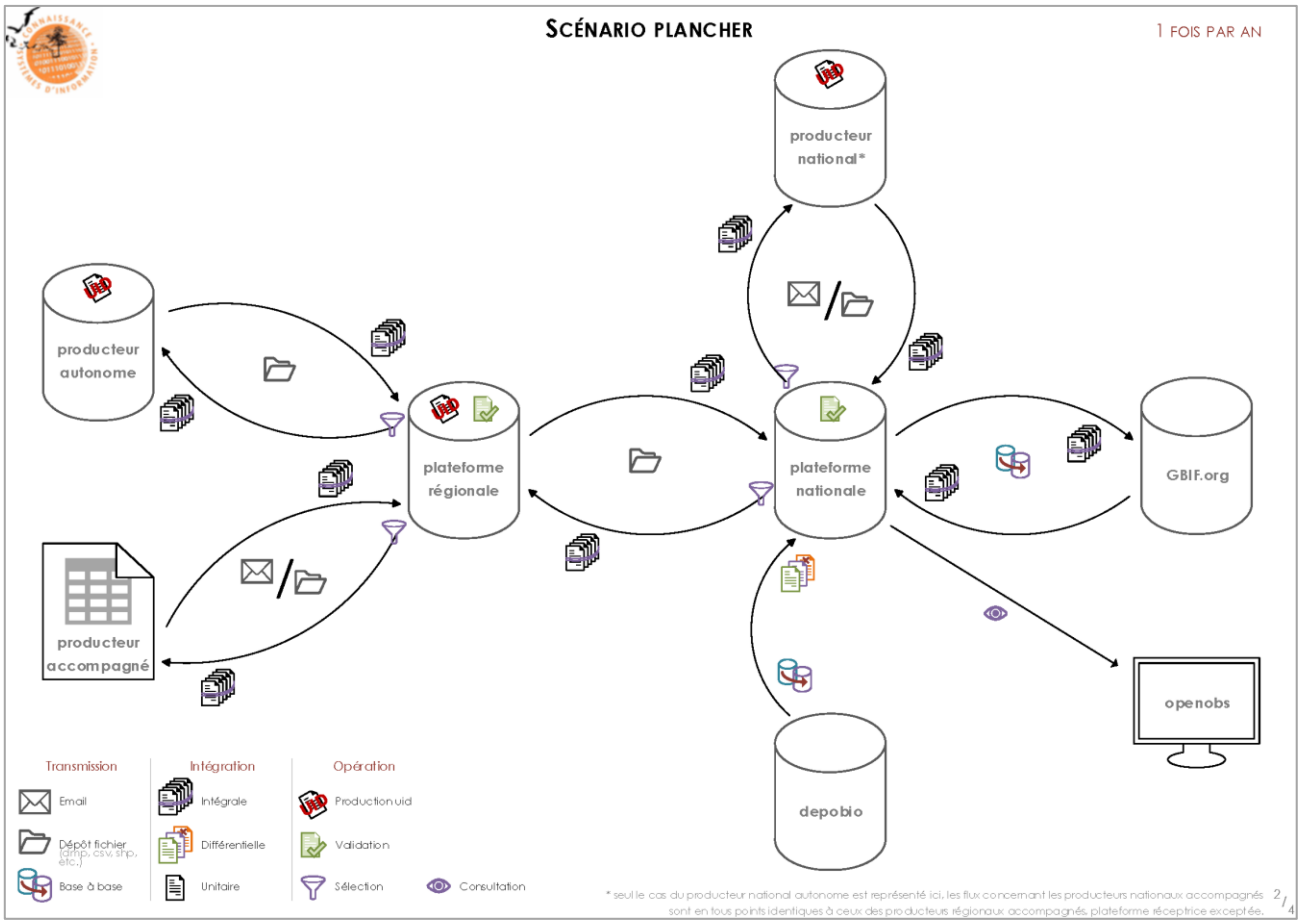
Les visuels des scénarios reproduits ci-après sont consultables en grand format dans l'annexe 1.

