IUT2 Université Grenoble Alpes
Département STID

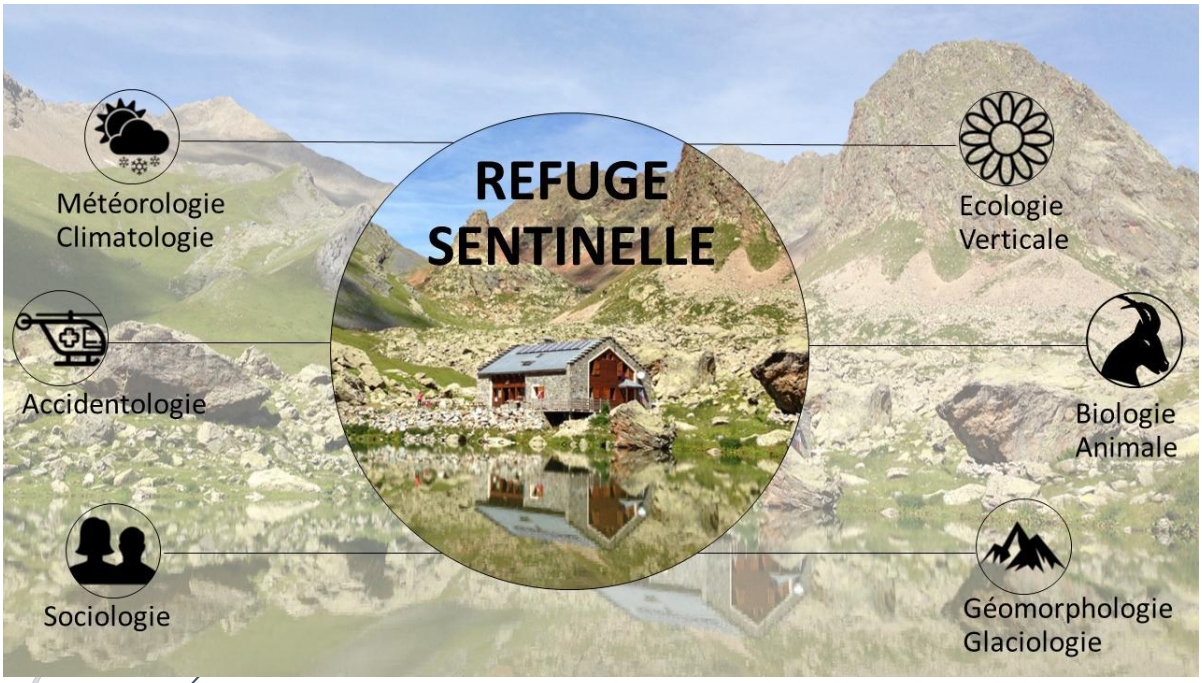
Pacte Laboratoire des sciences sociales

RefLab

2019 - 2020

Mémoire académique

Comment structurer des données multi-sources et multi-format afin de les interroger conjointement pour répondre à des problématiques d'analyse spatiale ?



BAKALARZ David
UGA - DUT STID – 2ÈME ANNEE

Remerciements

Je tiens à remercier le Laboratoire PACTE et le programme RefLab pour leur confiance et pour m'avoir permis de travailler avec eux dans le cadre de mon stage qui est à l'origine de ce mémoire.

Je remercie tout particulièrement Mme Raffaella BALZARINI, ma tutrice, pour son aide, sa confiance et son écoute lors de mon stage. Je remercie également Mme Mélanie MARCUZZI et M. Phillipe BOURDEAU, tous deux membres de RefLab, qui m'ont également fait confiance durant ce stage ainsi que M. Camille MONCHICOURT pour son intervention.

Je remercie aussi Mme Marlène VILLANOVA-OLIVER professeure tutrice pour son aide mais également Mme Edwige FAIN pour sa disponibilité durant la période difficile du COVID-19.

Table des matières

Liste des figures

I) Introduction

La problématique que j'ai choisie de présenter dans ce mémoire est : « Comment structurer des données multi-sources et multi-format afin de les interroger conjointement pour répondre à des problématiques d'analyse spatiale ? ». J'ai choisi cette problématique car elle est en lien avec mon stage dans le laboratoire des sciences sociales Pacte. Ce laboratoire est une unité de recherche du CNRS, de l'Université Grenoble Alpes et de Sciences Po Grenoble implanté principalement sur le site universitaire Grenoble Alpes. Pacte rassemble la majorité des géographes, politistes, sociologues et urbanistes du site et accueille également des économistes et historiens et a pour objectif de construire des connaissances sur les transformations de nos sociétés dans des dimensions politiques, territoriales, sociologiques et écologiques. Mon stage est intégré dans le cadre du programme Refuge Sentinelles. Ce programme vise à développer un dispositif expérimental d'observation du changement en haute montagne avec, comme lieu de mesure, les différents refuges. Les études s'orientent à la fois sur des processus géophysiques, climatiques et biologiques mais également sur les pratiques touristiques et sportives. L'enjeu de ce programme est également de construire des questions croisées et de réaliser des missions communes sur le terrain pour obtenir des résultats pluri et interdisciplinaires.

