2021

Conservatoire botanique national méditerranéen

Conservatoire botanique national alpin

Conservatoire botanique national de Corse

Processus d’intégration des données d'observations floristiques

SIMETHIS

Système d'information métier des Conservatoires botaniques nationaux alpin, méditerranéen et de Corse

SOMMAIRE

Préambule

Ce document décrit le processus d'intégration et de validation des observations floristiques dans le système d'information métier des Conservatoires botaniques nationaux alpin, méditerranéen et de Corse (SIMETHIS). Cet outil constitue l'outil mutualisé de gestion des observations floristiques et de leur qualification dans le cadre des activités propres aux trois conservatoires ainsi que dans le cadre de l'administration des données floristiques des plateformes SINP régionales SILENE (Provence-Alpes-Côte d'Azur), BiodivOcc (Occitanie - départements de l'ex Languedoc-Roussillon) et Biodiv'AURA (Auvergne-Rhône-Alpes - départements alpins).

1. Définitions

Intégration des données : processus permettant d'insérer les données dans la base réceptrice comprenant la vérification de la conformité et de la cohérence des données, le rattachement des données aux référentiels en vigueur et le contrôle des doublons. Une fois les données importées, elles peuvent passer par une étape de validation scientifique et de qualification et être diffusées dans les différents outils existants et faire l'objet d'analyses.

Validation des données (au sens du SINP) : la validation des données correspond au contrôle de la conformité et de la cohérence et à la validation scientifique des données.

Conformité (au sens du SINP) : la conformité désigne le respect des règles fixées par les standards de données et de métadonnées, autant sur les aspects physiques que conceptuels (renseignement des attributs obligatoires, format, type, utilisation des référentiels et des listes de valeurs). Par exemple, le champ « date » doit obligatoirement être renseigné et doit respecter la norme ISO 8601, comme spécifié dans le standard de données du SINP.

Cohérence (au sens du SINP) : la cohérence désigne le respect de la logique combinatoire des informations transmises (au sein des données, au sein des métadonnées, entre données et métadonnées). Par exemple, la date de début de l'observation doit être inférieure ou égale à la date de fin de l'observation : les champs « date début » et « date fin » sont cohérents entre eux.

Validation scientifique (au sens du SINP) : la validation scientifique est un processus d'expertise visant à renseigner sur la fiabilité de la donnée. On distingue : la validation scientifique dite automatique qui consiste en une validation faisant appel à des résultats d'expertise amont (des référentiels, des bases de connaissance, etc…) et la validation scientifique dite manuelle qui consiste en une validation des informations faisant appel à une expertise aval (avis d'expert suite à l'analyse des informations transmises).

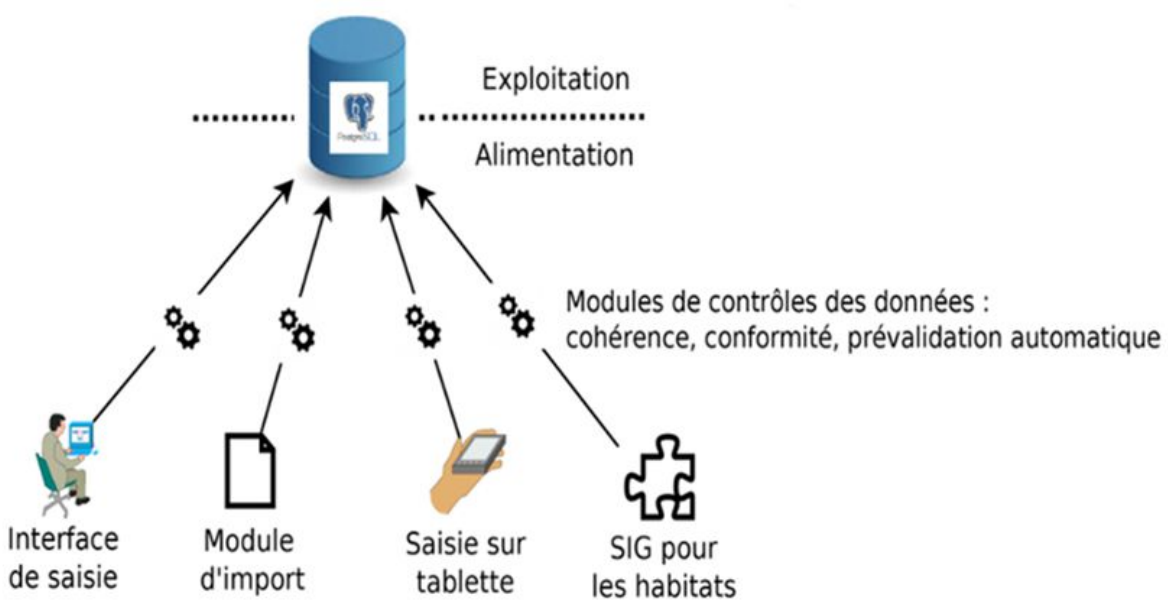
Qualification (au sens du SINP) : la qualification est l'aptitude à répondre aux usages. Elle est le résultat d'une sélection selon des critères inhérents à l'objectif visé, et est donc variable en fonction de cet objectif. Par exemple, une donnée qui ne comporte pas d'information sur les effectifs (ou autre évaluation quantitative de la population) d'une espèce végétale sera jugée utilisable pour réaliser un atlas (c'est l'information de présence de l'espèce qui est importante) mais non utilisable pour étudier la démographie d'une population.