1. Scénario plancher

Ce scénario formalise les échanges minimums attendus entre les différents acteurs du SINP pour assurer à la fois l'unicité des occurrences, la propagation multidirectionnelle des mises à jour pour assurer leur pérennité et leur valorisation, la formalisation de la structure et des caractéristiques des échanges entre les acteurs.

Il répond au premier enjeu du SINP : capter et diffuser les données de biodiversité. Aussi il attend de l'ensemble des acteurs qu'ils soient en mesure de partager leurs données au sein des différentes briques de ce système d'information décentralisé, quelle que soit la forme des échanges ou les méthodes employées.

Tous les acteurs du SINP qui participent déjà aux échanges de données au sein du réseau sont en conformité sur les principales attentes de ce scénario. Toutefois les flux descendants sont globalement moins mis en œuvre pour des raisons de moyens et de complexité technique (le scénario augmenté apporte par ailleurs une réponse pour lever ce frein).



Explicitation

Producteur → plateforme

Les producteurs de données collectent et agrègent des observations de terrain au sein de leurs systèmes. Les producteurs autonomes (cf. définitions au 3.1), quel que soit leur périmètre de collecte, apposent un identifiant unique sur chaque occurrence de données ou de métadonnées qui sera conservé et maintenu dans leur base de données.

Chaque producteur prépare un paquet de données et de métadonnées associées à intégrer et/ou mettre à jour au sein d'une plateforme SINP. Ce paquet contient l'ensemble de ses données d'occurrences (et

métadonnées) exprimées selon le format d'échange en vigueur. Selon les accords entre producteurs et plateformes, ce paquet peut être véhiculé dans un (ou plusieurs) fichier(s) (.xlsx, .csv, .txt., .xml, .json, .shp, etc.) ou un dump de base de données.

Les producteurs autonomes, quel que soit leur périmètre de collecte, privilégient une transmission de ces données à la plateforme SINP via un dépôt (serveur partagé FTP ou cloud, voire un guichet, cf. définition au 3.2). Si les producteurs accompagnés n'ont pas la possibilité de suivre cette démarche, ils peuvent utiliser le mail comme canal de transmission. Toutefois dans ce cas, la plateforme devra veiller à assurer une traçabilité de la livraison et accuser réception de l'envoi.

Les plateformes réceptrices, régionales ou nationale, ingèrent ensuite les données et métadonnées déposées par les producteurs selon une méthode intégrale (annule et remplace des occurrences de ce producteur en base de données). Les plateformes assurent également l'apposition d'un identifiant unique sur les occurrences qui n'en contiendraient pas nativement (cas potentiel des données provenant des producteurs accompagnés). Par ailleurs les plateformes assurent la validation des données en assignant à chaque occurrence un statut de validité (régional ou national selon la plateforme).

Plateforme régionale → plateforme nationale

Chaque plateforme régionale prépare un paquet de données et de métadonnées associées à intégrer et/ou mettre à jour au sein de la plateforme nationale, contenant l'ensemble des données d'occurrences (et métadonnées) transmises par ses producteurs adhérents. Ce paquet produit au format d'échange en vigueur peut être véhiculé dans un (ou plusieurs) fichier(s) (.json à privilégier) ou un dump. Il est transmis par le biais d'un dépôt sur un serveur partagé.

La plateforme nationale ingère ensuite ces données et métadonnées transmises par les plateformes régionales selon une méthode intégrale, puis elle assigne à chaque occurrence un statut de validité selon un double procédé (pré-validation automatique, validation manuelle par un cortège d'experts).

Les données et métadonnées de la plateforme nationale SINP sont diffusées sur OpenObs, en respectant les critères de sensibilité des espèces selon les territoires.

Plateforme nationale → GBIF.org

La plateforme nationale SINP prépare un paquet contenant l'ensemble de ses données et métadonnées et les transmet au format Darwin Core directement de base à base à la plateforme internationale du GBIF.

GBIF.org → plateforme nationale

La plateforme nationale SINP extrait de la plateforme internationale du GBIF les occurrences de données et de métadonnées décrites sur le territoire français et déposées par d'autres adhérents qu'elle. Elle intègre et/ou met à jour ces données par méthode intégrale après mise au format SOE des données récupérées en Darwin Core.

