



Coordination scientifique et technique
Système d'Information sur la Nature et les Paysages
Observatoire National de la Biodiversité



Titre / Objet	Pertinence du standard de données SINP Occurrence de taxon
Rapporteur	Vincent BOULLET
Contributeurs	Bernard CLÉMENT Benoît DAVID Olivier GILG Romain JULLIARD
Avis initial	23 octobre 2014 (V. Boulet)
Contributions	Bernard CLÉMENT (16 novembre 2014) Benoît DAVID (17 novembre 2014) Olivier GILG (4 novembre 2014) Romain JULLIARD (17 novembre 2014)
Avis consolidé	18 novembre 2014
Contribution post CST	Frédéric GOSELIN (15 décembre 2014)
Avis final	11 mars 2015

CONTEXTE

Le contexte de la saisine de la Coordination scientifique et technique (CST) du SINP sur la « pertinence du standard de données SINP Occurrence de taxon » est rappelé dans le mandat de saisine de la CST.

« Les données sur la nature et les paysages sont produites chaque année selon des méthodes et dans des formats différents par de nombreux acteurs sur le territoire français. Cela rend complexe les échanges de données et la nécessaire production de synthèses à différentes échelles. Confirmant cet état de fait, l'audit réalisé en mai 2009 par le Cemagref auprès des DIREN-DREAL et des établissements publics membres du SINP a mis en évidence une attente forte des acteurs en termes de mise en place de méthodes et d'outils pour le recueil et l'échange d'informations.

Afin de répondre à cette demande, le standard d'échange des observations naturalistes définit un format commun à l'ensemble des adhérents du SINP. Il permet de partager les données nature et paysages selon des concepts et des règles d'écriture communes, et de faciliter ainsi la bonne compréhension des données échangées et leurs réutilisations.

Conformément au protocole SINP, les Données Sources des producteurs sont transformées par les plateformes régionales et thématiques (ou par les producteurs de données eux-mêmes) en Données

Elémentaires d'Echange, c'est-à-dire en informations structurées suivant le standard d'échange SINP. Elles sont ainsi mutualisées au niveau de la plateforme nationale.

Ce format s'articule autour d'un cœur d'information présenté dans le standard de données « Occurrence de taxon ». Des extensions seront ensuite développées pour répondre aux besoins spécifiques des thématiques du SINP. »

SAISINE

Les trois questions posées à la CST sont les suivantes :

1. Le standard correspond-t-il bien aux informations minimales à échanger permettant de décrire l'observation naturaliste, ie l'occurrence de taxon ? Ces informations minimales concrétisant un cœur de standard qui sera commun à toutes les extensions de standard permettant d'échanger des informations sur d'autres sujets, comme par exemple, la végétation.

2 Le standard permet-il de véhiculer les informations nécessaires aux programmes listés dans le protocole SINP et concernant le standard « Occurrence de taxon » :

- Production des Atlas de la biodiversité communale (ABC) ;
- Élaboration du portrait de la biodiversité communale (PBC) ;
- Construction d'indicateurs de l'observatoire national de la biodiversité (ONB) ;
- Production des statistiques nationales relatives à la biodiversité (CGDD/SOeS/MNHN) ;
- Production des rapports au titre des directives DHFF, DO, DCSMM, SBUE ou des conventions ou accords internationaux ;
- Contribution au partage international de données GBIF, OBIS, WISE, CDB, etc. ;
- Orientations nationales de la trame verte et bleue, du schéma de création des aires protégées ou des documents d'urbanisme ;
- Production des données de référence de l'inventaire national du patrimoine national (INPN) du MNHN notamment la répartition d'espèces ;
- Espaces protégés (zonage, données réglementaires, espèces présentes) ;
- Inventaires ZNIEFF (zonage, espèces inventoriées) ;
- Natura2000 ;
- Évaluation de l'état de conservation des directives Habitats et Oiseaux
- Productions relatives aux paysages notamment atlas des paysages, carte des unités paysagères, sites classés et inscrits
- Réalisation de l'inventaire du patrimoine géologique.

3 Au-delà de ces usages cités ci-dessus, et considérant que les protocoles sont adaptés, pour quels usages le standard ne permet pas de véhiculer une information cohérente ou suffisante ? Si oui, pouvez-vous préciser le type de manque ou de problème ? (format, champs, qualification...).

Le rapportage de cette saisine s'est appuyé sur les documents mis à disposition de la CST :

- Standard de données SINP « Occurrence de taxon ». Version V1, 2013.
- Résultats du test du standard d'occurrence de taxon. Document de travail, octobre 2013.
- Identifiant permanent de la DEE / Définition opérationnelle dans le cadre du SINP pour la thématique occurrence du taxon. Avril 2014.
- Mandat du groupe de travail « Standard de données du SINP ». Version validée.
- Règles sémantiques d'échange de données entre le SINP et le GBIF / Principes généraux et mappings. Juin 2014.
- Le schéma GML / Gabarit physique du standard de données national SINP Occurrence de taxon.

DÉMARCHE ET REMARQUES INTRODUCTIVES

La démarche suivie du standard de données SINP « Occurrence de taxon » dans sa version « v1 » est, comme l'indique son sommaire, claire et concise : contexte, principe, concepts généraux, attributs, modèle conceptuel de données avec un récapitulatif final des attributs (pp. 25-26) et la déclinaison adaptée du standard de métadonnées du SINP au cas des DEE (chapitre 5).

Pour chaque attribut sont fournis les éléments suivants : nom, définition, caractère obligatoire, cardinalité, format, règle applicable, vocabulaire contrôlé.