2. Processus d'intégration des observations floristiques

Les données produites par les Conservatoires sont issues soit de la saisie par leur personnel des observations (inédites ou bibliographiques) directement dans les interfaces web de SIMETHIS, soit de leur saisie sur des outils nomades, soit d'import de jeux de données internes.

Les données externes reçues pour intégration proviennent principalement des réseaux de botanistes correspondants des CBN, de partenariats bilatéraux et de l'intégration des jeux de données dans le cadre SINP pour les régions Provence-Alpes-Côte d'Azur, Occitanie (partie ex-Languedoc) et Auvergne-Rhône-Alpes (départements alpins). Ces observations sont livrées principalement sous forme de tableurs ou de données SIG mais également par saisie directe depuis un ordinateur ou depuis des outils nomades. Une faible partie de ces données est livrée sous forme de communications orale ou écrite.

Figure 1 : Schéma général de réception des données



Les données saisies directement au travers des interfaces web de SIMETHIS ou des outils nomades associés sont directement soumises aux contraintes de conformité, cohérence et structure et sont également soumises aux règles de validation (cf. § 3). Le présent document s'attache à décrire plus précisément les étapes de traitement de données produites en externe.

2.1 Réception des données externes

Les fichiers reçus sont stockés localement sur les serveurs dans un dossier réservé aux données à intégrer. Un premier contrôle permet de s'assurer que les informations minimales (date, observateur (et organisme d'appartenance), taxon, localisation) sont bien disponibles et si besoin de revenir vers le fournisseur pour compléments d'informations. Les doutes sur l'organisme fournisseur et l'origine (publique/privée) de la données sont généralement traités à ce stade.

2.2 Contrôle des données

Trois contrôles sont réalisés sur les observations entrantes : conformité et cohérence, doublon et structuration. Ces contrôles font partie du processus de validation des données au sens du SINP. Une fois ces contrôles effectués, les informations sont rattachées aux référentiels en vigueur pour pouvoir être intégrées dans SIMETHIS.

2.2.1 Contrôle de conformité et de cohérence

Il s'agit de vérifier si les informations minimales requises sont présentes (**SINP : conformité**) :

- Quand : Date de l'observation (au minimum l'année)
- Qui : Observateur(s) et organisme(s) d'appartenance s'il existe
- Quoi : Nom scientifique du taxon observé. Les noms vernaculaires doivent être évités car ils peuvent correspondre à plusieurs noms scientifiques.
- Où : Localisation de l'observation (Ex : coordonnées).

L'absence d'une de ces quatre informations rendra l'observation impossible à intégrer.

Remarque : il est possible de fournir l'information sur la méthode de collecte (comment), le type d'habitats, l'effectif,… Ces informations ne sont pas obligatoires.