Depobio → plateforme nationale

La plateforme de dépôt légal de biodiversité reverse l'ensemble des données gérées à la plateforme nationale SINP. Ce versement s'opère par transferts de base à base de paquet de données selon une méthode différentielle.

Plateforme nationale → plateforme régionale

La plateforme nationale prépare à la demande un paquet complet de ses données et métadonnées déclarées sur le territoire géographique d'une plateforme régionale et n'ayant pas été transmises par la plateforme régionale – d'autres filtres peuvent être apposés au besoin. Ce paquet est transmis dans un lot de fichiers via un dépôt sur un serveur partagé.

La plateforme régionale réceptrice ingère ces données selon une méthode intégrale. Elle assigne ensuite aux nouvelles données transmises un statut de validité régionale. La conservation de l'identifiant unique tout au long de ces échanges lui permet d'éviter les doublons et de conserver les statuts déjà apposés.

Plateforme → producteur

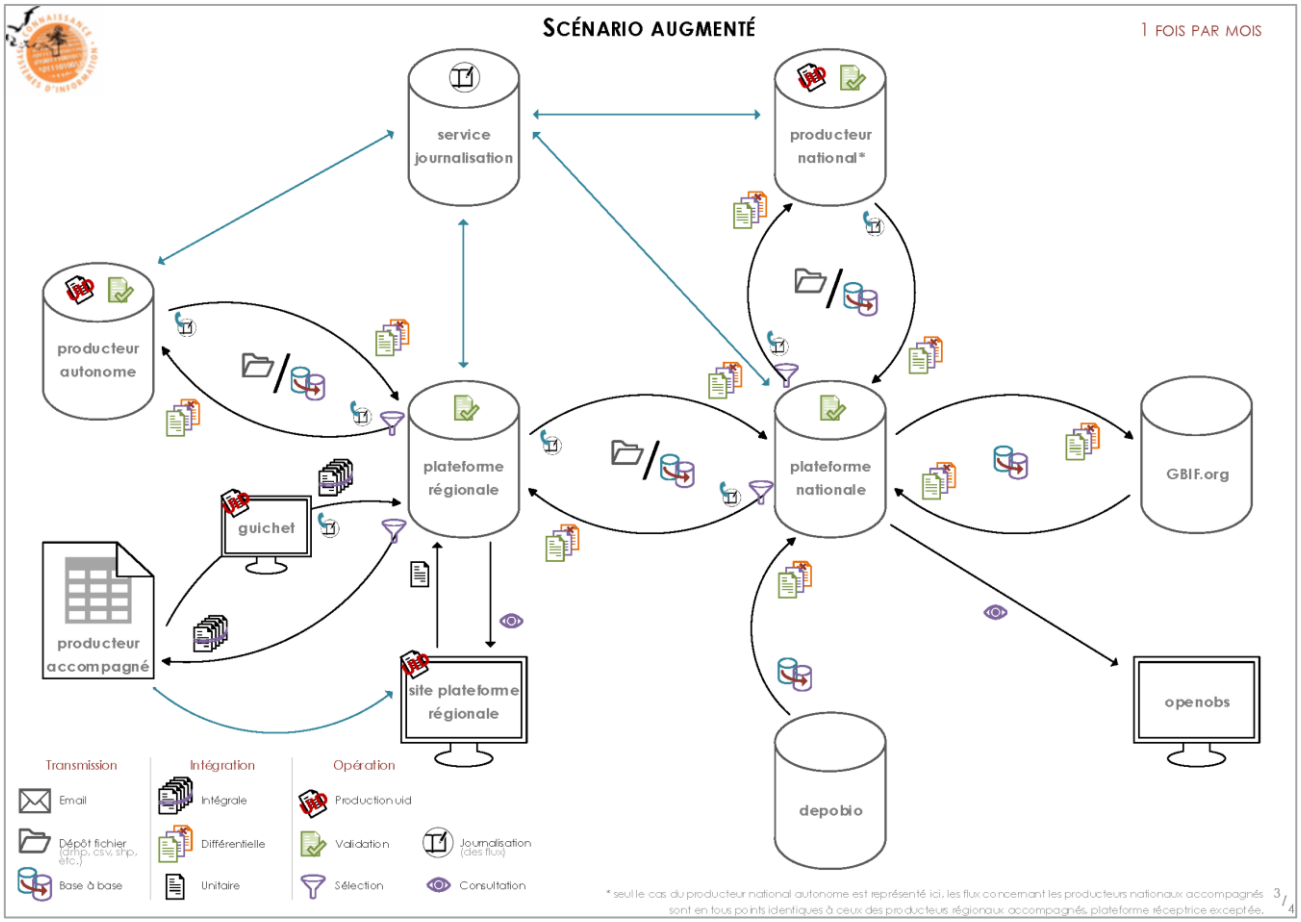
Chaque plateforme (régionale ou nationale) peut à la demande préparer un paquet complet des données et métadonnées qui intéressent un de ses producteurs adhérents. Ce paquet, transmis via un dépôt sur un serveur partagé et filtré selon des critères convenus entre les parties, contient l'identifiant unique de chaque occurrence que le producteur s'attachera à conserver. Les producteurs autonomes s'engagent à les stocker au sein de leurs bases de données lors de l'ingestion selon une méthode intégrale.