L'utilisation de données spatiales est donc totalement en accord avec les objectifs et enjeux de ce programme. L'information spatialisée constitue l'élément principal des analyses relatives à l'impact du changement climatique sur les représentations du milieu étudié, sur la vulnérabilité, le risque et sur la notion d'expérience touristique. Dans ce contexte pluridisciplinaire, une base de données correctement structurée permet aussi d'utiliser des données de sources et format divers de manière efficace.

Par ailleurs, mon stage s'inscrit dans la continuité du travail d'un autre stagiaire qui a travaillé sur la création d'une base de données que j'ai donc dû faire évoluer en base de données spatiale. Cette base a pour objectif d'organiser les données traitant des refuges de hautes et moyennes montagnes du parc naturel des Ecrins, qui sera le sujet d'étude principal de ce mémoire. Cette base était donc, en partie, déjà créé et mes objectifs étaient de :

- établir un diagnostic de la maquette BDD précédemment réalisée lors du stage de 2019 et concevoir l'évolution de la base

- enrichir et faire évoluer la base de données afin qu'elle soit totalement exploitable sur un logiciel de cartographie, sachant qu'un des objectifs à terme pour le programme Refuge Sentinelles est de réaliser une carte dynamique et interactive, accessible aux non-initiés.

Je vais donc expliquer les différentes étapes de mon stage que j'ai réalisé pour atteindre ces objectifs, en commençant par l'étude des schémas sous PostgreSQL, la réalisation du schéma relationnel de la base, la récupération et le nettoyage des données sous Excel, la création du script SQL sous NotePad++ pour la structuration et l'alimentation de la base, l'importation de ce script sur Postgre et la création du lien entre la base de données sur PostgreSQL et le logiciel de cartographie Qgis via l'extension Postgis.

II) Contexte

Comme énoncé précédemment, le travail sur la problématique de ce mémoire est en lien avec le stage que j'ai réalisé dans le cadre du programme Refuge Sentinelle. Ce programme existe depuis maintenant plus de quatre ans et est lui-même intégré dans un contexte de recherche scientifique très complet. Le programme est financé grâce au projet d'excellence LabEx ITEM (Innovation et Territoire de Montagne), porté par l'Université Grenoble-Alpes et le Parc national des Écrins qui est le lieu d'étude et d'expérimentation. Il est également partenaires avec de nombreux acteurs :

- **Partenaires 2017-2019 : Agence française pour la biodiversié, Parc national des Écrins, FFCAM, Jardin alpin du Lautaret, Office de tourisme La Grave-Villar d'Arène, Associations des gardien.ne.s des Hautes-Alpes et de l'Isère, Syndicat National des Gardiens de Refuges et gites d'étape, Educ'Alpes, Fondation Petzl, Musée Dauphinois, Collectif MuséoMix, LowTech Lab INPG Grenoble**
- **Laboratoires scientifiques impliqués 2017-2018 : PACTE, IGE, LECA, SENS, CERAG, INRIA, IRSTEA, EDYTEM, IREGE + Labex ITEM, Zone Atelier Alpes du CNRS et CDP Trajectories (IDEX UGA)**

Figure 1 : liste des partenaires du programme Refuge Sentinelle (source : Rapport scientifique 2019 de M. Philippe BOURDEAU)

Les enjeux de ce programme sont de développer les connaissances sur un double chainon manquant de l'observation en sciences de la nature et en sciences sociales :

- Tout d'abord sur la haute-montagne, un terrain sous-observé de part son accès restreint et ses contraintes climatiques mais pourtant témoin de phénomènes fortement liés aux changement climatiques.

- Mais aussi sur les refuges de ces montagnes, qui sont d'une part des sites idéaux d'études de part leurs ressources, la convergence de flux et la présence humaine mais également des sites en mutation de statut et de fonctions car les refuges sont contraints à évoluer et à s'adapter sur place aux changements.