Le choix thématique de regroupement des attributs ne permet cependant pas une lecture toujours fluide des attributs ; par exemple, en ce qui concerne la donnée source (DS), l'identifiant et code du jeu de données sont dans un chapitre, l'identifiant d'origine de la DS dans un autre, l'organisme gestionnaire du jeu de données dans un troisième.

Une première remarque générale tient au fait que le standard d'occurrence du taxon ne précise pas (ou tout au moins ne rappelle pas) d'emblée dans son introduction que chaque occurrence, c'est-à-dire chaque DEE (donnée élémentaire d'échange), est « monotaxon » et ne concerne donc qu'un seul taxon. Ce postulat est certes déductif via la cardinalité de certains champs comme le « CdNom » mais non explicite.

Les concepts et le vocabulaire associé ne nous paraissent également pas toujours constants (exemple : « localisation » versus « emprise spatiale ») ou clairs.

Notamment, l'observation (p. 4) est définie comme « *l'identification d'un taxon à une localisation à un moment donné par un ou plusieurs observateurs selon des méthodes d'observation directe (de visu) ou indirectes (empreintes, fèces...).* »

Cette définition est de fait un raccourci qui intègre (condensé qui se retrouve d'ailleurs ici et là de manière sous-jacente dans le standard) :

- le fait même d'observation à caractère unique et non reproductible (ce qui n'exclut pas sa répétition) et s'adressant à des objets concrets (vivants, morts ou produits du vivant) ;
- et le processus d'identification du taxon, évolutif en fonction des compétences du déterminateur et de l'évolution des connaissances taxonomiques, visant au classement des objets observés dans une catégorie taxonomique abstraite.

L'observation porte sur des objets vivants (ou leurs produits : cadavres, fèces, empreintes...), à savoir ici suivant l'échelle choisie du standard, des organismes, généralement un collectif d'individus représentant une population, pouvant à l'extrême se réduire à un unique individu. Observation et identification peuvent être concomitantes (reconnaissance taxonomique instantanée) ou décalée dans le temps (détermination ultérieure sur échantillon collecté, y compris moléculaire, ou autre matière descriptive : description littérale, dessin, photo...).

Le dénombrement d'une population observée, considérée dans le standard comme démarche facultative, renvoie inéluctablement à la question de la notion d'individu qui peut être claire pour la plupart des espèces animales, mais l'est beaucoup moins pour la flore (par exemple pour les espèces clonales ou, plus généralement, à multiplication végétative active) et la fonge. Dénombrer sans identifier clairement l'unité dénombrée n'a guère d'utilité et le standard actuel doit veiller, au-delà des notions théoriques d'individus (génétiques, biologiques, morphologiques), de préciser taxon par taxon l'unité (ou éventuellement les unités) de dénombrement. Le champ « **ObjetDénombrement** » prévoit à ce sujet une déclinaison typologique selon les thématiques systématiques du SINP qu'il serait intéressant de soumettre à la CST.

L'observation peut être positive, c'est-à-dire, qu'elle porte sur les objets présents (par abstraction sur les taxons présents) ou négative sur des objets non présents (par abstraction sur les taxons absents), ce qui sous-entend que soit défini, et de manière utile, le cadre

taxonomique des objets recherchés (principe de détection/ non détection sur base de liste complète d'espèces).

L'observation négative est, mis à part des protocoles ciblés, rarement une fin en soi, mais peut être sous-jacente selon le protocole d'observation à l'observation positive et ainsi déduite (cas, par exemple, des sondages exhaustifs de flore de type « atlas » sur base de bordereau référentiel avec la liste des taxons du territoire inventorié).

Cet aspect est abordé par l'attribut « StatutObservation » qui oppose la « présence » à la « non observation », éléments qui ne sont pas de même nature ; un individu peut être présent mais non observé, inversement, un individu peut être observé mort (par exemple un squelette) et donc non présent lors de l'observation (sauf à considérer que le statut disparu appartient à la présence). La clarification native de ces aspects d'occurrence est d'autant plus nécessaire qu'elle conditionnera, par la suite, les modèles statistiques utilisables pour l'exploitation des données.

Compte tenu de la très grande hétérogénéité de traitement de l'information source dans les systèmes d'information naturalistes, il est important de rappeler ici les grands principes d'intégration des informations attributaires sources de l'observation, selon les quatre situations suivantes :

- **archivage brut** : l'information originelle est retranscrite et stockée à l'identique de la forme source ; ce devrait être normalement le cas du champ « **Nomcite** » (exemples : « *Molinia coerulea* », « Reinette verte » avec l'orthographe originelle non corrigée) ;
- **archivage normé** : l'information originelle est simplement normée mais non modifiée dans sa valeur intrinsèque ; exemple : « douze pieds de jonquille » est normé « 12 » dans un champ « **DenombrementMin** » ;
- **interprétation** : l'information originelle est interprétée selon une analyse contextuelle ou une typologie ad hoc et l'information de stockage résultante peut différer de l'information originelle ; exemples : - le nom cité « BE » dans un bordereau de relevé de végétation de pelouses calcicoles est interprété « *Bromus erectus* » selon la notice du bordereau [analyse contextuelle] ; - « pieds » dans « douze pieds de jonquille » est interprété comme « individu » selon la typologie (vocabulaire contrôlé) de l'attribut « **TypeDenombrement** » ; le rattachement du nom cité à un référentiel nomenclatural est une forme d'interprétation : par exemple « *Bromus erectus* » à « *Bromus erectus* Huds. » présent dans Taxref ;
- **validation** : l'information, après expertise sur le fond, peut s'éloigner plus ou moins de l'information source. Ceci concerne aussi bien les contrôles de cohérence que les redéterminations ultérieures ; par exemples « *Brachypodium pinnatum* (L.) P.Beauv. » validé en « *Brachypodium rupestre* (Host) Roem. & Schult. » ou encore « Vipère aspic » absente du territoire d'où provient la mention, validé en Vipère péliade soit « *Vipera berus* (Linnaeus, 1758) ».