2. Scénario augmenté

Cette proposition se concentre sur deux points particuliers à fort enjeu pour le SINP :

- Production de l'identifiant unique au plus près de la collecte d'information :
  - Accompagnement des producteurs autonomes à systématiser son assignation au sein de leurs applicatifs ;
  - Mise à disposition de services pour permettre aux producteurs accompagnés d'apposer cet identifiant sur leurs jeux de données (et donc de le conserver pour leurs futurs usages) dans le même temps qu'ils les déposent pour diffusion sur une plateforme.
- Fluidité des échanges entre les acteurs :
  - Mise en place de protocoles maîtrisés, répétables et organisés entre les parties ;
  - Des échanges plus fréquents, donc une mécanique à la fois plus robuste et plus automatisée ;
  - Mise à disposition d'un service partagé de journalisation des échanges de données – ce service étant essentiel pour constituer les paquets d'occurrences créées, mises à jour ou supprimées au regard des précédentes données et métadonnées échangées, il apparaît opportun d'en construire un seul pour tous les acteurs.

Ce scénario est complémentaire au précédent, les nouvelles briques proposées n'étant pas exclusives des modes de fonctionnement évoqués au point 1. Il formalise par ailleurs des principes déjà mis en place par certains acteurs.



Explicitation

Producteur autonome → plateforme

Tous les producteurs autonomes, quel que soit le périmètre de leur dispositif de collecte, assignent à leurs données et métadonnées un identifiant unique et un statut de validité producteur.

Puis chaque producteur autonome prépare des paquets de données par métadonnée ne contenant que les nouvelles occurrences de données ou métadonnées, celles mises à jour, ainsi que le journal des occurrences supprimées s'il y a lieu, depuis la dernière transmission.

Chaque producteur autonome transmet ces paquets par le biais d'un dépôt sur un serveur partagé, ou si possible via un transfert de base à base. A la fin de la transmission, il déclare son dépôt au service de journalisation du SINP en mentionnant *a minima* la date, la plateforme réceptrice, le contenu et la nature du dépôt (lien vers la localisation du dépôt et/ou lien vers les métadonnées par exemple).

La plateforme réceptrice, prévenue du dépôt par le service de journalisation, ingère les données et métadonnées transmises par méthode différentielle. En fin de traitement, elle confirme son intégration du dépôt au service de journalisation, puis elle peut lancer son processus d'apposition d'un statut de validité sur les données.

