Le schéma GML

Gabarit physique du standard de données national SINP Occurrence de taxon

Julie Chataigner (MNHN/SPN), Frédéric Vest (MNHN/SPN), Dimitri Sarafinof (IGN)

1 Introduction 1
2 Structure générale 2
3 Contenu du fichier 3
4 Le Modèle Logique de Données du standard 6
5 Les modèles du Standard 8

1 Introduction

L'objectif de ce document est de présenter l'implémentation du dictionnaire de données en Geography Markup Langage (GML), le format choisi par le GT Standard de données (cf CR du 14 octobre 2013). Les fichiers GML, comme les XML, sont créés à partir d'un schéma de référence en XSD, appelé dans ce document « schéma GML ». Les classes et les attributs sont matérialisés par des balises.

L'utilisation du GML 3.2.1 permet de respecter les normes ISO (norme ISO 19136 publiée en 2007) et Inspire (format préconisé par Inspire).

Ce travail a fait l'objet d'un sous-groupe rassemblant des structures ayant expérimentées la mise en place de fichiers d'échange dans le cadre de système d'information : le Mnhn et l'IGN. Le Sandre (Dimitri Meunier) intervenant sur le Système d'Information sur l'Eau (SIE) a été consulté .

Remarque : en avril 2014, le SIE a décidé de temporiser l'utilisation de format GML pour remplacer le format XML à cause de l'impact que cela a sur les outils existants du SIE, développés depuis de nombreuses années et gérés par différents partenaires. Le SINP n'est pas dans la même configuration car les plateformes sont en cours de création et le standard d'échange est un nouvel élément du SINP.

2 Structure générale

2.1 Découpage des fichiers

Un fichier GML échangera zéro à plusieurs observations représentant une partie ou la totalité d'un jeu de données. Ce découpage pourra être optimisé selon les performances des plateformes dans l'échange de données.
Dans les fichiers d'échange, la balise englobante du jeu de données est « Collection » et la balise de chaque observation est « feature Member » et « Member ». Elles n'apparaissent pas dans le schéma GML (xsd) mais doivent être ajoutées dans les fichiers GML. Cf fichier d'exemple.

2.2 Flexibilité de la structure du GML

La structure du gabarit GML peut être plus ou moins verrouillée :
- dans la présence de toutes les balises, même si elles sont vides. En effet, dans un GML/XML, un champ vide peut se concrétiser par une balise vide ou par une absence de balise. Ainsi, si a, b, c sont des balises d'un GML et que la balise b est vide, car l'attribut est facultatif par exemple, alors la présence de toutes les balises n'est pas obligatoire : la structure du fichier est pour autant toujours conforme.

Fichier avec les balises a, b, c :	**b est facultatif et non renseigné :**	**ou**
<a> 	<a>xxx 	<a>xxx
		<c>zzz</c>
<c> </c>	<c>zzz</c>	

- dans l'ordre des balises. L'ordre des balises n'est pas fixé. Par exemple : les balises a, b, c peuvent se présenter en b, a, c ou c, a, b etc ; pour autant, la structure du fichier est toujours conforme.

Fichier avec les balises a, b, c :	**Autres possibilités :**	
<a> 	<a>xxx	yyy
	yyy	<c>zzz</c>
<c> </c>	<c>zzz</c>	<a>xxx
	yyy	<c>zzz</c>
	<a>xxx	yyy
	<c>zzz</c>	<a>xxx etc

Le choix de verrouiller ces aspects de la structure est impactant pour la validation de la conformité des fichiers et leurs utilisations. En effet, plus la structure est fixée, et plus la validation du fichier

peut se faire avec des parseurs simples et plus la récupération et l'interrogation des données sont facilitées.