Ce programme a donc comme objectif : « d'observer les changements en haute montagne, en faisant converger sur un panel de refuges des travaux de sciences de la nature et de sciences sociales, en partenariat avec les parties prenantes professionnelles, faciliter et développer de nouveaux croisements formels et informels entre sciences sociales et sciences de la nature et expérimenter un dispositif interdisciplinaire et participatif d'intelligence climatique, territoriale et culturelle. » (Source : Présentation Refuge sentinelle mars 2017).

Comme l'explique M. Philippe BOURDEAU dans son rapport scientifique de 2019, les travaux conduits depuis 2017 portent sur des données à la fois qualitatives mais aussi quantitatives :

- données sur la fréquentation des refuges et la pratique estivale de la montagne (relevés de destinations post-refuges, nuitées à l'année et jour par jour)

- données sur la clientèle des refuges (questionnaires avec 1800 réponses)

- interviews sur le métier de gardien, l'évolution de la culture professionnelle en lien avec les changements climatiques et culturels

- interviews sur l'expérience touristique en refuge (n = 80)

- données de géolocalisation de pratiques hors-sentiers et itinéraires, et de bivouacs

- 12 mémoires de master sur un large éventail de thèmes

Problématique :

Afin de réaliser ces travaux, le programme Refuge Sentinelle travaille directement avec les refuges du parc des Ecrins. Chaque année, les refuges s'engagent sur des points précis comme participer aux enquêtes fréquentation, accepter d'installer de l'équipement météorologique, participer aux interviews, etc. Mais ces engagements changent suivant les années (un refuge peut revoir ses engagements), et certains protocoles mis en place par le programme évoluent aussi (séparation d'un questionnaire global en plusieurs questionnaires plus précis et complets, nouveaux équipements scientifiques, etc). De plus, toutes ces actions sont menées « au même niveau », sans logique de « flux de données », il se peut donc que certaines données se répètent dans les différents protocoles.

C'est donc dans ce contexte d'études des effets croisés des changements environnementaux et culturels que la problématique « Comment structurer des données multi-sources et multi-format afin de les interroger conjointement pour répondre à des problématiques d'analyse spatiale ? » prend tout son sens. Étant donné le nombre de protocoles de récoltes de données, le contexte pluridisciplinaire de ce programme et l'absence de logique de « flux de données », les données deviennent donc multi-sources et imposent donc au programme Refuge Sentinelle, un traitement spécifique pour celle-ci. Mais elles sont aussi multi-format : quantitatives, qualitatives mais aussi géographiques, un nouveau format de données que je n'ai pas encore pu beaucoup étudier à STID. Dans la problématique, je parle aussi de les interroger conjointement. En effet, dans ce contexte pluridisciplinaire, on ne peut envisager de les interroger autrement que conjointement. Le croisement des disciplines implique un croisement des données si l'on veut obtenir les meilleurs résultats possibles. Mais le croisement de données multi-sources et multi-format nécessite une réflexion plus profonde qu'un simple croisement de données puisque le risque d'erreur est beaucoup plus important.

La principale problématique était de passer d'une logique d'expérimentation de terrain, menée sur plusieurs fronts mais pas aux mêmes moments ni aux mêmes endroits et traduite, au mieux, dans une logique de "fichiers Excel", à une logique de vie de la donnée qui la rend exploitable et reproductible.

III) Méthodologie

Dans cette partie, je vais vous présenter la méthodologie que je propose afin de répondre à la problématique de ce mémoire, que j'illustre par des exemples issus de mon stage, où j'ai également essayé d'appliquer cette méthodologie.

1) Présentation des schémas

Comme énoncé précédemment mon premier objectif était de reprendre le travail du précédent stagiaire, le comprendre et commencer à avoir des idées pour faire évoluer la base de données. La base était hébergée sous PostgreSQL, un outil nouveau pour moi. PostgreSQL possède une particularité que certains systèmes de gestion de base de données non pas. Il s'agît de schémas. C'est un élément qui, hiérarchiquement, est au-dessus des tables dans le schéma relationnel. Un schéma peut donc contenir plusieurs tables. En termes de code SQL, la création des schémas ressemble beaucoup à celle des tables :

```
CREATE SCHEMA Exemple;
```

Figure 2 : Exemple de script pour la création des schémas

Ces schémas sont principalement utilisés pour « ranger » les tables de manière organisée, dans le but de simplifier la vision d'ensemble de la base, mais aussi de simplifier certaines requêtes SQL.

