Règles sémantiques d'échange de données entre le SINP et le GBIF

Principes généraux et mappings

Julie Chataigner (MNHN/SPN), Solène Robert (MNHN/SPN), Laurent Poncet (MNHN/SPN), Frédéric Vest (MNHN/SPN)

Echanges avec Sophie Pamerlon (GBIF France)
Relecture au GT Standard de données

Juin 2014

1 Principes généraux

Ce document a pour objectif d'expliciter aux producteurs de données comment leurs données sont diffusées au GBIF via la plate-forme nationale (INPN) et d'informer les utilisateurs des données du SINP sur les données du GBIF qu'ils sont susceptibles de requêter dans le SINP. Les données du GBIF intègrent le SINP par la plateforme thématique « occurrence de taxon ».

Le SINP et le GBIF sont deux systèmes d'information (SI) ayant des objectifs différents.
- Le GBIF est dans une logique de diffusion d'un maximum de données au niveau mondial. De ce fait, la communauté a opté pour un standard peu contraignant : le Darwin Core (Dwc). Dwc a peu d'attributs obligatoires (5) et peu de règles à suivre (recommandations, i.e non obligatoires).
Lien sur le Darwin Core : http://rs.tdwg.org/dwc/terms/
- Le SINP a pour première vocation de permettre la réutilisation des données dans le cadre des exploitations nationales décrites dans le protocole du SINP (directive habitat, indicateurs ONB, atlas de la biodiversité des communes etc). Pour répondre à ces besoins, le standard de données est contraint par des attributs obligatoires (13), et des règles à suivre comme des référentiels à utiliser. Cela permet de favoriser la réutilisation des informations échangées.

Lien sur le standard DEE SINP : http://www.naturefrance.fr/actions/groupe-de-travail-standardisation-des-donnees-biodiversite-du-sinp
Les règles pour échanger les données entre ces deux systèmes d'Information sont présentées ci-dessous et prennent en compte les spécificités de ces systèmes d'information.

Remarque : Le standard Dwc pris en compte est celui de décembre 2013. Il n'y a pas de communication sur les modifications de ce standard d'échange. Celles-ci se font à la volée (pas de fréquence de mise à jour) et ne sont pas documentées (pas de versioning par exemple), il est donc difficile de garder les mappings à jour. Néanmoins pour garder la cohérence entre les deux SI, il peut être décidé de réviser les mappings proposés une fois par an.

1.1 Règles générales dans l'échange des données SINP vers le Dwc

1. Renseigner les champs du DWC donnant une meilleure visibilité de la donnée sur les sites du GBIF :

- Générer des coordonnées X,Y pour chaque observation, même lorsque l'objet n'est pas ponctuel.
En effet, sans ces coordonnées, l'observation n'apparait pas sur les cartes de résultats des sites internet du GBIF. Ainsi, pour que les données apparaissent, la localisation de l'observation sera dégradée au centroïde de l'objet représentant l'observation (lignes, polygones, mailles, communes, espaces naturels).
En complément, les différents attributs ci-dessous seront ajoutés :
- La précision correspondant à la longueur du rectangle d'encombrement de l'objet est diffusée dans le champ « precision »
- L'objet géométrique est diffusé dans le champ « footprintWKT »
- La commune, si elle est renseignée, est diffusée dans le champ « municipality »
- Le champ dataGeneralizations, permettant d'indiquer que plus d'informations sont disponibles chez le fournisseur, est renseigné pour indiquer que des informations complémentaires existent dans le SINP. Une valeur par défaut sera définie. Exemple « Informations complémentaires disponibles dans le format SINP»

- Renseigner le champ « Country ». Une des requêtes classiques du site GBIF est de filtrer les observations par pays. Ce champ renseigné, les observations sont ainsi sélectionnées lorsque l'utilisateur du portail GBIF requête sur le nom du pays.

- Bien que TAXREF (identifiant 0e61f8fe-7d25-4f81-ada7-d970bbb2c6d6) soit déclaré en CheckList (liste de référence), le GBIF France conseille d'échanger les champs de classification dans l'enregistrement transmis. Cette organisation est donc retenue.

2. Diffuser les informations uniquement lorsque le champ DWC correspondant au champ du standard SINP existe, tout en limitant l'effort de standardisation.