Pour le SINP, nous proposons de verrouiller la présence de toutes les balises dans le fichier. Pour cela, nous avons utilisé dans le MLD le stéréotype « voidable », créé par Inspire. Tous les attributs sous le stéréotype Voidable peuvent être Null, ou vides mais la balise doit être présente.
Dans le fichier GML, cela est représenté par les propriétés suivantes des balises : minOccur = 1 maxOccur =1 nillable = true

Nous proposons aussi de fixer l'ordre des balises. Dans le schéma GML, cela est représenté par la balise <xs:sequence>. Remarque : conformément au concept du GML, les balises objet correspondant aux classes sont laissées libres. La classe SujetObservation est flaggée « Root ».

Remarques :
- Dans le cadre du Système d'Information sur l'Eau (SIE), le Sandre a fait le choix de verrouiller la présence et l'ordre de toutes les balises objets pour faciliter leurs traitements.
- Les contrôles de conformité des fichiers au gabarit peuvent être faits par les plateformes.

3 Contenu du fichier

3.1 Balises

Les balises correspondent aux classes et aux attributs du dictionnaire de données.
Des choix d'implémentation ont été faits pour implémenter le Modèle Conceptuel de Données en Modèle Logique de Données, qui lui-même est traduit en Modèle Physique de données (le schéma GML).
Les choix d'implémentation sont présentés au chapitre 4.

Les définitions de chaque élément (classe, attribut, énumération, codeListe) sont ajoutées dans le schéma GML.

Pour l'attribut géographique, le GML rend obligatoire l'échange d'un identifiant unique de l'objet.
Cet identifiant n'est pas définit par le GT Standard de données. Il conviendra aux plateformes R/T de diffuser le leur s'il existe ou d'en générer un s'il n'existe pas.

3.2 Vocabulaire contrôlé

Le vocabulaire contrôlé représente les valeurs de référence à utiliser pour renseigner un champ. Il peut s'agir :
- d'un référentiel géré au niveau national dans le cadre du SINP (TaxRef) ou hors SINP (Commune par l'INSEE). Elles sont considérées comme des *CodeListes* en langage UML.
- d'une liste de valeur interne comme le vocabulaire contrôlé de statutSource ou statutObservation.
    - Ces listes peuvent être extensibles. C'est le cas, par exemple, de la liste « ObjetDenombrement » : des valeurs peuvent être ajoutées par les utilisateurs. Ces énumérations extensibles sont représentées par des *CodeListes* en langage UML.
    - Les listes fermées de valeurs sont des énumérations, représentées par des *Enumeration* en langage UML.

Ces différentes codelistes ou énumérations peuvent faire l'objet de mises à jour plus ou moins importantes. Les référentiels comme TaxRef et Commune sont mis à jour chaque année alors que le vocabulaire contrôlé de statutObservation : « Présent, Non Observé, Ne sait pas » n'a pas vocation à évoluer.

Nous proposons de faire référence aux vocabulaires contrôlés dans le fichier GML, ce qui permet potentiellement de vérifier que la valeur transmise dans le champ est une valeur autorisée par l'énumération ou la CodeListe. Elle permet en outre d'accéder aux valeurs sans avoir recours au dictionnaire de données PDF.

Concrètement, cela peut se traduire de de deux façons :
- Soit la nomenclature est inscrite directement dans le schéma. La nomenclature est directement consultable mais la mise à jour de la nomenclature implique de créer une nouvelle version du schéma GML pour prendre en compte les modifications.
- Soit la nomenclature est appelée via une URL : le fichier fait le lien avec le référentiel disponible dans une forme interrogeable (en XML par exemple).

Avec cette solution, la prise en compte des mises à jour des listes de vocabulaire se fait automatiquement. Cependant, elle est plus lourde à mettre en place et à utiliser :

- la ressource doit être disponible sur la plateforme nationale sous une forme exploitable pour le parseur, ce qui n'est actuellement pas le cas et peut être lourd à mettre en place pour les référentiels complexes et de grande volumétrie comme TAXREF, Commune, EspaceNaturel, Maille et MasseEau.
- il est alors nécessaire d'utiliser un parseur spécifique pour valider la donnée ou de développer du code pour récupérer la liste de valeurs et la comparer au fichier.