Dans le cadre de ce contrôle un test de cohérence (**SINP : cohérence**) va être réalisé sur la date et la localisation :

- Pour la date (**cohérence temporelle**), elle doit toujours être antérieure à la date de réception des données.
- Pour la localisation (**cohérence spatiale**), lorsque la commune est fournie, une analyse spatiale permet de vérifier la concordance entre la commune citée et la commune obtenue par analyse spatiale :
  - Si la commune citée est différente de la commune obtenue par analyse spatiale, le fournisseur est contacté afin de définir l'action à réaliser (correction de la commune, correction de la localisation, rejet de l'observation,…).
  - Si la commune citée est identique à la commune obtenue par analyse spatiale la localisation sera alors qualifiée (voir paragraphe 2.3.3.2)

- Pour la localisation (**cohérence spatiale**), lorsque la commune n'est pas fournie, elle est déduite automatiquement à partir des coordonnées géographiques.

2.2.2 Présence de doublons

Ce contrôle permet de déterminer avant intégration si les observations à intégrer sont déjà en base de données. La vérification est basée sur une comparaison de : la date, de(s) observateur(s), du taxon et de la localisation entre les observations à intégrées et celles déjà en base de données.

Les doublons repérés par cette méthode ne seront pas intégrés en base de données ou provoqueront une mise à jours de la données préexistante en accord le producteur.

Remarque : dans le cas où une observation est transmise deux fois : une fois avec une localisation précise et une autre fois avec une localisation floutée, il sera considéré qu'il s'agit de deux observations différentes quand il n'y a pas d'autre information permettant d'affirmer qu'il y a présence de doublons.

2.2.3 Structuration

Lors de la fourniture de données sous forme de tableur ou de couche SIG, le contrôle se fera aussi sur la structuration des données. En effet des données bien structurée seront facilement intégrables et demanderons peu de modifications alors que des données mal structurées demanderons un travail important de normalisation.

Dans le cas de données mal structurées (dates non formatées ; observateurs non individualisés ; etc.), si le travail est trop important, les données ne seront pas intégrées ; il sera demandé au fournisseur de transmettre des données mieux structurées.

Si besoin, les Conservatoires feront un retour au fournisseur de données sur la structuration des données et les points à améliorer.

2.3 Intégration des observations floristiques dans SIMETHIS (SI métier des CBN)

2.3.1 Délai d'intégration

Hors problème majeurs et jeux de données très conséquents, le délai d'intégration est de six mois au maximum entre la réception et l'intégration pour les données fournies en fichier. Pour les observations demandant une saisie manuelle, il n'y a pas de délai garanti.

2.3.2 Rattachement des informations aux référentiels et mise au format

L'intégration en base de données des observations floristiques va permettre de formater, de standardiser et de rattacher les informations aux référentiels (taxonomiques, géographiques, observateurs, source, …). Ainsi, chaque composante de l'observation sera rattachée aux référentiels lui correspondant ou structurée selon un format défini :

- Quoi : si le code taxon (cd_ref) n'est pas fourni, rattachements au référentiel taxonomique grâce à un outil de rattachement automatique au référentiel TAXREF en vigueur en se basant sur la comparaison du nom du taxon avec le nom du référentiel. Une post-expertise par l'expert peut intervenir sur les rattachements les moins probables.
- Où : conversion de la localisation dans l'une des deux projections acceptée dans SIMETHIS (Lambert 93, WGS 84) si la localisation est livrée dans une autre projection.
- Quand : la date de l'observation doit être livrée au format année-mois-jour (aaaa-mm-jj), si ce n'est pas le cas, celle-ci est convertie. La date est gérée dans deux champs correspondant aux dates de début et de fin de l'observation. Ce système de gestion de la date permet de gérer des périodes (du 18 juillet 2020 au 27 août 2020), des dates incomplètes (2019 => 2019-01-01 ; 2019-12-31, mars 2019 => 2019-03-01 ; 2019-03-31) des dates approximatives (bibliographie).
- Qui : rattachement au référentiel observateurs et organismes.
- Comment : rattachement au référentiel des protocoles de collecte si l'information est présente.

Remarques :

Dans certains cas, il ne sera pas possible de rattacher le taxon au référentiel taxonomique. Le rattachement sera alors fait aux valeurs de mise en attente du référentiel (Taxon à vérifier). Ces données en attente pourront être réétudiées par la suite en vue de les rattacher aux référentiels.