- Ne pas utiliser le champ dynamicProperties, permettant d'ajouter des informations dans le Dwc sous la forme : « NomAttribut = ValeurAttribut ; NomAttribut2 = ValeurAttribut2 ». Seules les informations dont les champs sont explicitement prévus dans le Dwc sont diffusées.

- Le champ HigherGeography, permettant de stocker les différents espaces géographiques où est située l'observation, est compliqué à renseigner : il nécessite un requêtage sur plusieurs champs et la gestion de plusieurs cas spécifiques. Le champ résultant dans le Dwc est une valeur concaténée difficilement exploitable. Exemple : HigherGeography = « Amérique du sud ; Argentine ; Patagonie ; Parque Nacional Nahuel Huapi ; Neuquén ; Los Lagos ». L'effort est important pour peu de plus-value : ce champ n'est pas renseigné.

- Pour des raisons similaires : ne pas diffuser « habitat »

- Par contre le champ du dénombrement « IndividualCount » du Dwc ne nécessite qu'une requête simple par filtrage des dénombrements des individus dans le SINP (« ObjetDenombrement » = Individu). Il est donc renseigné quand l'information est disponible dans le SINP.

- Malgré le peu de visibilité dans le Dwc de l'attribut permettant de caractériser si l'occurrence concerne une observation ou une non-observation, il est décidé de transmettre ces deux types d'occurrence et de renseigner systématiquement le champ « OccurrenceStatus ».

- Le champ dataGeneralizations est utilisé avec en valeur par défaut : « Informations complémentaires disponibles dans le format SINP »

3. Mise en format explicite des données véhiculées dans le standard

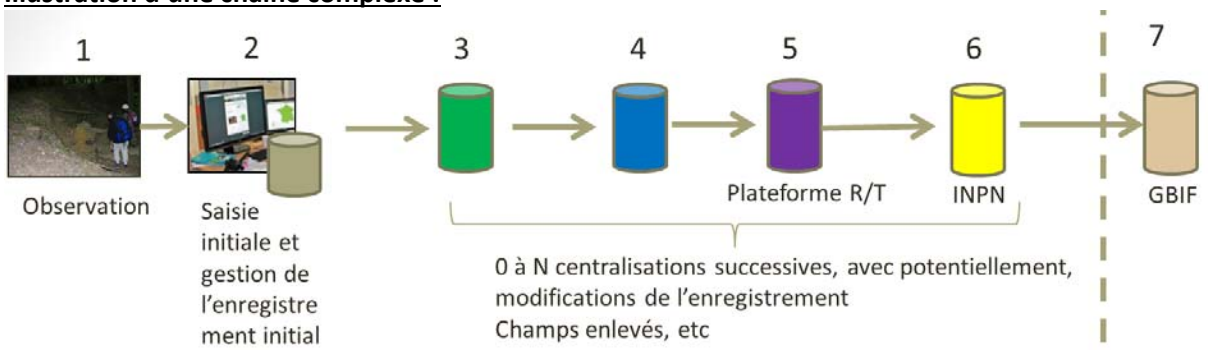
Lorsque des informations sont codées dans le SINP, c'est le libellé qui est transmis au GBIF. Exemple : le mnémonique « Li » est diffusée en « Littérature ».

4. Valoriser les différents acteurs de la production et de la diffusion des données.

Dans le SINP, la chaîne de production et de diffusion de la donnée peut être complexe et faire interagir différents acteurs. L'objectif ici est de valoriser les acteurs intervenant dans la production, la gestion de la donnée source (DS) et dans la standardisation de la DS en DEE.

Les différents rôles du Dwc sont consultables à l'adresse suivante : http://rs.gbif.org/vocabulary/gbif/agent_role.xml. Les définitions sont assez floues et ne prennent pas en compte les chaines de diffusion de données faisant intervenir plusieurs intermédiaires.

Illustration d'une chaîne complexe :



L'INPN étant la plateforme nationale du SINP.