Une solution plus légère possible : l'URL peut aussi simplement diriger vers une page informative sur le référentiel à utiliser.

En prenant en compte les particularités des référentiels, et de l'opérationnalité de la mise en œuvre, il est proposé pour cette première version d'intégrer les référentiels à faible volumétrie et faible mise à jour directement dans le gabarit. En attendant que les référentiels plus complexes soient en format interrogeables par URL, les liens ne seront pas intégrés ou dirigeront vers une page informative sur les référentiels à utiliser (à voir avec la plateforme nationale).

Dans un premier temps, les liens URL ne dirigeront pas vers les référentiels sous une forme lisible machine. A terme, il est prévu de rendre les référentiels disponibles dans un format appelable par URL de la plateforme nationale afin de faciliter la vérification de conformité des fichiers. Les nomenclatures plus simples pourront être laissées dans le schéma.

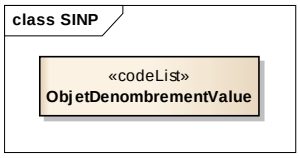
La validation de conformité du fichier peut être effectuée au niveau des plateformes régionales et thématiques et en complément au niveau de la plateforme nationale.

Nom de la nomenclature	**Référentiel Externe au standard**	**Volumétrie** +++ : importante - : peu	**MAJ ou Extensibilité** +++ : très variable - : peu variable	**Intégration dans le schéma GML pour la V1**
HabitatRef	OUI	-	-	**Oui**
StatutObservation	NON	-	-	**Oui**
StatutSource	NON	-	-	**Oui**
ObjetDEnombrement	NON	-	++	**Non**
TypeDenombrement	OUI	-	-	**Oui**
Type EN	OUI	-	-	**Oui**
DSPublique	NON	-	-	**Oui**
NatureObjetGeo	NON	-	-	**Oui**
Code EN	OUI	++	+++	**Non**
TAXREFValue	OUI	+++	+++	**Non**
CodeHabitat	OUI	++	+++	**Non**
CodeCommune	OUI	+++	+++	**Non**

CodeMasseEau	OUI	+++	+++	**Non**
CodeMaille	OUI	+++	-	**Non**
CodeIDCNPDispositif	OUI	+++	+++	**Non**
NomOrganisme	OUI	++	++	**Non**

Figure 1. *Liste et caractéristiques des vocabulaires contrôlés utilisés dans la version 1 du standard occurrence de taxon.*

Cas particulier : Le vocabulaire contrôlé « ObjetDenombrement » a 2 possibilités connues « IN » et « NSP » mais elle peut être extensible donc elle est implémentée en codeList et non pas enumération. Cela pourra être modifié dans une V2, lorsque la liste de valeur sera définie.



Remarques :
- Dans le cadre du Système d'Information sur l'Eau (SIE), le Sandre a fait le choix d'inclure dans les XSD les vocabulaires contrôlés à faible volumétrie, les référentiels plus complexes ne sont pas liés aux XSD.
- Les nomenclatures dans le schéma GML seront de plus, disponibles en CSV pour intégration dans les bases de données.

3.3 Annotation

Une balise de documentation peut être insérée en annotation pour présenter les caractéristiques du schéma. Il est proposé de rajouter une balise d'annotation présentant des informations sur la création et le sujet du schéma.
Ces informations sont en en-tête du fichier de schéma XSD mais elles n'apparaissent pas dans les fichiers de données XML et GML. Ci-dessous la balise proposée :

```
<xs:annotation>
<xs:documentation source = « nom »>Occurrence de Taxon</xs:documentation>
<xs:documentation source = « versionDictionnaire »>1.0</xs:documentation>
<xs:documentation source = «versionSchemaXSD »>1.0</xs:documentation>
<xs:documentation source = «auteurs»>SINP</xs:documentation>
<xs:documentation source = «statutDoc»>En cours validation</xs:documentation>
<xs:documentation source = «description »>Ce schéma permet de d'échanger les données de biodiversité « occurrence de taxon » entre les plateformes régionales/thématiques et la plateforme nationale du Système d'Information Nature et Paysages</xs:documentation>
</xs:annotation>
```

Ci-dessous, pour exemple, la balise d'annotation du schéma des standards d'échange du SIE.