Producteur accompagné → plateforme

Tous les producteurs accompagnés, quel que soit le périmètre de leur dispositif de collecte, partagent leurs données avec les plateformes :

- Soit au travers d'un guichet mis à disposition par la plateforme. Un guichet est une interface permettant d'assurer des dépôts de paquets de données tracés et sécurisés, incitant au respect du format d'échange et à la collecte de métadonnées. Le guichet se charge de l'apposition d'un identifiant unique sur l'ensemble des occurrences de données et métadonnées transmises, ainsi que de la déclaration du dépôt au service de journalisation.
- Soit au travers d'un site de collecte de données et métadonnées, mis également à disposition par la plateforme, où le producteur accompagné peut saisir occurrence par occurrence ses informations, les conserver et les consulter. Ce site de collecte assure l'assignation d'un identifiant unique à chaque occurrence.

Service de journalisation

Tous les acteurs du SINP peuvent contacter le service de journalisation ou s'abonner à des notifications pour recevoir des alertes en cas de dépôts de données qui pourraient les intéresser (sur un territoire donné, pour un groupe taxonomique, d'un groupe de producteur particulier, etc.). Ce service permet de fluidifier les échanges en concentrant l'ensemble des informations concernant l'existence de corpus de données et en facilitant l'accès à leur dépôt. Ainsi un acteur intéressé par un jeu de données peut ensuite faire la demande de sa transmission et/ou de sa diffusion au sein du réseau SINP. Par ailleurs la journalisation des flux permet à tous les acteurs de disposer d'un outil de référence pour suivre et piloter ses propres envois et mises à jour.

Plateforme régionale → plateforme nationale

Chaque plateforme régionale prépare des paquets de données par métadonnée ne contenant que les nouvelles occurrences de données ou métadonnées, celles mises à jour, ainsi que le journal des occurrences supprimées s'il y a lieu, depuis la dernière transmission. Pour concevoir ce lot, la plateforme peut s'appuyer sur le service de journalisation des échanges du SINP et retrouver la date de dernière transmission, les lots de cadre d'acquisition ou jeux de données sélectionnés, etc.

Chaque plateforme régionale transmet ces paquets par le biais d'un dépôt sur un serveur partagé, un transfert de base à base sera à privilégier tant que possible. A la fin de la transmission, elle déclare son dépôt au service de journalisation du SINP en mentionnant *a minima* la date, la plateforme réceptrice et le contenu du dépôt (dont le lien vers ses métadonnées).

La plateforme nationale ingère ces données et métadonnées mise à disposition par méthode différentielle.

En fin de traitement elle confirme son intégration du dépôt au service de journalisation, puis elle assigne à chaque occurrence un statut de validité en suivant le double procédé mentionné au point 4.1.

Plateforme nationale → GBIF.org → plateforme nationale

Les flux de données entre la plateforme nationale SINP et la plateforme internationale du GBIF s'opère au travers de paquets de données encapsulés par métadonnée et selon une méthode différentielle.

Plateforme → plateforme ou producteur

À la demande, toute plateforme prépare des paquets de données encapsulés par métadonnées, contenant l'ensemble des occurrences de données et de métadonnées créées, mises à jour ou supprimées depuis la dernière transmission à ce demandeur pour ces mêmes critères. Le transfert de ces données s'effectue via un dépôt sur un serveur partagé, et si possible de base à base. Lorsque le dépôt est finalisé, la plateforme le déclare au service de journalisation des échanges du SINP.

Le demandeur, qu'il soit producteur ou plateforme régionale, ingère ces données selon une méthode différentielle au sein de son système d'information et confirme son intégration au service de journalisation.