D'où les correspondances suivantes :

Rôle SINP	**Rôle GBIF**	**Acteur dans la chaîne de production** *Correspondant au numéro de l'illustration ci-dessus*
DWC (données)		
OrganismeGestionnaireDonnees	institutionCode "The name (or acronym) in use by the institution having custody of the object(s) or information referred to in the record." *correspond à l'organisme qui gère et à la garde de la donnée*	2 ou 5
	OwnerinstitutionCode "The name (or acronym) in use by the institution having ownership of the object(s) or information referred to in the record." *correspond à l'organisme propriétaire de la donnée*	2 ou 5
EML (métadonnées)		
IdentiteObservateur OrganismeObservateur	ResourceCreator/Originator "an agent who originally gathered/prepared the dataset" *Correspond à l'agent qui a créé le jeu de données source*	2
	Custodian Steward "an agent who is responsible for/takes care of the dataset" *Correspond à un agent qui gère le jeu de données source (ou DEE ?)*	2 ou 5
OrganismeStandard	Processor "an agent responsible for any post-collection processing of the dataset" *Correspond à un agent qui a modifié puis transmis la donnée*	5 Potentiellement 3,4
	Distributor (transmet sans modifier la donnée) "an agent involved in the publishing/distribution chain of a dataset" *Correspond à un agent qui a transmis la donnée sans modifier la donnée*	6 Potentiellement 3,4
	contentProvider "an agent who contributed content to a dataset (the dataset being described may be a composite)" *Correspond à l'agent qui a contribué à créer l'information du jeu de données source*	1
	MetadataProvider "an agent responsible for providing the metadata" *Correspond à l'agent qui a fourni les métadonnées du jeu de données au GBIF.*	6
	Owner "an agent who owns the dataset (may or may not be the custodian)" *Correspond à l'agent qui possède le jeu de données source (ou DEE ?)*	2 ou 5
	pointOfContact	6, 5 ou 2 ?

	"an agent to contact for further information about the dataset" *Correspond à la personne point de contact sur le jeu de données transmis au GBIF*	
	PrincipalInvestigator "a primary scientific contact associated with the dataset" *Correspond à la personne point de contact scientifique du jeu de données*	1

Et pour les bases de données/collection :

jddCode jddId	collectionCode "The name, acronym, coden, or initialism identifying the collection or data set from which the record was derived." collectionID "An identifier for the collection or dataset from which the record was derived." *Correspond au code et au libellé (identifiant) du jeu de données source dans le 5, idéalement de l'IDCNP.*	code et identifiant du jeu de données dans le 5 du 2

Remarques : afin de traduire la complexité de la chaîne de diffusion des données de la biodiversité, le terme « agent » est interprété comme « organisme ayant un rôle » et non pas une « personne ayant un rôle ».
Le tableau propose des correspondances : il faudra définir avec chaque Plateforme Régionale et Thématique quels acteurs sont à attribuer aux rôles et quels rôles sont à échanger avec le GBIF (seul le rôle InstitutionCode est obligatoire).

5. Nommer le jeu de données afin de valoriser le jeu de données initial en premier mais aussi le SINP
Le nom du jeu diffusé est la concaténation du nom du jeu de données source et du nom de la plateforme nationale **« via SINP/INPN »**
Exemple : « Données des campagnes de Suivi Aérien de la Mégafaune Marine (SAMM) de France métropolitaine via SINP/INPN »

1.2 Règles générales dans l'échange des données Dwc vers SINP

Si les informations obligatoires du SINP ne sont pas retrouvées dans l'occurrence Dwc alors l'occurrence du GBIF n'est pas reprise dans le SINP. Exemple : une occurrence sans date.
Remarques :
- Il faudra s'assurer que l'observation n'a pas déjà été déposée en tant que DEE dans le SINP
- Les données disponibles sur le GBIF France et sur le GBIF International sont les mêmes : même jeux de données, même contenu et même métadonnées.

Limiter l'effort de standardisation :
- Se concentrer sur la reprise des attributs obligatoires dans le SINP
- Ne pas reprendre les attributs Dwc ne correspondant pas exactement à un champ du standard SINP.
Exemple : Ne pas essayer de reprendre le dénombrement du Dwc qui est susceptible de véhiculer d'autres informations « 2 male, 2 female » (*définition modifiée en 2014*) et rendant impossible sa récupération automatisée. Autre exemple, ne pas reprendre la donnée de l'habitat : dans le Dwc, il manque l'information sur le référentiel Habitat utilisé et il n'est pas demandé d'en utiliser un.
- Ne pas essayer de reprendre des informations du champ non standardisées « dynamicProperties », qui est un champ de concaténation d'information (cf 1.1 point 2)

- Reprendre l'observation même si le « scientificName » (nom scientifique du taxon observé) n'est pas retrouvé dans TaxRef : il faudra alors le soumettre à l'expertise du MNHN (équipe TAXREF du SPN) afin d'analyser la possibilité de rattacher l'observation au référentiel taxonomique national.