En plus de ces cinq informations, l'intégration donne lieu au renseignement de champs supplémentaires : source de la donnée (organisme fournisseur de la donnée), type de donnée (bibliographique, herbier, terrain, …), altitude (obtenue d'après la donnée d'origine et/ou par croisement avec le modèle numérique de terrain), ….

2.3.3 Qualification des données

La qualification des données a pour objet d'apporter une valeur ajoutée à l'observation floristique en lui associant une précision pour certains champs.

Cette qualification facilitera ultérieurement l'utilisation des données en apportant de nouvelles possibilités de tris.

Sur les 4 informations obligatoires (quoi, où, quand, qui) composant l'observation floristique, seules sont qualifiées les données correspondant au « où » (précision), au quoi (validation scientifique) et au « quand » (qualification de la date).

2.3.3.1. Qualification de la date

Le champ booléen « date_avere » permet de qualifier la date :

- Si la date ou période livrée correspond à la période de terrain alors ce champ prend la valeur « vrai »
- Si la date ou période livrée est déduite approximativement à partir d'autre informations alors ce champ prend la valeur « faux »

2.3.3.2. Qualification de la localisation

La localisation géographique peut être issue :

- de données GPS précises de quelques centimètres à quelques mètres selon le type de GPS utilisé
- de pointages sur carte topographique ou photographie aérienne
- de description manuscrite (lieu-dit, commune, massif, …).

Le tableau ci-dessous liste les différentes qualifications choisies pour la localisation :

Description	Libellé qualification
Localisation GPS, objet ponctuel. Objet linéaire ou surfacique pour lequel l'observation est en tous points de celui-ci.	Pointage précis(P)
Observation localisée au lieu-dit ou objet linéaire ou surfacique pour lequel l'observation n'est pas en tous points de l'objet.	Pointage approximatif (T)
Observation localisée au centroïde de la commune	Commune (C)

A la localisation est associée la résolution en mètres qui correspond à la distance de précision de la donnée :

- précision de la mesure du GPS.
- précision du pointage sur carte ou sur photographie aérienne.
- estimation du rayon du cercle dans lequel peut se trouver l'observation (localisation approximative)
- pour les localisations surfaciques ou linéaires, rayon du cercle englobant le polygone ou la ligne.

2.3.4. Identifiant unique

SIMETHIS permet de gérer 3 identifiants uniques :

- l'identifiant unique propre à SIMETHIS.
- l'identifiant unique dans la base de données source fourni par le producteur. Cet identifiant permet de revenir vers le producteur si besoin.
- l'identifiant unique de la donnée au format uuid. Cet identifiant est soit fourni par le producteur si sa base de données permet de le générer, soit créé lors de l'import dans SIMETHIS. Dans ce dernier cas, l'uuid sera alors envoyé au producteur pour import dans sa base ou archivage.

2.3.5. Intégration

Utilisation de l'outil d'"Import par lots" de SIMETHIS pour réaliser l'intégration. Après le choix du fichier et la description des données, le système procède alors automatiquement aux opérations suivantes et s'interrompt en cas de non conformité :

- chargement du fichier sur le serveur SIMETHIS et vérification des données (présence des champs obligatoires, encodage, formatage des données) ;
- analyse de cohérence des données au regard des référentiels ;
- intégration des données, calcul de l'altitude, de la commune ;
- pré-validation scientifique automatique (voir ci-après) ;
- mise à jour de l'historique des intégrations et création d'un fichier bilan de la pré-validation scientifique.

3. Validation scientifique des observations

La validation scientifique est le mécanisme de contrôle des observations floristiques permettant d'attribuer à chaque observation un niveau de validité.