Un des problèmes récurrents est la traçabilité des modifications de l'information attributaire source, pour lesquelles, les bases de données prêtent souvent peu d'attention et engendrent des situations confuses (modifications non contrôlées de l'information source).

Le standard d'occurrence des taxons ne tient pas compte de ces principes qui, dans un certain nombre de cas, mériteraient d'être précisés afin d'améliorer l'homogénéité des jeux de données.

QUESTION 1

« *Le standard correspond-t-il bien aux informations minimales à échanger permettant de décrire l'observation naturaliste, ie l'occurrence de taxon ?* »

Les **informations minimales visées**, c'est-à-dire permettant la description et l'échange des données naturalistes, doivent nécessairement intégrer les éléments suivants :

- l'identité de la donnée, ce qui permet de la repérer et de la mobiliser rapidement dans le jeu de données du SINP ;
- la source de la donnée (répondant à la question « d'où ? ») qui permet, en outre, de revenir à la source documentaire *sensu lato* de la donnée : collection, littérature, autre (dénommée dans le standard « terrain ») ;
- la trajectoire de la donnée dans le processus SINP (et notamment la relation DS ► DEE) (répondant à la question « vers ? ») ;
- l'objet de l'observation (défini ici comme le sujet et traduit en termes catégoriels de taxon) (répondant à la question « quoi ? »), auquel on peut ajouter le « statut de vie » de l'objet observé, que l'on peut concevoir comme une réponse à quelques questions biologiques simples, mais essentielles pour l'exploitation future de l'observation, telles que spontanéité versus cultivé (par exemples, pouvoir faire la différence entre une plantation de sapins et une sapinière naturelle pour la flore vasculaire, entre migrateur versus nicheur pour les oiseaux, etc.) ;
- les circonstances de l'observation (homme / espace / temps) (répondant aux questions « qui ?, où ?, quand ? »), en rappelant que le « où » est à la fois un « où géographique » et un « où stationnel », reconstituant en d'autres termes le noyau conceptuel central de la notion d'habitat (localité + station) ;
- la méthode d'observation (répondant à la question « comment ? »), c'est-à-dire le protocole d'observation qui a généré la donnée ; cet aspect, bien trop souvent négligé dans les méthodes naturalistes et les systèmes d'information, est essentiel pour garantir la pertinence de nombreuses applications attendues pour le SINP.

On peut résumer l'information minimale précédente en deux corpus :

- un **corpus descripteur de l'observation** et répondant aux cinq questions fondamentales : qui ?, quand ?, quoi ?, où ?, comment ?
- un **corpus d'identification de la donnée** permettant à la fois son échange et sa traçabilité (identité, source, trajectoire).

D'un point de vue des concepts (chapitre 3.1), le standard d'« occurrence de taxon » appréhende correctement, dans sa sémantique propre, le corpus d'information minimal descripteur de l'observation : sujet de l'observation (quoi), localisation (où), emprise temporelle (quand), acteurs et rôles (qui), protocoles d'acquisition (comment).

Le corpus d'identification de la donnée ne fait pas l'objet de présentation spécifique et est intégré au « quoi » (chapitre 3.2.1) avec l'intitulé « Attribut de l'observation ». D'une part, cet intitulé ne nous paraît pas approprié, car tous les autres groupes d'attributs du standard (emprise temporelle, emprise spatiale, acteurs et rôles, autres, attributs additionnels) se rapportent aussi à l'observation. D'autre part, l'identification de la donnée ne représente pas le sujet de l'observation. L'organisation de la présentation des attributs pourrait, de ce point de vue, être améliorée.

Des commentaires détaillés des attributs proposés pour les DEE sont donnés en annexe de cet avis. Nous y avons intégré diverses remarques issues du test du standard réalisé sur 13 jeux de données.

Nous n'en présentons à la suite qu'une synthèse en réponse à la question posée.

Description de l'observation

La cible taxonomique du standard, c'est-à-dire le « sujet de l'observation » ou « quoi ? » repose fondamentalement les limites, déjà débattues, entre standards d'« occurrence de taxon » et de « validation de taxon ».

Le standard actuel d'occurrence de taxon acte le principe que la DEE, élaborée dans les plateformes régionales, véhicule à la fois :

- le nom cité « **NomCite** », c'est-à-dire l'archivage brut de l'information taxonomique citée ;

- le « **CdNom** » (selon Taxref), c'est-à-dire l'interprétation nomenclaturale (sous forme de code) du nom cité sur la base de Taxref (rattachement du nom cité à un nom existant dans Taxref) ;
- le « **CdRef** » (selon Taxref), c'est-à-dire la détermination / validation taxonomique au niveau de la plateforme régionale de l'observation. À ce stade et sans autre indication, il s'agit bien d'une information volontaire issue d'un processus de validation taxonomique, à l'exclusion de toute règle automatique de synonymisation (transformation automatique d'un synonyme en son nom valide dans le référentiel usité, c'est-à-dire attribution automatique du CdRef du CdNom).

Cette trilogie attributive (citation / interprétation nomenclaturale / validation taxonomique régionale) répond aux remarques faites antérieurement au niveau de la CST et confirme l'importance, à l'échelon régional, d'une DEE expertisée d'un point de vue taxonomique (sur la base du référentiel commun Taxref) avant son intégration dans le dispositif national.

Ce point mériterait une plus ample explication dans le standard d'occurrence en espérant que notre analyse correspond bien aux objectifs recherchés de la version 1 du standard. Cette étape régionale de détermination/ validation correspond à une phase anticipée du standard de validation de taxon envisagé et il importe de veiller à la cohérence globale de l'ensemble.