Lors de notre réflexion sur les schémas de la base, on m'a demandé de réfléchir également à une hiérarchie pour les schémas. L'objectif étant de différencier les schémas avec des données structurelles ou d'administration qui ne risquent pas de changer dans le temps, avec les schémas dit de « mesures » qui seront les schémas utilisés pour les études du programme Refuge Sentinelles, avec des données issues principalement d'enquêtes et qui seront mises à jour dans le futur. Cette hiérarchie de schémas n'existe pas réellement dans PostgreSQL, c'est une organisation subjective de schémas pour mieux les comprendre. La logique voudrait qu'il y ait deux schémas pour distinguer les deux types de données énoncées précédemment. Mais j'ai préféré séparer encore en deux les données structurelles et administratives en mettant dans le 1er niveau les données identitaires des refuges et dans le 2nd des données complémentaires sur les refuges. Je suis donc arrivé à trois niveaux de schémas.

- 1er niveau avec des données sur l'identité des refuges, leur localisation, quel type de refuge…
- 2ème niveau avec des informations complémentaires sur les refuges comme leurs médias, leurs organisations administratives, la règlementation, les périodes d'ouverture…
- 3ème niveau avec des données issues de différentes enquêtes, c'est le niveau des schémas « mesures »

Chaque niveau est composé d'un ou plusieurs schémas :

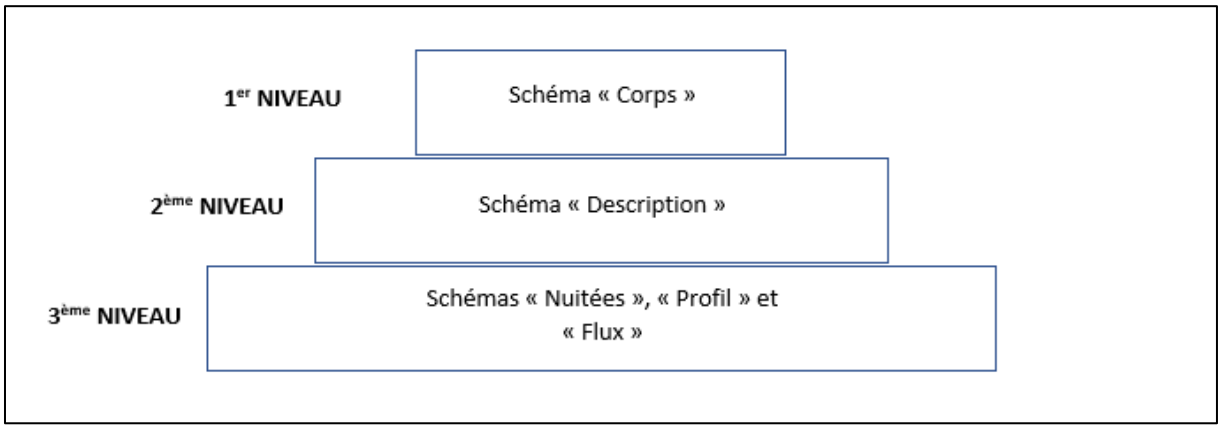


Figure 3 : Hiérarchie des schémas

Après ce travail en amont sur l'organisation des schémas, j'ai travaillé sur le schéma relationnel de la base. En effet, la base de données avait déjà été en partie créée par le précédent stagiaire. J'ai donc ajouté les nouveaux schémas et tables que nous avions imaginées sans oublier les clés primaires de chaque table ainsi que les clés référentielles qui permettent l'intégrité référentielle de la base. L'intégrité référentielle d'une base de données est quelque chose de fondamental pour qu'une base de données soit correctement construite. C'est une contrainte que l'on s'impose pour que les données restent utilisables et cohérentes et que l'on ne puisse pas supprimer des données qui dépendent les unes des autres.

Ce schéma m'a également permis de me rendre compte que la table « Identité » dans le schéma « Corps » est certainement la table la plus importante car pratiquement toutes les tables font référence à celle-ci (via l'identifiant du refuge).

De plus, en travaillant sur ce schéma relationnel, j'ai remarqué que pratiquement toutes les tables de la base ont une cardinalité 1-1 avec la table « Identité » (c'est-à-dire qu'une occurrence de la première entité peut correspondre à une et une seule occurrence de la seconde entité). Cela signifie que théoriquement, on pourrait regrouper l'intégralité des tables ayant une cardinalité 1-1 avec la table « Identité » en une seule et même table. Le choix de séparer les données en plusieurs tables est encore une fois un choix subjectif de notre part.

2) Présentation des tables

Dans cette partie seront détaillées les différentes tables de la base. Le schéma relationnel simplifié de cette base de données peut être observé en annexe. Aussi, pour mieux comprendre les données présentes dans ces tables, vous pourrez observer en annexe, le modèle de donnée relatif au volet « Fréquentation et pratiques ».