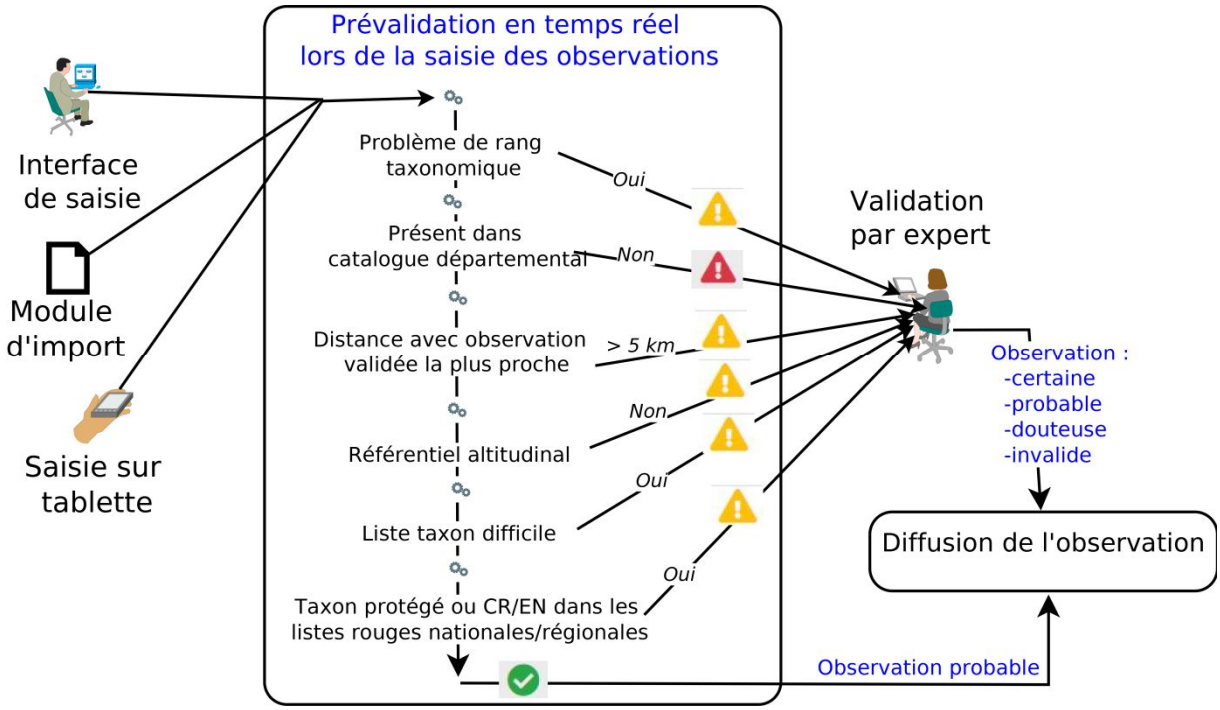
Le tableau ci-après décrit les niveaux de validité de SIMETHIS en lien avec ceux du SINP.

Code	Libellé	SINP
NULL	Non réalisée	Non évalué
0	En cours de validation	Non évalué
1	Certain	Certain - très probable
2	Probable	Probable
3	Douteux	Douteux
4	Invalide	Invalide
5	Non validable	Non réalisable

Remarque : en plus du niveau de validité, sont stockées dans la base de données les informations de : date de validation, type de validation (automatique / manuelle), validateur, commentaire.

La validation scientifique est réalisée en plusieurs étapes d'abord automatisées pour évaluer la probabilité de pertinence de l'observation et si nécessaire manuelles par les experts botanistes référents.

Figure 2 : Schéma général de description des étapes du processus de validation scientifique appliqué par les Conservatoires botaniques nationaux alpin et méditerranéen.



3.1 Validation scientifique automatique

La validation scientifique automatique fait appel à six niveaux d'analyse :

- **Rang taxonomique :** rattachement à un rang spécifique ou infra-spécifique.
- **Catalogue départemental :** comparaison avec le catalogue de référence listant la présence des taxons par département.
- **Cohérence territoriale :** vérification si le taxon a déjà été observé récemment (>= 2000) dans un rayon de 5 kilomètres autour du lieu de l'observation.
- **Cohérence altitudinale :** vérification que l'altitude de l'observation est dans la gamme altitudinale connue pour le taxon (statistique sur les observations valides et précises géographiquement présentes dans SIMETHIS).
- **Détermination :** absence de la liste des taxons complexes à déterminer définie par les experts (listes départementales).
- **Enjeu réglementaire ou de conservation :** taxon non protégé et non fortement menacée (EN, VU dans les listes rouges régionales).