On fera également remarquer que ce trio d'attributs « taxonomiques » répondant au quoi ne reflète que partiellement la chaîne d'information complète nécessaire au traitement de tous les cas :

nom cité [information originelle archivée] ► nom reconnu [interprétation contextuelle] ► nom référencé [rattachement nomenclatural au référentiel] ► nom validé [rattachement taxonomique au référentiel].

Exemple : « Bra pin » ► « *Brachypodium pinnatum* » ► CdNom « 86289 » [= *Brachypodium pinnatum* (L.) P.Beauv., 1912] ► CdRef « 86301 » [= *Brachypodium rupestre* (Host) Roem. & Schult., 1817].

Il ne paraît pas nécessaire d'intégrer à la DEE le nom reconnu, néanmoins il convient de préciser ce qu'on entend par « nom cité » afin de véhiculer une information cohérente au niveau national.

Il faut ensuite clarifier le processus détermination/validation, tant dans sa conception que dans sa description dans le standard. Certes, le GT Validation du SINP contribuera à cette clarification (mention p. 21), mais il paraît important que ce point soit amélioré dans le standard actuel d'occurrence du taxon.

Actuellement le standard propose un attribut « déterminateur » et un attribut « valideur ». Il distingue donc clairement les deux rôles (« détermination taxonomique de l'observation » versus « validation formelle de la détermination taxonomique », mais ne propose qu'un seul attribut de date pour ces opérations « **DateDéterminationObs** » qui est précisée comme « date de la dernière détermination du taxon de l'observation ».

Si les deux fonctions « détermination » et « validation » sont reconnues et véhiculées dans le standard elles doivent être cohérentes : à chaque fonction doivent être associés un acteur et une date d'action.

En conclusion de ce point, il faut :

- **préciser les fonctions de détermination et de validation ;**
- **intégrer clairement leur caractère évolutif dans le standard.**

Ces deux points doivent être traités de manière globale avec le GT validation du SINP. Pour notre part, la distinction entre détermination et validation pourrait s'inscrire dans l'orientation suivante :

- la détermination, comme processus engageant les compétences et responsabilités scientifiques individuelles ;

- la validation, comme processus engageant la responsabilité scientifique d'un organisme, notamment lorsqu'il met à disposition des données, telles que les DEE.

Les différentes étapes associées à ces processus seraient alors :

- la première détermination (ou *determinavit* originel) attachée au taxon observé (cette détermination n'est pas obligatoirement concomitante de l'observation même, mais peut être menée ultérieurement) ; dans le cas d'observations instantanées d'"objets volants ou courants...", cette primo-détermination est en fait la seule, car il est impossible d'y revenir (pas d'échantillonnage ou de confirmation ultérieure possible) ;
- les déterminations ultérieures (ou *determinavits* successifs) éventuelles, qui peuvent confirmer ou infirmer les identifications préalables (les *determinavits* successifs sont fréquents en botanique pour les observations associées à des collectes d'herbier) ; on peut associer à cette phase, les résultats des homologations de données (via un comité d'homologation par exemple) ;
- la validation, que l'on peut concevoir comme une certification conforme de la détermination taxonomique (pouvant aussi rappeler, par certains aspects, les processus d'homologation de données mis en œuvre pour certains groupes systématiques).

Déterminations et validations sont des informations évolutives et multivariées qui posent donc la question à la fois de contenu et de mise à jour des DEE.

Le traitement actuel du standard de données paraît à ce sujet quelque peu confus et nécessiterait d'être amélioré.

La règle essentielle de respecter le niveau taxonomique originel (mais pas forcément le rang car celui-ci peut évoluer selon les traitements systématiques retenus) dans la chaîne de détermination/validation est fort justement rappelée dans le standard, sauf, bien sûr, pour les collectes lorsque le matériel collecté permet d'affiner le niveau taxonomique de détermination.

Concernant l'état du « quoi ? », le standard, après avoir testé puis abandonné l'information sur le sexe des individus, n'aborde que les aspects de dénombrement mais omet totalement l'information minimale sur les statuts de vie (spontanéité ou non, migrateur versus nicheur, etc.) ; certes, les attributs additionnels (et donc facultatifs) permettent de gérer au besoin de telles informations, mais leur caractère facultatif entraînera obligatoirement une forte hétérogénéité de l'information et donc des exploitations limitées.

Sur le dénombrement lui-même, la structuration de l'information dans le standard ne permet pas d'intégrer telles quelles des classes de dénombrement, mais invite à considérer les bornes de ces classes comme les nombres maximum et minimum d'individus. Il nous paraît dans ce cas important de compléter la typologie de dénombrement pour intégrer ce mode opératoire.

Enfin, le standard actuel, hors attributs additionnels, ne permet que de traiter les aspects d'abondance (dénombrement), mais pas ceux de dominance ou d'abondance-dominance, comme le coefficient d'abondance-dominance de Braun-Blanquet, pourtant le plus utilisé dans le monde entier. Dans la perspective d'intégration des données « taxons » des données phytosociologiques (notamment des relevés de végétation), cet aspect aurait toute sa place dans le standard de base.

Les questions du « où ? » et du « quand » n'appellent pas de remarques majeures quant aux aspects de localisation et de date des observations. Le standard, sur ces aspects, nous paraît structurer et véhiculer les informations minimales.