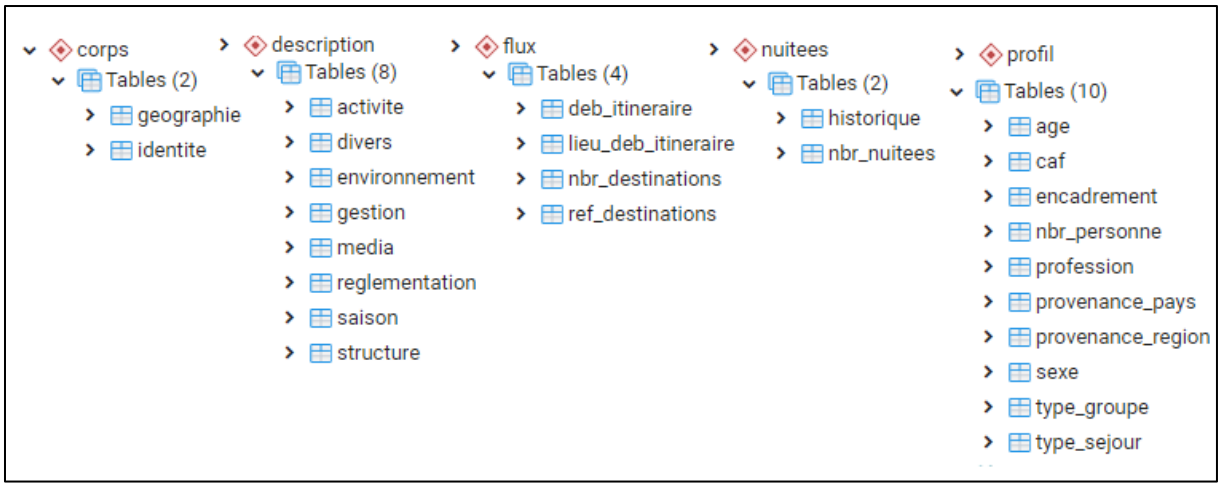


Figure 4 : Liste des tables par schéma

a) Schéma Corps

Table « Identité » : La table au cœur de la base, elle lie principalement les refuges à un ID qui sera utilisé dans les autres tables.

Table « Géographie » : Cette table apporte des informations géographiques sur les refuges comme l'altitude, le massif, la région, …

Un élément à noter sur ces tables : elles contiennent toutes deux des données sur des refuges nationaux (différent du parc des Ecrins). Ces données sont présentes afin de préparer le travail qui sera réalisé dans les prochains mois voire années, pour élargir la base à tous les refuges de France.

De plus, la table « Identité » possède les seules données géographiques présente pour le moment dans cette base. Il s'agît des coordonnées géographiques des différents refuges que l'on utilisera pour localiser les refuges sur nos cartes d'analyses spatiales.

b) Schéma Description

Table « Environnement » : Cette table parle principalement des sources d'énergies utilisées par le refuge, la gestion des déchets, l'approvisionnement en eau, …

Table « Réglementation » : Cette table traite de l'accueil des mineurs ainsi que du statut de protection PN et PNR.

Table « Média » : Cette table contient les différents sites internet des refuges.

Table « Gestion » : Cette table est composée des éléments sur l'administration du refuge, l'aspect financier, …

Table « Saison » : Cette table traite des données sur les périodes d'ouverture.

Table « Structure » : Cette table donne des informations sur la date de construction du refuge et l'équipement dont elle dispose (couvertures, douches, chauffage, …).

Table « Divers » : Une table d'informations diverses.

Table « Activité » : Cette table traite des données sur le type d'activité proposé dans les différents refuges.

c) Schéma Flux

Table « Début itinéraire » : Cette table traite les informations sur le type de lieu de départ pour aller vers un refuge donné (depuis un parking, un autre refuge ou un autre point de départ).

Table « Lieu du début d'itinéraire » : Cette table complète les informations de la table « Début itinéraire » en précisant le nom du lieu de départ.

Table « Ref_Destination » : Il s'agit d'une table qui référence toutes les destinations atteignables par les refuges.

Table « Nbr_Destination » : Cette table donne la mesure du nombre de personne ayant choisi une destination depuis un refuge donné. Pour le moment, elle ne contient que les données de 2019 mais elle pourrait être complétée avec les années précédentes et les années futures.