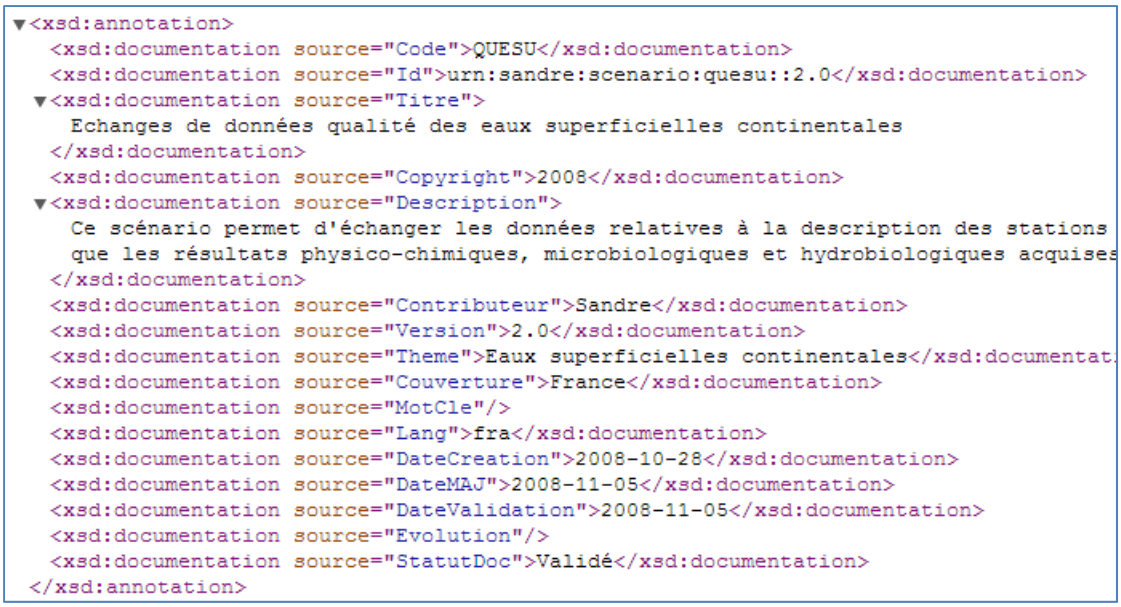


Figure 2. *Extrait du XSD QUESU du SANDRE : illustration de la séquence annotation*

4 Le Modèle Logique de Données du standard (MLD)

Le passage du Modèle Conceptuel de Données (MCD) au MLD requiert des choix d'implémentation. Ils sont présentés ci-dessous.
Le passage du MLD au MPD se fait sans option d'implémentation mais selon les règles des formats techniques, ici selon les règles du standard GML notamment.

4.1 Implémentation des classes

Les classes « Source » et « SujetObservation » sont liées par une association 1-1 : elles pourraient être fusionnées. Cependant, il est choisi d'implémenter les deux classes en balises afin de séparer les attributs de traçabilité issus de la donnée source et caractérisant celle-ci, des attributs de l'observation en elle-même.
A cette fin, l'attribut « statutSource » dans « SujetObservation » est déplacé dans la balise « Source », l'attribut « nomCite » est déplacé de la classe « Source » à la classe « SujetObservation ».

« statutSource » permet de reprendre en un attribut la généralisation de « Source » en Terrain, Littérature, Collection. Le vocabulaire contrôlé de « StatutSource » faisant référence à ces trois types de source : Te, Li, Co. Cet attribut est implémenté dans la balise « Source ». « ReferenceBiblio » y est aussi implémenté.

Les classes « Emprise temporelle » et « SujetObservation » sont liées par une association 1-1 : les 2 classes sont fusionnées : les attributs d' « Emprise temporelle » sont placés dans la balise « SujetObservation ».