Considérer certaines données par défaut
- L'objet géographique
L'objet peut être la localisation précise de l'observation comme son aire d'étude. On considère par défaut le cas le plus large et indique que le taxon se situe « quelque part » dans l'objet et non « partout » dans l'objet : NatureObjetGeo = « Inventoriel ».
- Utiliser les possibilités « Ne sait pas » ou « inconnu » du standard SINP

2 Mappings

2.1 Mapping SINP → Dwc

Correspondance entre les attributs d'occurrences de taxon de la plateforme nationale du SINP et les attributs du standard DarwinCore du GBIF pour diffusion des données du SINP au GBIF selon le format DarwinCore.

Champ SINP	Champ DarwinCore	Commentaire
StatutSource		Ne pas diffuser car pas de champ Dwc
ReferenceBiblio	associatedReferences	Attribut plus strict dans le SINP : la référence est la source de l'observation.
JddId	CollectionId	Identifiant du jeu de données des DEE du SINP
JddCode	CollectionCode	Identifiant du jeu de données des DEE du SINP
IdentifiantOrigine	CatalogNumber	-
IdentifiantPermanent	OccurrenceID	-
DSPublique		Ne pas diffuser car pas de champ Dwc
CodeIDCNPDispositif		Ne pas diffuser car pas de champ Dwc
OrganismeStandard	Processor (EML)	
StatutObservation	OccurrenceStatus	Echanger les non-observations : Indiquer « non observé » dans ce cas, « Present » si présence
NomCite	ScientificName	-
CdNom	taxonID	-
CdRef		Choix de diffuser le CD_NOM
Sensible		Pas de diffusion des données sensibles donc champ sans objet. Pourra évoluer lors de la V2 du standard de données avec la prise en compte des modalités prévues par le GT donnée sensible.
DenombrementMin	Individual Count	Filtrer dans le SINP sur ObjetDenombrement = Individu – le champ Dwc correspond au dénombrement d'individu minimum par défaut.
DenombrementMax		Ne pas diffuser car pas de champ Dwc.
TypeDenombrement		Ne pas diffuser car pas de champ Dwc.
ObjetDenombrement		Par défaut « **Individual** »
IdentiteObservateur	recordedBy	Règle de concaténation des 2 champs IdentiteObservateur & OrganismeObservateur
OrganismeObservateur	recordedBy	
OrganismeGestionnaire Donnees	InstitutionCode	Cf 1.1 point 4
Determinateur	identifiedBy	
Validateur		Ne pas diffuser car pas de champ Dwc
Commentaire	occurrenceRemarks	
DateDebut	eventDate	Règle de formatage pour gérer suivant la norme ISO