L'approche stationnelle est réduite à une référence synthétique via le référencement à une ou plusieurs typologies d'habitats (champs « [CodeHabitat](#) » et « [RefHabitat](#) »), ce qui paraît suffire à l'objectif attendu pour les DEE. On notera cependant à ce niveau, la confusion entre végétation et habitats, informations qui ne sont pas de même nature. Même

si la cardinalité permet de gérer plusieurs codes associés à un référentiel, la séparation en deux corpus d'information « végétation » d'une part, « habitat », d'autre part œuvrerait à la pédagogie et la compréhension de l'information naturaliste.

La complexité de l'information géographique associée aux observations naturalistes sur une longue période historique est abordée dans le standard dans ses dimensions géométriques et territoriales pragmatiques et opérationnelles. La prise en compte obligatoire de la nature stationnelle et inventorielle des objets géographiques associés à la DEE, permet de garantir la pertinence des exploitations spatiales ultérieures des données.

La prise en compte de périodes et d'heures pour les dates autorise une grande souplesse et restitue, sans interpréter, l'éventail utile des informations temporelles des observations naturalistes.

La structuration de l'information concernant les observateurs et les organismes éventuellement associés (le « qui ? ») répond aux objectifs attendus d'information minimale et paraissent suffire aux besoins du SINP. Le standard ne prévoit pas d'information contrôlée pour les noms d'observateurs. Le risque est d'avoir *in fine* une pléthore de doublons de noms d'auteurs mêlant des noms orthographiés différemment (avec des problèmes récurrents d'accentuation), des prénoms entiers ou abrégés, des noms alias, etc. Dans la mesure où le modèle de données ne gère pas cette complexité structurelle des noms de personne, des règles d'écriture devraient être, *a minima*, exigées : respect de l'accentuation, minuscules sauf initiales de noms propres, etc. L'idéal serait à terme de s'appuyer sur un référentiel des noms d'observateurs/déterminateurs/validateurs sous forme de liste ouverte et contrôlée.

La question du « comment ? », c'est-à-dire du protocole d'observation est abordée par un unique attribut, sous forme de code IDCNP. Si le principe de renvoi vers ce référentiel est une façon efficace de gérer cette information dans le standard d'occurrence de taxon, cela suppose la capacité du standard de métadonnées de décrire correctement le protocole d'observation. L'exemple fournit dans le standard d'occurrence de taxon (chapitre 5.2) et la consultation de quelques fiches sur le site du SINP à l'adresse indiquée montrent que le standard de métadonnées a bien vocation à jouer ce rôle mais que sa structuration actuelle, mise en place bien avant la réflexion sur le standard des DEE, ne répond que partiellement au besoin en s'attachant plus au « pourquoi ? » qu'au « comment ? ». Des adaptations et extensions du standard de métadonnées sont en conséquence nécessaires afin de fournir le cadre nécessaire d'information sur le protocole d'observation utilisé, permettant en outre de lier le protocole au programme et rapprocher le « comment ? » du « pourquoi ? ».

Identification de la donnée

Le principe retenu d'identification de la donnée, basé sur un identifiant d'origine de la donnée source, associée à l'indication codée du jeu de données source (attributs « JddId » et « JddCode ») et un identifiant permanent de la DEE, **apporte, de notre point de vue, les garanties nécessaires d'unicité et d'origine de la donnée.**

Les indications de l'organisme gestionnaire (« **OrganismeGestionnaireDonnees** ») et de l'organisme ayant participé à la standardisation (« **OrganismeStandard** ») complètent par ailleurs la traçabilité de la DEE. Dans la mesure où plusieurs organismes standardisateurs interviennent (cardinalité multiple du champ concerné), il serait intéressant de noter leur rôle respectif (nécessité alors d'un attribut supplémentaire).

Si le modèle conceptuel de données du standard actuel nous paraît apporter toutes les garanties d'identité de la DEE et de traçabilité attendues dans un cadre national et international, il paraît malgré tout utile de pouvoir stocker et véhiculer les éventuels identifiants régionaux attribués par les plateformes régionales, d'une part parce qu'ils diffèrent généralement des identifiants d'origine de la base source et que parfois cet identifiant d'origine de la DEE n'existe pas ou n'est pas unique dans la base source.

QUESTION 2

« Le standard permet-il de véhiculer les informations nécessaires aux programmes listés dans le protocole SINP et concernant le standard « Occurrence de taxon » ? »

Une réponse détaillée à cette question nécessiterait une connaissance précise et un travail d'analyse complet des nombreux programmes listés dans la question, ce qui ne nous semble pas réalisable dans le cadre de la CST.

Néanmoins, les réponses à la question 3, en indiquant les limites d'usages du standard d'occurrence de taxon, sont une façon indirecte de répondre à la question 2 en laissant la possibilité d'examiner, au cas par cas, la portée de ces limites d'usages pour chaque programme.

QUESTION 3

« Au-delà de ces usages cités ci-dessus, et considérant que les protocoles sont adaptés, pour quels usages le standard ne permet pas de véhiculer une information cohérente ou suffisante ? Si oui, pouvez-vous préciser le type de manque ou de problème ? (format, champs, qualification...) »

Dans les réponses que l'on peut apporter à cette question, nous ne reprendrons pas les remarques et suggestions d'amélioration de la version actuelle du standard d'occurrence de taxon, mais considérerons les contenus non abordés par la version v1 du standard d'occurrence de taxon.

1. La clé d'entrée du standard est d'ordre taxonomique du vivant ; il n'a donc pas nativement vocation à répondre aux questions du patrimoine géologique, mis à part peut être les éléments associés à la paléontologie (fossiles). Néanmoins, une bonne part du contenu actuel du standard doit pouvoir constituer l'ossature d'un standard d'observation géologique. La place des fossiles, depuis les fossiles récents (comme les coquilles de mollusques associés aux horizons pédologiques) jusqu'aux fossiles des temps géologiques, ou, dans un même ordre d'idée, celle des données de palynologie, d'antracologie, de macrorestes, etc. devront être précisées.