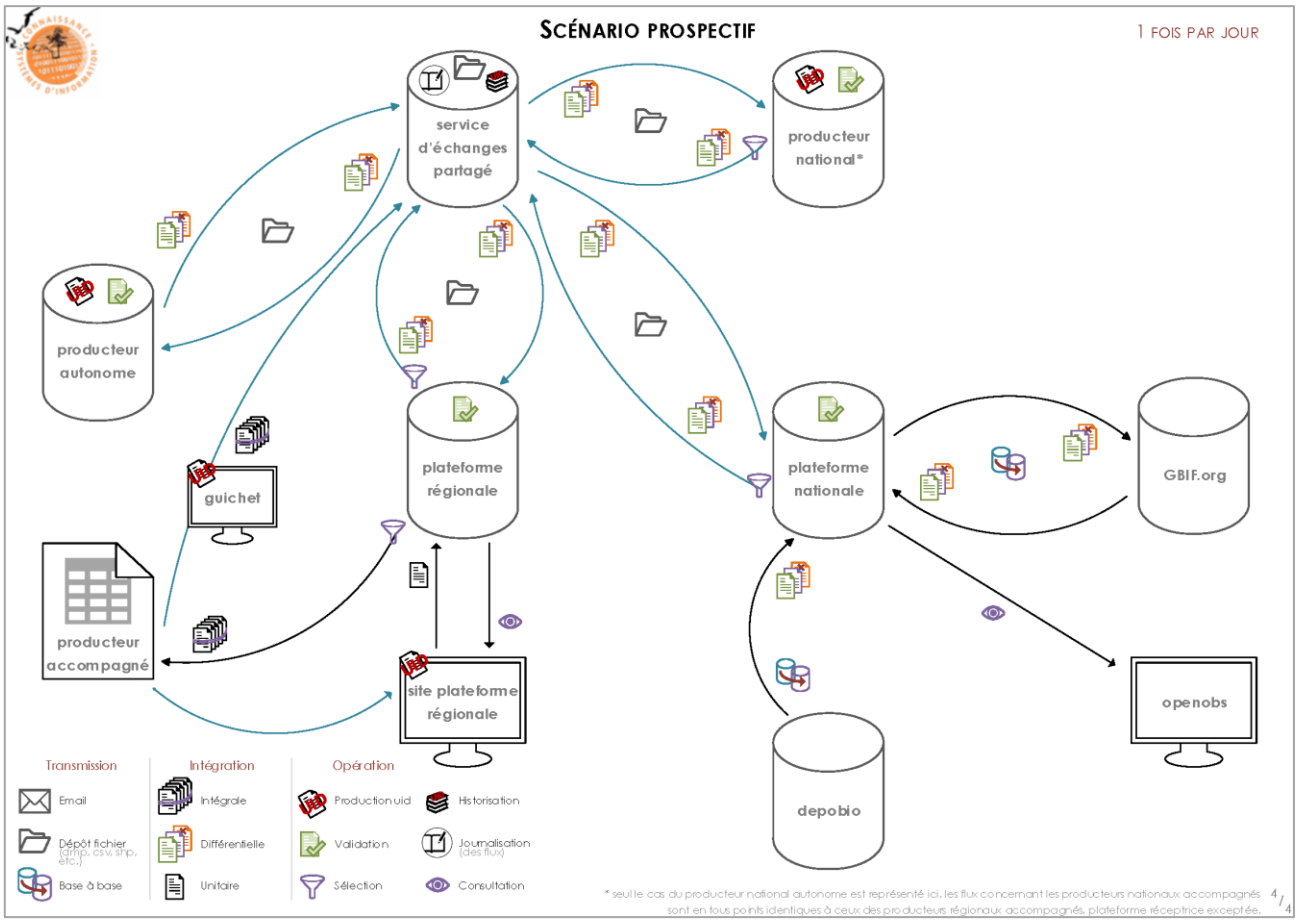
3. Scénario prospectif

Afin de s'abstraire des problématiques de flux générées par l'architecture stratifiée du dispositif SINP, ce scénario explore la possibilité de changer de paradigme : de passer d'une vision ascendante/descendante à une approche centrée autour de l'essence de ces flux, i.e. le partage de données.

En proposant la création d'un service mutualisé de gestion des échanges de données et métadonnées, le modèle présenté encapsule et assume la complexité due à l'organisation décentralisée du SINP et concentre les relations entre les acteurs autour de deux processus :