Les listes de taxons pour lesquelles toute nouvelle observation devra systématiquement être contrôlée de manière manuelle par un botaniste référent sont les suivantes :

- **Taxons protégés :** protection nationale, régionale et sa déclinaison départementale.
- **Taxons très menacés :** dont la cotation UICN régionale (si disponible) ou nationale est dans l'une des catégories UICN suivantes :

Code cotation	Libellé cotation
EX	Eteinte au niveau mondial
EW	Eteinte à l'état sauvage
RE	Eteinte au niveau régional
CR*	Non revu récemment, disparition probable mais pas certaine
CR	En danger critique
EN	En danger

- **Taxons de détermination complexe :** listes départementales élaborées à dire d'expert contenant les taxons difficiles à déterminer, mal connus, souvent confondus, etc.

Les observations répondant positivement à l'ensemble des critères sont automatiquement tagguées au niveau de validité "probable" et non spécialement considérée par les experts. Les observations répondant négativement à au moins l'un des critères sont orientées vers la validation manuelle.

3.2 Validation scientifique manuelle par un botaniste référent

Suite à la validation automatique, toutes les observations à contrôler sont analysées par un expert botaniste référent qui statuera du niveau de validité en fonction de :

- ses connaissances ;
- des éléments de contexte (expérience de l'observateur, période d'observation, condition écologiques, etc.) ;

- des données existantes ;
- des éventuels échanges avec l'observateur.

Celui-ci pourra décider d'intervenir sur le rattachement taxonomique pour préciser un rang infra-spécifique par exemple. L'intervention sur le rattachement taxonomique relance le processus de pré-validation automatique. Le nom du taxon d'origine, fourni par le producteur, est systématiquement conservé.

Enfin, il décidera d'attribuer manuellement à l'observation un des niveaux de validité suivant :

- Certain : si une preuve de l'observation (part d'herbier, photographies) a été vue par l'expert.
- Probable : si les éléments en sa possession lui rendent l'observation crédible.
- Douteux : si les éléments en sa possession lui rendent l'observation peu crédible sans toutefois pouvoir le vérifier dans des délais raisonnables).
- Invalide : si les éléments de vérification prouvent une erreur (cas de données anciennes publiées par exemple).
- Non réalisable : en cas d'absence totale d'éléments pouvant aider à une vérification (nom inconnu ou non identifiés dans les référentiels taxonomiques en vigueur).

L'expert dispose des éléments de synthèse des résultats de l'analyse automatique et d'un système d'alertes (vert / orange / rouge) pour doser son effort de contrôle :

- vert : taxons répondant à tous les critères de l'analyse automatique mais protégés ou menacés
- orange : taxons ne répondant pas à au moins un des critères de l'analyse automatique (hors enjeu réglementaire ou de conservation)
- rouge : taxons non connus ou non confirmés (>= 2000) dans le département

Les alertes rouge font l'objet d'un double contrôle (expert + référent validation). Leur validation entraîne l'actualisation automatique du catalogue départemental.

4. Bilan de l'intégration et de la validation scientifique

Dans le cas d'intégration de jeux de données de structures tierces, après intégration et validation (a minima automatique) des observations transmises, un bilan d'intégration peut être transmis sur demande au fournisseur. Ce bilan intègre les informations suivantes :

- Date de réception,
- Date d'intégration,
- Nombre d'observations reçues,
- Nombre d'observations intégrées,
- Nombre d'observations par niveau de validation ou fichier bilan de la validation,
- Problèmes rencontrés et si besoin fichier(s) des observations concernées.

Le fichier bilan de la validation comprend les informations suivantes :

- Id_source : identifiant dans la base de données source
- Id_permanent : identifiant permanent de la donnée au format SINP
- Taxon source : nom du taxon fourni par le producteur
- Cdref_retenu : cd_ref utilisé après validation
- Taxon retenu : nom du taxon après validation
- Modif_rattach_taxon : le rattachement du taxon a-t-il été modifié
- Methode_qualification : qualification manuelle ou automatique
- Resultat qualification : niveau de validité
- date_validation : Date à laquelle a été faite la validation
- nom_validateur : Nom du validateur si qualification manuelle