Ce schéma ne possède pas de données géographiques. Mais une discussion a récemment commencé sur l'idée d'utiliser ces données dans un cadre d'analyses spatiales en ajoutant de nouvelles données ou en transformant des données existantes en données géographiques. En effet, il pourrait être intéressant d'étudier le trajet des pratiquants / clients des refuges. Nous pourrions alors avoir des données de type « point » pour les points de départ et destinations possibles, mais aussi des données poly-lignes pour représenter le parcours des pratiquants.

d) Schéma Nuitées

Table « nbr_nuitées » : Cette table donne des informations sur le nombre de nuitées annuelles pour une année donnée ainsi que le nombre de jour d'ouverture total.

Table « Historique » : Cette table est un historique du nombre de nuitées annuelle pour les refuges depuis 1946.

e) Schéma Profil

Pour les tables du schéma Profil, elles contiennent les mesures liées à une enquête liée à l'étude socio-démographique réalisée sur les pratiquants / clients des refuges. Ces tables seront complétées à mesure que des enquêtes seront réalisées.

	uid smallint	nom character varying (220)	type character varying (220)	ecrin boolean	position geometry
1	1537	ADELE PLANCHARD	gardé	true	01010000008CDB6...
2	1488	AIGLE	gardé	true	0101000000CEAAC...
3	1529	ALPE DU PIN	gardé	true	010100000036E50...
4	1547	ALPE DU VILLAR D ARENE	gardé	true	0101000000EFC9C...
5	1538	BANS	gardé	true	0101000000C442A...
6	2166	BERARDE	gardé	true	01010000003FA9F6...
7	6024	CARRELET	cabane non gardée	true	0101000000E0BE0E...
8	1669	CEZANNE	cabane non gardée	true	01010000007CED9...
9	1767	CHABOURNEOU	gardé	true	0101000000C38190...
10	1276	CHALANCE	non gardé	true	010100000036CD3...
11	5572	CHAMOISSIERE	gardé	true	0101000000986E12...
12	1753	CHATELLERET	gardé	true	0101000000FBE8D...
13	1552	CLOT (XAVIER BLANC)	gardé	true	010100000099F562...
14	2319	CLOTS	gardé	true	0101000000B35E0...
15	1540	DU SELE	gardé	true	0101000000726DA...
16	1535	ECRINS	gardé	true	01010000000E677E...
17	1741	EVARISTE CHANCEL	gardé	true	010100000012A0A...
18	2216	EYCHAUDA	gîte d étape	true	0101000000645DD...

Figure 5 : Exemple de table : Table Identité

J'ai choisi de vous montrer comme exemple la table « Identité » car en plus d'être la table la plus importante de la base, elle possède des données géographiques.

Comme on peut le voir sur l'image, cette table possède cinq colonnes. La 1ère est l'identifiant associé au refuge. La 2nd est le nom du refuge. La 3ème colonne est le « type » du refuge, c'est-à-dire refuge gardé ou non, cabane, gîte d'étape… La 4ème permet de préciser si le refuge fait partie ou non du parc naturel des Ecrins et la dernière contient les données géographiques sur la localisation des refuges. On peut observer un œil dans la colonne position qui permet d'afficher les données géographiques sous la forme choisie (ici sous forme de points), comme on peut le voir sur la figure 6.

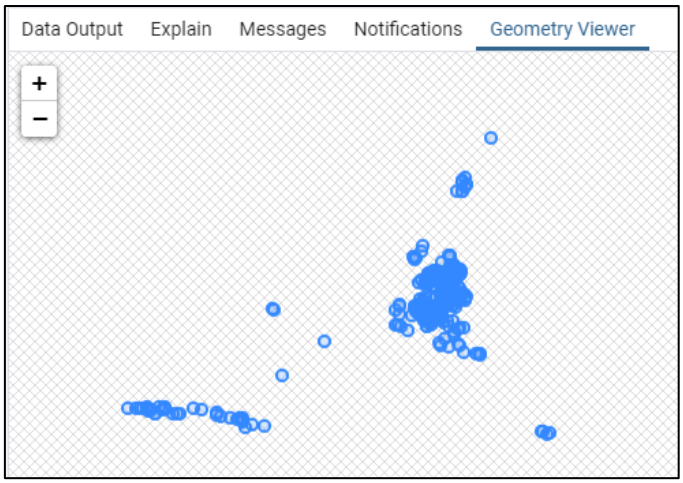


Figure 6 : Visualisation des données géographiques de la colonne position de la table « Identité »

3) Récupération et nettoyage des données

Les données m'ont été transmises au fur et à mesure de mon stage. Les premiers fichiers avaient déjà été traités et nettoyés auparavant. Ma mission consistait à les insérer dans la base. Il s'agissait des fichiers liés aux schémas Corps et Description.