2. Concernant les aspects de démographie et de biologie des populations, le standard actuel se limite à la seule question de la taille des populations observées via le dénombrement des « individus » dans un cadre méthodologique qui doit d'ailleurs être clarifié (voir la possibilité de référencer ou décrire le protocole de dénombrement dans le standard de métadonnées via le code IDCNP) et complété (champ « [ObjetDnombrement](#) » à développer selon les thématiques). L'indication du sexe, proposée initialement dans un champ spécifique a été abandonnée suite aux tests effectués.

D'une part, en ce qui concerne notamment la flore, la taille aréale des populations (surface d'occupation des populations) est à la fois plus pratique dans sa mise en œuvre et plus cohérente dans ses applications dans la mesure où elle évite d'aborder la notion d'individus et le dénombrement toujours difficile de populations à différents stades phénologiques pas toujours maîtrisés... Cet aspect, à la base notamment de l'évaluation des menaces selon les critères de l'UICN, manque totalement dans le standard actuel. L'intégration dans le standard paraît assez facile puisque les 4 champs actuels dédiés au dénombrement pourraient être utilisés, en autorisant leur cardinalité (plusieurs types de « dénombrement » possibles), en utilisant le champ « [ObjetDnombrement](#) » pour choisir une entrée « surface » et en imposant une unité de surface (sinon il faudrait un attribut supplémentaire pour le choix de l'unité de surface).

D'autre part, le standard actuel ne permet pas de sous-dénombrement thématique (par sexe, par stade phénologique, par classe d'âge, etc.). Là encore, la multichiffrement des 4 champs précédents et l'adaptation typologique du champ « [ObjetDenombrement](#) », permettrait sans difficulté d'intégrer ce type de données.

En conclusion, le standard actuel ne véhicule que la seule taille globale des populations en nombre d'individus, mais possède conceptuellement un potentiel d'intégration plus précis de données démographiques.

Le manque principal concerne, comme indiqué précédemment, l'absence d'information sur le statut simple des populations (spontanéité ou non, migrateur versus nicheur, etc.) qui limitera fortement, de notre point de vue, la pertinence et la portée des DEE du SINP. Il nous semble donc important d'intégrer nativement, et non via un ou plusieurs attributs additionnels, cet aspect.

Le cadre du SINP ne nous paraît pas, par ailleurs, réellement approprié pour stocker, échanger et véhiculer d'autres informations scientifiques plus spécialisées sur la biologie des populations. La possibilité d'attributs additionnels le permet de toute manière mais il faudrait que cet effort d'information réponde à une demande cohérente et coordonnée (objectifs, concepts et protocoles partagés) au niveau national.

3. Concernant les aspects écologiques des populations (le « où ? stationnel »), le standard se limite aux seuls aspects d'habitats, avec intrication des notions de végétation et d'habitats qu'il est nécessaire de clarifier.

Ici aussi, le cadre du SINP ne nous paraît pas le mieux approprié pour stocker, échanger et véhiculer des informations scientifiques plus spécialisées sur l'écologie des populations. La possibilité d'attributs additionnels le permet de toute manière mais il faudrait que cet effort d'information réponde à une demande cohérente et coordonnée (objectifs, concepts et protocoles partagés) au niveau national. On pourrait d'ailleurs imaginer des « DEE écologiques » véhiculant les informations écologiques d'un compartiment stationnel donné (ce qui permettrait d'échanger notamment les nombreuses informations recueillies dans des programmes écologiques des domaines de la phyto-écologie, de l'écologie forestière et pastorale, de la pédologie, etc.). Les DEE d'occurrence de taxon pourraient ainsi être liées à des DEE écologiques.

4. Sur la question des menaces pesant sur les populations, thématique importante dans l'évaluation des menaces (démarches listes rouges, directive « Habitats, Faune, Flore », stratégie de conservation, etc.), le standard actuel n'intègre aucune information et ne permet donc aucune exploitation sur cette thématique dans les démarches de liste rouge ou de rapportage Natura 2000.

DISCUSSION COMPLÉMENTAIRE

Deux thèmes sont abordés en complément des commentaires précédents.

En premier lieu, la question des liens entre DEE. Le SINP va véhiculer des données similaires, issues de la même observation initiale, mais qui par la nature diversifiée de leur support (manuscrit de terrain, collecte associée, publication scientifique) vont constituer autant de DEE, puisque de sources différentes. La pratique de l'accumulation de telles données montre fréquemment des déviations (corrections ultérieures, erreurs de transcriptions ou de localisation, etc.) de l'information entre ces différentes données, pouvant entraîner confusions et erreurs d'appréciation, et souvent même amplifier artificiellement l'occurrence d'un taxon... S'il est toujours possible, par comparaison des attributs respectifs de ces données, d'espérer retrouver les liens entre de telles DEE, et de répéter la démarche à chaque besoin, il est plus opérationnel de prévoir techniquement un lien possible entre DEE dans le standard de données lui-même, en précisant (par un champ dédié) la nature du lien.

Un autre type de lien entre DEE qui pourrait être intégré dans ce cadre, concerne les DEE liées par un même protocole (par exemples, données d'un même bordereau de levé d'espèces, ou données d'un transect, etc.). Dans le standard actuel, le rattachement de DEE à une même source ou un même protocole (métadonnée) permet certes d'établir des liens déductifs entre DEE, mais une information claire sur l'existence et la nature des liens entre DEE seraient assurément plus pertinente.