- Processus de dépôt (côté fournisseur) :
  - Interrogation de la base de journalisation (dernière date d'échange, liste des uuid, version de standards, etc.) ;
  - Conception du paquet de données et métadonnées (selon la méthode différentielle) ;
  - Transmission du paquet au standard en vigueur sur un serveur de dépôt partagé ;
  - Déclaration du nouveau dépôt dans la base de journalisation (date, liste des uuid modifiés, créés, supprimés, version de standard, etc.) ;
  - Information automatique des destinataires par le service de gestion des échanges.
- Processus d'intégration (côté consommateur) :
  - Récupération du paquet de données et métadonnées (dont analyse de cohérence, conformité référentielle, liste des uuid) ;
  - Intégration (par défaut en méthode différentielle) ;
  - Déclaration de l'intégration dans la base de journalisation (dont confirmation liste des uuid modifiés, créés, supprimés, etc.) ;
  - Information automatique des intéressés de cette mise à disposition de contenu (sur la base des territoires, groupes taxonomiques couverts déclarés dans les jeux de données par exemple).

La mise en œuvre d'un tel service mutualisé de gestion des échanges permettrait également d'offrir la possibilité aux acteurs du SINP d'automatiser leurs flux, et ainsi de pouvoir imaginer des échanges hebdomadaires voire journaliers entre les principales plateformes. Cette diffusion rapide des nouvelles informations garantirait à toutes les acteurs de bénéficier quasiment dans le même temps des corrections, modifications, validations, etc. ; ce qui assurerait à la fois l'atomicité des occurrences de données et de métadonnées, donc la robustesse du SINP, et permettrait aussi un accès ouvert et facilité aux dépôts ainsi qu'une meilleure connaissance et maîtrise des flux au sein du dispositif.



Explicitation

Service d'échanges partagé

Le concept de plateforme de service est étendu à l'ensemble du flux, de la journalisation au dépôt en passant par la gestion de l'historisation des identifiants uniques et de leurs statuts, de la notification des nouveaux dépôts voire à l'automatisation de leur transmission.

Tout acteur du SINP qui souhaite transmettre ou récupérer de la donnée dispose de cette plateforme de services dès qu'il répond à deux critères : capacité d'assignation et de conservation d'un identifiant unique sur toutes les occurrences de données et métadonnées, respect du format d'échange.

Les producteurs accompagnés bénéficient d'un service de guichet ou de saisie en ligne comme précédemment, ce sont les plateformes qui mettent à disposition ces services qui gèrent les interactions avec le service d'échanges partagé.

Avant toute intégration de données au sein de leurs systèmes d'information, les acteurs du SINP ont la possibilité de consulter l'historique des identifiants pour s'assurer de l'existence de l'identifiant unique et de son statut (actif, supprimé), de sa date de première création, etc.

Producteur ou plateforme → service d'échange partagé

Le déposant prépare des paquets de données encapsulés par métadonnées, composés des occurrences de données et métadonnées créées, mises à jour ou supprimées depuis sa dernière transmission pour ces cadres d'acquisition ou jeux de données considérés.

Le déposant déclare sur le service d'échange le contenu, le contexte de sa transmission et son destinataire principal. Le service lui délivre une clé pour transférer ses paquets sur le dépôt partagé et sécurisé de la plateforme de service (à noter qu'un échange base à base pourrait aussi être envisagé à terme). A la fin du traitement il reçoit une confirmation, le journal des échanges est automatiquement

enrichi et le destinataire du dépôt reçoit une notification ainsi que les abonnés au service si le dépôt peut les concerner.

Il est envisageable pour le récipiendaire de la notification d'enclencher (potentiellement automatiquement) un processus de récupération de ces nouveaux paquets disponibles.

Service d'échanges partagé → producteur ou plateforme

Tout producteur adhérent et toute plateforme du réseau SINP peut demander au service d'échange la transmission de paquets de données mis à disposition, selon son périmètre de responsabilité. Il mentionne les lots existants qui l'intéressent au service d'échange (ou fait une requête à sa plateforme de rattachement pour production des paquets et mise à disposition sur le service d'échanges partagé). Si sa demande est recevable et validée, il reçoit une clé.

Le demandeur utilise cette clé pour télécharger les paquets attendus depuis le dépôt partagé (à noter qu'un échange base à base pourrait aussi être envisagé à terme) et les ingère selon la méthode différentielle au sein de son système d'information. Puis il confirme leur intégration au service d'échanges partagé.