Pour les données étant liées aux « mesures », mon travail avant de les insérer, consistait à les préparer et à les nettoyer. Ce travail était principalement réalisé sous Excel. La première étape était de regrouper l'ensemble des données issues de différents fichiers dans une seule et même feuille Excel. Ces fichiers provenaient d'enquêtes réalisées par différentes équipes liées au programme Refuge Sentinelle (soit des équipes de RefLab soit du parc des Ecrins). Ils pouvaient aussi provenir de scraping d'informations sur les différents sites web ou encore avoir été donnés par des entités partenaire du parc des Ecrins. Enfin, certaines données sont issues de plusieurs protocoles réalisés par le programme Refuge Sentinelle. Ces protocoles pouvaient être réalisés en parallèle, c'est-à-dire réalisé par différentes personnes à différents endroits. Toutes ces sources nous amènent à la seconde étape qui était de supprimer les doublons existants. En effet, les données provenant de sources multiples, il arrivait souvent que certaines données soit présentes plusieurs fois. Or, dans le cadre d'une réalisation d'une base de données relationnelle (géographiques ou non) il faut impérativement supprimer les doublons pour éviter tout problèmes avec les clés primaires et étrangères. Pour rappel, pour que l'intégrité référentiel soit respectée, la clé primaire ne doit avoir que des valeurs uniques. Pour faire cette suppression de données, j'ai utilisé l'outil « Supprimer les doublons » dans Excel.

Enfin la dernière étape était de supprimer tout caractère spécifique (apostrophes, parenthèses, virgules, …) qui aurait pu générer des erreurs dans le code. En effet, dans le script SQL les données sont, par exemple, séparées par des virgules. Les chaînes de caractères, elles, commencent et se terminent par des apostrophes. Pour cette étape, j'ai beaucoup utilisé l'outil « Rechercher Remplacer ». Dans cette étape, il m'arrivait également de mettre les chaînes de caractères entre apostrophes pour gagner du temps sur la partie SQL. Pour cela, j'utilisais la formule présentée dans la figure 7.

L'objectif principal du passage des données dans Excel est de pouvoir préparer le plus proprement possible ces données avant d'être ajoutées dans le script SQL. Lorsque les données étaient propres, j'enregistrais la feuille Excel en format csv que j'ouvrais ensuite sous Notepad++. Ainsi, toutes les données étaient déjà placées dans le script, propres, et je n'avais qu'à ajouter la syntaxe pour l'insertion des données dans la base (qui sera présentée dans la prochaine partie).

Ce travail m'a également permis de mettre en évidence le manque de données, le manque de « contexte » ou encore de « temporalité » de certaines tables. Par la suite, nous avons pu corriger certains de ces manques (lorsque c'était possible).

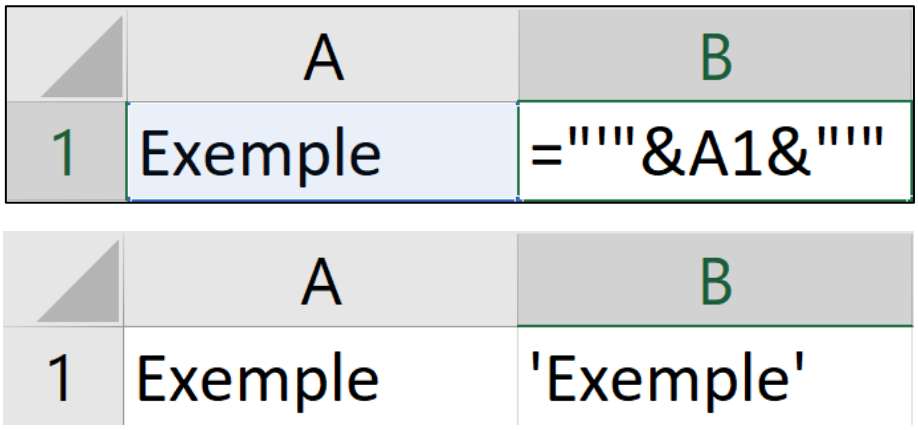


Figure 7 : Formule Excel pour ajouter les apostrophes

4) Création du script SQL

Pour simplifier et alléger les fichiers, le script est séparé en deux parties. Une pour la structure de la base, c'est-à-dire les schémas, les tables, les clés primaires et étrangères. La seconde partie est composée du script d'insertion des données de toutes les tables dans la base.