DateFin	eventDate	(date, période, heure dans le même champ)
HeureDebut	eventDate	
HeureFin	eventDate	
DateDeterminationObs	dateIdentified	
AltitudeMin	minimumElevationInMeters	
AltitudeMax	maximumElevationInMeters	
ProfondeurMin	minimumDepthInMeters	
ProfondeurMax	maximumDepthInMeters	
Habitat		Ne pas diffuser car règles complexes
RefHabitat		Ne pas diffuser car pas de champ Dwc
Geometrie	footprintWKT	Pas de champ projection nécessaire, le WKT gère cette information.
Precision	coordinateUncertaintyInMeters	Si la localisation est un point : le champ Precision alimente le champ dwc. Pour toutes autres localisations : ce champ est alimenté par la longueur du rectangle d'encombrement de l'objet géométrique ou du territoire de rattachement
NatureObjetGeo		Ne pas diffuser car pas de champ Dwc
CodeCommune		Ne pas diffuser car pas de champ Dwc
NomCommune	municipality	Concaténer dans le champ si plusieurs communes, séparer par une virgule
CodeEN		Ne pas diffuser car pas de champ Dwc
TypeEN		Ne pas diffuser car pas de champ Dwc
CodeMaille		Ne pas diffuser car pas de champ Dwc
CodeME		Ne pas diffuser car pas de champ Dwc
Attributs additionnels		Ne pas diffuser : effort de standardisation trop important
ThematiqueAttribut		Ne pas diffuser : effort de standardisation trop important
NomAttribut		Ne pas diffuser : effort de standardisation trop important
DefinitionAttribut		Ne pas diffuser : effort de standardisation trop important
ValeurAttribut		Ne pas diffuser : effort de standardisation trop important
UniteAttribut		Ne pas diffuser : effort de standardisation trop important
Autre champ obligatoire dans le Dwc		
	basisOfRecord	« terrain », « bibliographie » = « occurrence », « collection » : « preservedSpecimen »
Autres Champs		
	NameAccordingTo	« TAXREF »
	country	« France » cf Tableau en Annexe
	countryCode	https://www.iso.org/obp/ui/#search FR pour l'ensemble des territoires français. Cf Tableau en Annexe
	kingdom	A minima Kingdom et Family – Une fois le lien avec TAXREF fait, l'ajout d'un ou plusieurs rang de la classification n'ajoute pas de difficulté, d'où l'ajout des rangs taxonomiques ci-contre
	phylum	
	class	
	order	
	family	
	genus	
	X	Coordonnées du centroïde de l'objet permettant de localiser l'observation
	Y	Coordonnées du centroïde de l'objet permettant de localiser l'observation
	dataGeneralizations	Valeur par défaut à définir indiquant que des données complémentaires sont disponibles dans le SINP : « Informations complémentaires disponibles dans le format SINP »

2.2 Mapping Dwc → SINP

Correspondance entre les attributs d'occurrences de taxon du GBIF au format DarwinCore et les attributs de standard SINP pour intégration des données GBIF dans la plateforme thématique « occurrence de taxon » du SINP.

Champ DarwinCore	Champ SINP	Commentaire
	StatutSource	L'information n'est pas connue : **cela pose problème**. Noter « NSP »
associatedReferences	ReferenceBiblio	La définition d'associatedReferences est sujette à interprétation. Elle n'est pas forcément la référence d'où l'occurrence de taxon a été initialement gérée (=ReferenceBiblio du SINP). **Ne pas reprendre**
collectionID	JddId	Remarque : Ne sera pas probablement pas celui de l'IDCNP.
collectionCode	JddCode	
catalogNumber	IdentifiantOrigine	
occurrenceID	IdentifiantPermanent	La plateforme Thématique lui en attribue un.
	DSPublique	« NSP » pour Ne sait pas
-	**CodeIDCNPDispositif**	Obligatoire non récupérable. Possibilité de créer un dispositif de collecte pour « donnée obtenue par GBIF » ou analyse plus fine : dépend du projet IDCNP. **A définir**
Processor & « GBIF »	OrganismeStandard	Concaténer la valeur « GBIF » à l'information de Processor si elle existe
OccurrenceStatus	StatutObservation	Si Occurrencestatus n'est pas renseigné alors est valorisé dans StatutObservation en « présent ». Si le champ DWC est « absent » alors ne pas reprendre car l'occurrence est potentiellement une donnée de synthèse
ScientifiqueName	NomCite	
-	CdNom	Retrouver le CD_NOM s'il existe par la plateforme
-	CdRef	Retrouver le CD_REF s'il existe par la plateforme
-	Sensible	Problématique de la plateforme qui attribue la sensibilité de la donnée
Individualcount	DenombrementMin	
-	DenombrementMax	Le champ n'existe pas
	TypeDenombrement	« NSP » pour Ne sait pas
	ObjetDenombrement	Par défaut le count concerne l'individu : « Individu »
recordedby	IdentiteObservateur	L'information sera à diviser dans les deux champs ce qui va probablement demander un effort de standardisation non négligeable car non automatisable mais nécessaire car les 2 informations sont obligatoires
recordedby	OrganismeObservateur	
InstitutionCode	OrganismeGestionnaireDonnees	Vu la flexibilité du DarwinCore, cette correspondance n'est pas forcément toujours vraie.
identifiedby	Determinateur	
-	Validateur	Ne pas essayer de renseigner
eventremarks	Commentaire	Ne pas traduire
eventdate	DateDebut	Peut présenter juste une année, juste un mois, qu'il faudra modifier en une période d'incertitude dans le SINP. Les périodes peuvent être traitées dans 5 formats et les dates dans 4 avec possibilité de traiter l'heure dans eventDate.
eventdate	DateFin	
eventdate	HeureDebut	Vérifier dans eventdate si l'heure est présente
eventdate	HeureFin	
dateIdentified	DateDeterminationObs	