Les classes « Localisation » et « SujetObservation » sont liées par une association 1-1 : il est choisi de fusionner « Localisation » dans « SujetObservation ». Cela concerne aussi la classe « ObjetGeographique ».
Les informations du territoire de rattachement sont par contre gérées dans des balises à part (association N-N).
Les classes filles « EspaceNaturel », « Commune », « Maille10x10 », « MasseEau » héritent des attributs et associations de la classe abstraite « Territoire de rattachement ». L'attribut « Code » est ainsi hérité en « codeEN », « codeCommune », « codeMaille », « codeMasseEau ».

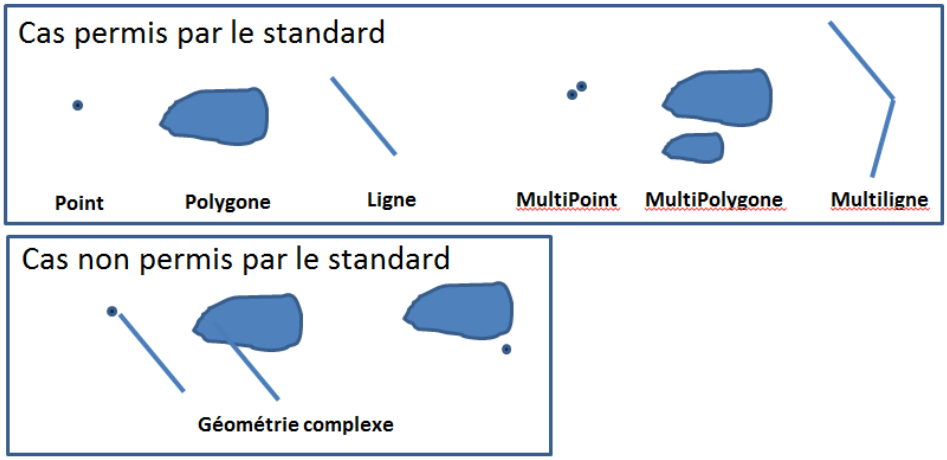
La classe « AttributAdditionnel » est implémentée en une balise englobante « AttributAdditionnel ».

« Validateur » et « Determinateur » sont gérées chacun en une balise dans la balise englobante « SujetObservation », avec concaténation des noms et organismes, conformément au dictionnaire de données. Il a été choisi de traiter différemment les observateurs afin de séparer les individus observateurs des organismes observateurs pour valoriser ces informations plus facilement. « Observateur » est défini en 2 balises « Identite » et « Organisme ». « Observateur » est remis dans « SujetObservation ».

De même les différents rôles de la classe « Organisme » sont intégrés dans les balises englobantes
- « Source » pour « OrganismeGestionnaireDonnees » et
- « SujetObservation » : « OrganismeStandard ».

« Organisme » fait l'objet d'un vocabulaire contrôlé (à terme), ce qui n'est pas le cas des identités des personnes, cf chapitre 3.3.

Remarque : Afin de simplifier le format du fichier GML, la géométrie est mise en GM_object (toute géométrie permise). Cela fait que le schéma permettra de véhiculer des objets complexes, ce qui normalement n'est pas permis par le standard, mais cela simplifie beaucoup le format. Il sera juste nécessaire que ce contrôle soit effectué par les plateformes.



4.2 Gestion des attributs facultatifs

Les champs facultatifs sont notés voidable, cf chap 2.2

Remarques
L'attribut «IDCNPDispositif » est obligatoire dans le dictionnaire de données, cependant, vu les difficultés et le taux d'utilisation du référentiel IDCNP pour les dispositifs de collecte, cet attribut est mis en « voidable » dans cette première version pour ne pas bloquer la mise en œuvre du standard, même si, officiellement cette information doit être et reste obligatoire.

L'attribut « Sensible » n'est pas voidable, conformément au protocole du SINP : Si l'information sur la sensibilité n'est pas connue, alors elle est estimée à « Non » ou « False ». Par défaut, la valeur est donc « Non » (False) et ne requiert pas le traitement en voidable.