Une seconde interrogation concerne les liens entre DEE « taxon » et DEE d'autres échelles du vivant, comme celles des communautés et complexe de communautés. Le cas des liens taxons/syntaxons concernant les échelles de flore (populations) et végétation (communautés végétales) est particulièrement représentatif et essentiel puisqu'il véhicule, au travers notamment des relevés phytosociologiques de végétation, un nombre considérable (plusieurs millions) de données flore (en outre très précieuses puisque précisément informées sur le plan écologique) constituant autant de DEE Taxon et de données végétation (plusieurs centaines de milliers) à vocation de DEE syntaxon, les premières étant liées aux secondes et vice versa.

L'étroite coordination entre ces deux échelles taxon/syntaxon et la solidarité de leurs traitements dans le SINP afin d'une part, d'éviter les doublons, et, d'autre part, d'assurer la cohérence des déterminations, des validations et de leurs mises à jour nous paraît une absolue nécessité.

CONCLUSION

Le standard de données SINP Occurrence de taxon répond bien globalement bien aux informations minimales à échanger permettant de décrire et de véhiculer l'observation naturaliste. L'articulation et les limites respectives des standards d'occurrence de taxon et de validation des données restent cependant à clarifier et préciser.

Sur le plan de la description de l'observation naturaliste, des améliorations sont principalement attendues sur la thématique des déterminations et des validations en lien avec la remarque précédente. Sur les aspects de bibliographie et de nomenclature des personnes, l'appui sur des référentiels communs et conformes aux normes recommandées nous paraît indispensable.

En ce qui concerne les protocoles d'observation, si le principe retenu par le standard de renvoi de l'information, via un code IDCNP, vers une fiche de métadonnées, paraît pertinent, le standard actuel de métadonnées du SINP ne nous paraît, par contre, pas répondre en l'état à la gestion de cette information.

Sur le plan de l'identification des données, il apporte des garanties d'unicité et de traçabilité dans un cadre national et international, mais pourrait, dans un souci d'opérationnalité régionale, intégrer une identification complémentaire à l'échelle des plateformes régionales.

La possibilité de lier entre elles les DEE issues de la même observation ou appartenant à un même protocole d'observation mériterait d'être intégrée au standard. Celui-ci trouve, par ailleurs, ses limites dans l'absence d'informations plus spécialisées en matière de biologie et d'écologie des populations, mais le cadre du SINP ne paraît pas le meilleur lieu pour échanger et véhiculer ces informations sauf à répondre à un programme national coordonné et de méthodologie parfaitement cadrée. Une autre option serait de créer des « DEE écologiques » permettant l'échange de descriptions écologiques de compartiments stationnels.

Enfin, il est rappelé la nécessité d'une gestion cohérente entre données d'échelle différentes du vivant, et notamment, compte tenu du volume de données concernées, des relevés phytosociologiques de végétation avec leur double entrée taxon / syntaxon.

ANNEXE - COMMENTAIRES DÉTAILLÉS DES ATTRIBUTS PROPOSÉS

Remarque générale - Les cardinalités multiples (0..N, 1..N) ne sont pas gérées de la même manière dans le format. Pour certains champs (« **IdentiteObservateur** », par exemple), l'information multivariée est concaténée dans une entrée unique (avec séparateur prédéfini).

Source

StatutSource - Libellés du vocabulaire contrôlé à harmoniser ; en outre, le terme « terrain » est-il bien approprié ?

ReferenceBiblio - Cet attribut n'est pas normé avec une simple recommandation du format ISO 690. La question d'un référentiel bibliographique commun au SINP est abordée dans le test et évoquée dans un second temps.

Ce champ ne donne que la référence bibliographique en elle-même et non l'emplacement de l'observation dans la source (par exemple, une pagination). Il nous paraît pourtant utile de véhiculer cette information si elle est connue, afin d'éviter de fastidieuses répétitions de recherche pour retrouver le texte correspondant à l'observation.

JddId - Attribut de signification pas très claire et d'interprétation variable (cf. résultats du test). Articulation avec l'attribut suivant « **JddCode** » à revoir. Qui structure et fournit cette information ? La règle « Si non, il peut être complété si l'information est connue » est peu compréhensible.

JddCode - Attribut de signification pas très claire et d'interprétation variable (cf. résultats du test). Articulation avec l'attribut précédent « **JddId** » à revoir. Qui structure et fournit cette information ? Le nom de l'attribut « **JddCode** » n'est pas approprié puisqu'il peut s'agir d'un code, d'un nom, d'un acronyme, d'une initiale... La frontière avec l'attribut « **JddId** » n'est pas évidente. La règle « Si non, il peut être complété si l'information est connue » est peu compréhensible.

Identifiant et traçabilité

Identifiant origine - RAS.

Identifiant permanent - Voir la proposition d'un identifiant de type URL dans le document annexe « Identifiant permanent de la DEE - Définition opérationnelle dans le cadre du SINP pour la thématique occurrence du taxon (CHATAIGNIER *et al.*, avril 2014) ». Cette proposition intègre un UUID (identifiant aléatoire à caractère unique généré par algorithme) pour chaque occurrence de taxon.

La question d'un troisième identifiant, à caractère régional, complémentaire à l'identifiant origine et l'identifiant permanent, est posée dans le document cité précédemment.

DSPublique - Les résultats des tests mentionnent un effort important de standardisation et de documentation de cette information et interpellent l'objectif et la définition de cet attribut. La définition et typologie actuelles n'appellent cependant pas, pour notre part, de remarques particulières. Cet attribut est essentiel à la mise en œuvre du SINP et son absence dans les bases actuelles ne nous paraît pas un argument suffisant pour le remettre en cause.

Code IDCNP - Le principe de renvoi vers le référentiel IDCNP est une façon efficace de gérer l'information du protocole associé dans le standard d'occurrence de taxon ; mais cela suppose la capacité du standard de métadonnées de décrire correctement le protocole d'observation, ce qui n'est pas le cas actuellement.