minimumElevationInMeters	AltitudeMin	
maximumElevationInMeters	AltitudeMax	
minimumDepthInMeters	ProfondeurMin	
maximumDepthInMeters	ProfondeurMax	
	Habitat	Ne pas essayer de renseigner
	RefHabitat	
decimalLongitude, decimalLatitude ? et / ou footprintwkt	Geometrie	Transformer dans le format de la plateforme SINP
coordinateUncertaintyInMeters	Precision	
	NatureObjetGeo	Par défaut « Inventoriel » car doute dans les données GBIF si l'objet représente l'aire de prospection et non l'observation
	CodeCommune	Valoriser par le codeCommune par la plateforme
municipality	NomCommune	Vérifier que cela corresponde à une commune INSEE
HigherGeographyID	CodeEN	Ne pas reprendre
HigherGeographyID	TypeEN	Ne pas reprendre
HigherGeographyID	CodeMaille	Ne pas reprendre
HigherGeographyID	CodeME	Ne pas reprendre
-	Attributs additionnels	Ne pas essayer de les renseigner à partir des fichiers Dwc.
-	ThematiqueAttribut	Ne pas essayer de renseigner
-	NomAttribut	Ne pas essayer de renseigner
-	DefinitionAttribut	Ne pas essayer de renseigner
-	ValeurAttribut	Ne pas essayer de renseigner
-	UniteAttribut	Ne pas essayer de renseigner
Champs des métadonnées		
NameAccordingtoID	Id. du ref. taxonomique	0e61f8fe-7d25-4f81-ada7-d970bbb2c6d6
NameAccordingto	Référentiel Taxonomique	Si le CdNom est bien renseigné alors sera valorisé à « TAXREF »
Autres Champs		
	kingdom	Les champs de classification peuvent aider à retrouver le code taxon dans TAXREF mais ne sont pas repris en tant que tel
	phylum	
	class	
	order	
	family	
	genus	

Annexe : Code Pays

Country : code https://www.iso.org/obp/ui/#search Code « Officially assigned » (sauf pour FX)

Code	**Short name**	
FR	France	Comprises: Metropolitan France, French Guiana, Guadeloupe, Martinique, La Réunion, Mayotte, Saint Barthélemy, Saint Martin, Saint Pierre and Miquelon, French Polynesia, French Southern Territories, New Caledonia, Wallis and Futuna. Includes: Clipperton Island.
TF	French southern territories	Bassas da India, Crozet Archipelago, French scattered Indian Ocean Islands, Europa Island, Glorioso Islands , Juan de Nova Island, Kerguelen Islands, Saint-Paul Island, Tromelin Island, Amsterdam Island Comprises: Amsterdam Island, Crozet Archipelago, Kerguelen Islands, Saint Paul Island and French scattered Indian Ocean Islands formed by Bassas da India, Europa Island, Glorioso Islands, Juan de Nova Island and Tromelin Island
GP	Guadeloupe	Includes: la Désirade, Marie-Galante, les Saintes
MQ	Martinique	
GF	French Guiana	
YT	Mayotte	
RE	Réunion	
NC	New Caledonia	Includes: Loyalty Islands
MF	Saint Martin (french part)	The island of Saint Martin is divided into the northern French part and the southern Dutch part
WF	Wallis and Futuna	Futuna, Hoorn Islands, Uvea, Wallis Islands. Comprises: Hoorn Islands (Principal island: Futuna), Wallis Islands (Principal island: Uvea)
PM	Saint Pierre and Miquelon	
PF	French Polynesia	Austral Islands , Gambier Islands, Marquesas Islands , Society Archipelago , Tahiti, Tuamotu Island Comprises: Austral Islands, Gambier Islands, Marquesas Islands, Society Archipelago (Principal island: Tahiti), Tuamotu Islands.
FX	*France, Metropolitan*	*Exceptionnaly reserved (Not Officially assigned)* *Refers to Metropolitan France and reserved at the request of France*