4.3 Gestion des attributs obligatoires conditionnels

Afin de gérer au mieux les attributs obligatoires conditionnels dépendant d'autres attributs, les balises englobantes suivantes sont créées :
- DenombrementType
- HabitatType
- ObjetGeographiqueType
- PersonneType pour Observateur.

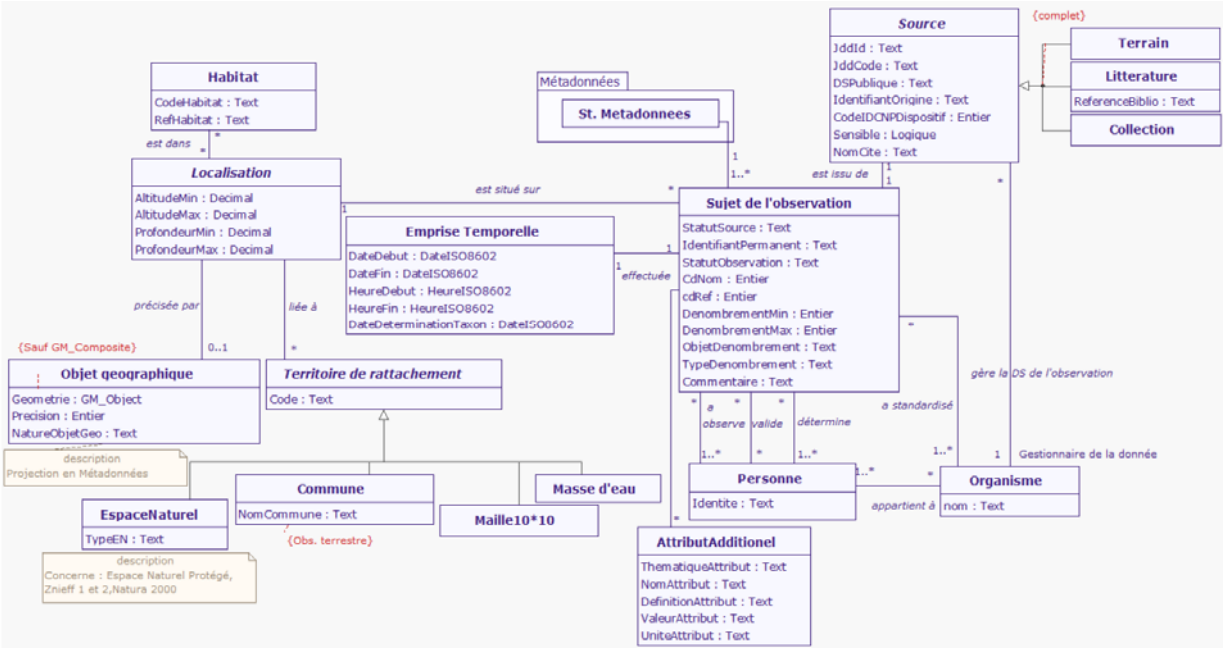
Les autres attributs obligatoires conditionnels sont notés « voidable ». Cependant, cela ne veut pas dire qu'ils sont facultatifs, il faut se référer aux règles pour savoir quand ils sont potentiellement non renseignés. Par exemple, dans le cas où une nouvelle espèce a été observée en France, le taxon n'est alors pas encore référencé dans TAXREF : le cdNom peut être vide pour cette observation. Le schéma permet cette possibilité.