Sujet de l'observation

StatutObservation - Les libellés de l'attribut ne sont pas logiques : « présent » ne s'oppose pas à « non observé ». Les règles d'intégration des observations négatives doivent être précisées. Voir plus haut la remarque générale sur le sujet.

NomCite - Voir plus haut la discussion sur les attributs taxonomiques du quoi.

CdNom - Les commentaires du test doivent être actualisés avec l'ajout depuis du champ « CdRef ». Les règles d'utilisation des « CdNom » et « CdRef » doivent être clairement explicitées. En théorie, le CdNom qui est un rattachement nomenclatural du nom cité est stable, sauf nouvelle interprétation de la donnée.

CdRef - Les règles d'utilisation des « CdNom » et « CdRef » doivent être clairement explicitées. En théorie, le CdRef correspond à la dernière détermination validée. L'échelon de gestion et de mise à jour, de même que la procédure de mise à jour de cet attribut, doivent être explicités entre les niveaux régional et national. Deux logiques doivent se conforter dans une procédure à caler : la cohérence taxonomique nationale, d'une part, et la réactivité des plateformes régionales aux évolutions taxonomiques, d'autre part.

Sensible - Voir les conclusions finales de la démarche « données sensibles ».

DenombrementMin - RAS ; répond aux demandes d'amélioration issues du test.

DenombrementMax - RAS ; répond aux demandes d'amélioration issues du test.

ObjetDenombrement - Typologie à établir selon les thématiques. Voir la remarque générale plus haut concernant la notion d'individu.

TypeDenombrement - RAS.

Emprise temporelle

DateDebut - Les tirets de la norme ISO8601 ne sont pas utiles ; le format abrégé de cette norme serait plus simple.

DateFin - Les tirets de la norme ISO8601 ne sont pas utiles ; le format abrégé de cette norme serait plus simple.

HeureDebut - RAS.

HeureFin - RAS.

DateDeterminationObs - Voir commentaire de la question 1.

Emprise spatiale

Remarque - La DEE peut avoir plusieurs localisations, donc plusieurs objets géographiques associés (géométrie multiple).

AltitudeMin - RAS.

AltitudeMax - La formulation de la règle « si une seule mesure d'altitude moyenne est mesurée » n'est pas correcte ; en outre il y a confusion entre une valeur unique d'altitude (cas habituel des données ponctuelles) et l'altitude moyenne qui suppose la moyenne d'une

plage de valeurs ; le standard à deux attributs d'altitude mini et maxi permet de gérer une plage d'altitude, une altitude unique (mini = maxi), mais pas une valeur moyenne qui nécessiterait un autre champ.

ProfondeurMin - RAS.

ProfondeurMax - Même remarque que pour l'attribut « AltitudeMax ».

CodeHabitat - Voir remarque de l'attribut suivant.

RefHabitat - On retrouve de nouveau la confusion entre végétation et habitats, informations qui ne sont pas de même nature. Même si la cardinalité permet de gérer plusieurs codes associés à un référentiel, la séparation en deux corpus d'information « végétation » d'une part, « habitat », d'autre part œuvrerait à la pédagogie et la compréhension de l'information naturaliste.

Attribut de l'objet géographique

Geometrie - RAS ; cette information peut cependant poser des problèmes de définition de l'objet géométrique (voir résultats du test).

Precision - Attribut de définition actuellement confuse et de standardisation difficile. Sa pertinence est remise en cause par les résultats des tests. À traiter, soit comme un champ libre de remarque ou à restructurer en plusieurs champs permettant une normalisation de l'information.

NatureObjetGeo - RAS.

Attribut de la maille

La grille nationale 10 x 10 km utilisée ne possède pas de caractère évolutif (hormis un changement de grille) et les données de maille sont figées.

Ce n'est pas le cas des communes dont les limites administratives peuvent évoluer dans le temps. Cette évolution n'est pas prise en compte dans le référencement des communes.

CodeMaille - RAS.

CodeCommune - RAS. Pour les indications de commune non issues de l'exploitation d'un objet géographique, il serait utile de structurer et d'homogénéiser la prise en compte administrative des données « littérales » de communes qui ne concernent pas forcément la commune mentionnée !

NomCommune - Indication de format erronée : il s'agit d'un format « Text » et non « Entier ».

Attribut de l'espace naturel

Même remarque générale que pour les communes concernant le caractère évolutif des limites des espaces naturels et l'absence de prise en compte dans le standard actuel.

CodeEN - RAS.

TypeEN - RAS.

Attribut de la masse d'eau

CodeME - Même remarque que pour les communes et espaces naturels.

Acteurs et rôles

Les subdivisions de ce chapitre (une par attribut) paraissent superflues.

IdentiteObservateur - Si les observateurs d'une observation sont concaténés dans une entrée unique de l'attribut, la cardinalité effective n'est pas « 1..N ». Voir par ailleurs le commentaire de la question 1 et la nécessité, à terme, d'un référentiel de personnes.

OrganismeObservateur - Même remarque que pour « **IdentiteObservateur** ».

Determinateur - Voir commentaire de la question 1.

Validateur - Voir commentaire de la question 1.

OrganismeGestionnaireDonnees - RAS.

OrganismeStandard - RAS.

Autres

Commentaire - RAS.

Attributs additionnels

Les cinq attributs proposés (**NomAttribut**, **ThematiqueAttribut**, **DefinitionAttribut**, **ValeurAttribut**, **UniteAttribut**) constituent une suite logique et cohérente d'information. Il manque malgré tout une information sur le protocole utilisé (protocole de mesure par exemple).