5. Autres spécificités liées au SINP

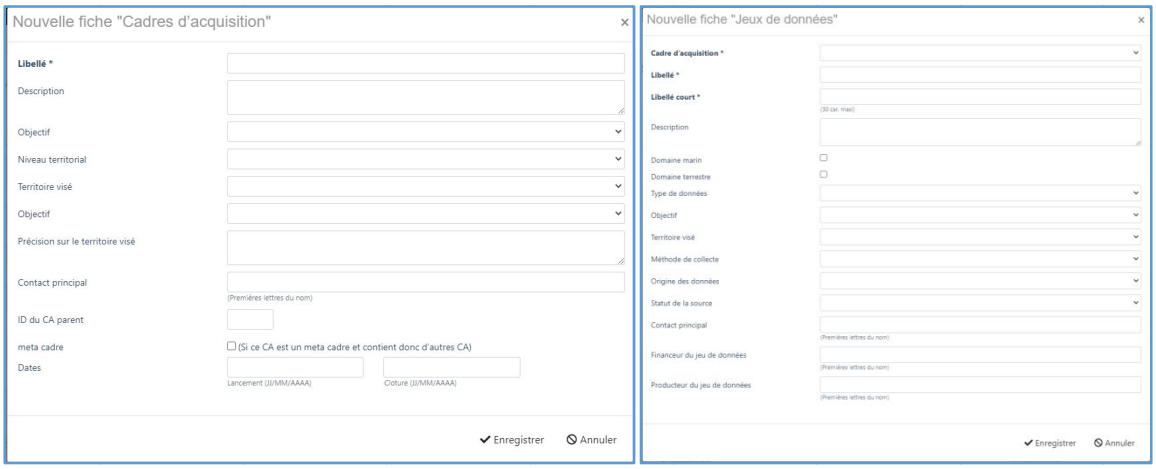
5.1 Génération des uuid

L' identifiant unique universel (UUID) de l'observation est une information obligatoire dans le standard occurrence taxon du SINP. Dans le cas où les données fournies ne contiennent pas d'UUID ceux-ci sont générés lors de l'intégration des observations floristiques dans SIMETHIS pour les observations et les relevés comme évoqué en 2.3.4.

5.2 Génération des métadonnées

SIMETHIS dispose d'un module de gestion des métadonnées permettant de saisir les informations relatives aux jeux de données (JDD) et aux cadres d'acquisition (CA).

Les informations saisies sont compatibles avec le standard SINP Métadonnées et sont les suivantes :



Ces informations doivent être fournies en même temps que les données. Si elles sont incomplètes, les fournisseurs sont contactés pour complément. Dans le cas où le JDD et/ou le CA livrés possèdent des identifiants uniques universels (uuid) ils seront intégrés à la place des UUID générés automatiquement.

5.3 Flux des données de SIMETHIS vers les SINP régionaux

Au sein de SIMETHIS, des fonctions permettant de convertir les données au format des SINP régionaux peuvent être lancées à la demande. Elle permettent le transfert des données soit par la génération de fichiers texte (.csv) soit en écrivant directement dans une base de données par l'intermédiaire de FDW, un système de flux entre base de données.

Dans les 2 cas, la liste des données envoyées est archivée afin de permettre la gestion des insertions, mises à jours, suppressions lors des envoies successifs.

Enfin, une table de SIMETHIS archive le nombre de données envoyé vers les SINP pour un suivi quantitatif dans le temps des données transmises :

bd_cible	date_export	nb_insert	nb_update	nb_delete
SINP OCC	2019-11-19	1404039	0	0
SINP OCC	2020-12-01	1878350	0	1404039
SINP PACA	2021-02-24	4959704	0	0
SINP AURA	2021-04-14	5075435	0	0
SINP OCC	2021-06-16	2052025	0	1878350
SINP OCC	2021-06-18	4593	20066	0

Conservatoire botanique national méditerranéen

34 avenue Gambetta

83400 HYÈRES

04 94 16 61 40

contact.siege@cbnmed.fr

Conservatoire botanique national alpin

Domaine de Charance

05000 GAP

04 92 53 56 82

cbnaatcbn-alpin.fr

Conservatoire botanique national de Corse

Office de l'Environnement de la Corse

14, Avenue Jean Nicoli – 20250 CORTE

04 95 45 04 00

cbnc@oec.fr