4.4 Fusion d'attribut

Les attributs d'heure : heureDebut et heureFin sont gérés avec les attributs de date d'observation : dateDebut et dateFin, grâce au type DateTimeISO8601 (YYYY-MM-DDThh:mm:ss)

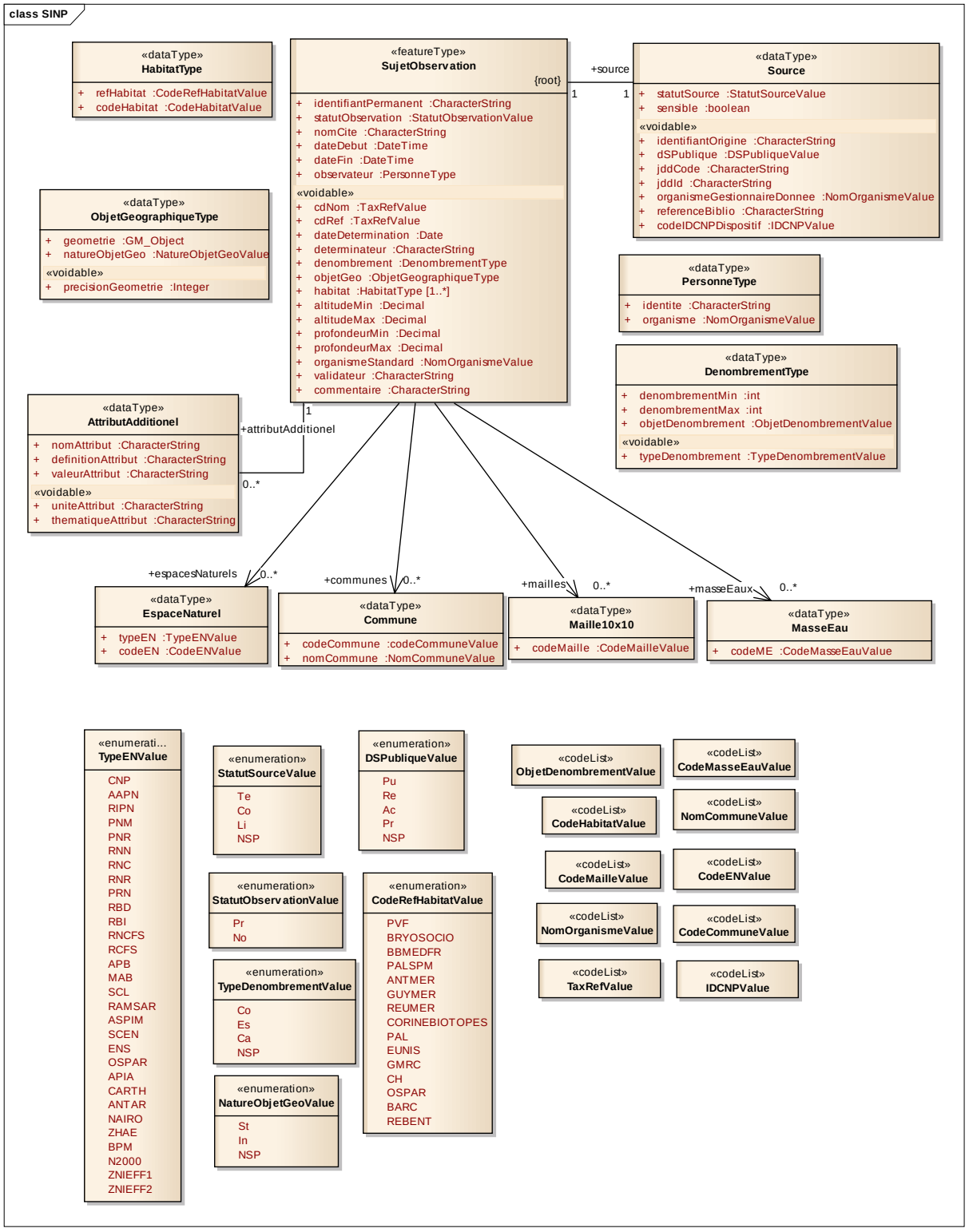
5 Les modèles du Standard

5.1.1 Modèle Conceptuel de Données



5.1.2 Modèle Logique de Données

Le MLD produit à partir d'Enterprise Architect est disponible en format propriétaire EA.



5.1.3 Modèle Physique de Données (MPD)

Le MPD est disponible en schéma GML (OccurrenceTaxon1beta.xsd).