a) Partie Structure

Ce premier script est également composé de plusieurs parties : le début du script permet de créer les différents schémas et l'extension Postgis si elle n'existe pas encore dans la base. L'extension Postgis permet de pouvoir utiliser des données géographiques, en proposant de la syntaxe supplémentaire pour traiter ces données par exemple, mais elle est surtout utile pour pouvoir créer un lien entre notre base de données sous PostgreSQL et Qgis. Cette extension peut être installée comme « éléments supplémentaires » lors de l'installation de PostgreSQL. Sur la figure 8, on peut voir à la ligne 6 un exemple de syntaxe pour la création d'un schéma (dans ce cas, le schéma « Description ») et à la ligne 16 la ligne de code permettant de créer l'extension Postgis (seulement si elle n'existe pas).

```
3
4   --Création des schémas
5
6   CREATE SCHEMA Description;
7
8   CREATE SCHEMA corps;
9
10  CREATE SCHEMA Nuitees;
11
12  CREATE SCHEMA Profil;
13
14  CREATE SCHEMA Flux;
15
16  CREATE EXTENSION IF NOT EXISTS postgis WITH SCHEMA corps;
```

Figure 8 : Exemple de script pour la création des schémas et l'extension Postgis

La seconde partie permet de créer les différentes tables de la base. Pour illustrer cette partie j'ai choisi de vous montrer la table « identité » car elle possède une colonne avec des données géographiques. Ce genre de colonne est quelque chose de nouveau pour moi puisque ces éléments ne sont pas étudiés dans le programme du DUT STID.

Pour créer une table, on utilise donc la syntaxe « CREATE TABLE » (ligne 1 de la figure 9) en précisant par la suite dans quel schéma on souhaite la créer (ici corps), et comment on veut la nommer (ici Identité). Ensuite, sur les lignes 2 à 5 de la figure 9, on crée les différentes colonnes de la table en précisant le nom de la colonne ainsi que son type de données (boolean, varchar, text, …). Pour la colonne « uid », on précise également que la valeur ne doit pas être NULL. Enfin, pour créer la colonne des données géographiques, j'ai dû utiliser une des syntaxes que propose l'extension Postgis et que l'on peut voir sur la dernière ligne de la figure 9. Ce code permet de créer la colonne géométrique en précisant le schéma, la table, le nom de la colonne, le srid (identifiant de référence spatiale associé à un système de coordonnées géographiques), le type de forme géométrique et le nombre de dimension de la donnée (2D, 3D, …). Dans cet exemple, on décide de créer la colonne dans le schéma « corps », la table « identité », on la nomme « position », on utilise le srid 0, on veut que les données soient de type « POINT » et en deux dimensions. Dans le but de rendre le code plus simple à « lire », j'ai décidé de mettre en commentaire le nom de la colonne « position » à la suite des autres colonnes créées avec la syntaxe de base. Cela me permet, quand je travaille sur le script, de me rappeler rapidement que la colonne position est bien présente dans la table « identité ».

```
CREATE TABLE corps.Identite (
    uid smallint NOT NULL,
    nom VARCHAR(220),
    type VARCHAR(220),
    Ecrin boolean
    --position
);

SELECT AddGeometryColumn ('corps','identite','position',0,'POINT',2);
```

Figure 9 : Exemple de script pour la table Identité

La dernière partie permet d'ajouter les différentes clés primaires et clés étrangères de la base. Sur l'exemple de la figure 10, on peut voir toutes les clés créées pour les deux tables du schéma « corps ». Sur la ligne 3 de ce code, la syntaxe « ALTER TABLE ONLY » permet de changer la définition d'une table donnée, ici il s'agît de la table « Identité » du schéma « corps ». Sur la ligne 4, j'utilise la syntaxe « ADD CONSTRAINT … PRIMARY KEY (…) » pour créer la clé primaire de la table « identité ». Après « ADD CONSTRAINT », je donne un nom à la clé, toujours en essayant de garder une logique dans la nomination (en général, j'utilise le nom de la table suivi de _pkey pour les clés primaires et le nom de la table suivi de _fkey pour les clés étrangères). Ensuite, après « PRIMARY KEY » je précise entre parenthèse quelle donnée sera la clé primaire : ici j'ai choisi « uid » car il s'agît de l'identifiant du refuge. Puis, sur les lignes 7 et 8 j'ai utilisé la même syntaxe pour créer la clé primaire de la table « Géographie ».

Enfin, sur les deux dernières lignes, j'ai créé une clé étrangère avec la syntaxe « FOREIGN KEY … REFERENCES … ». La différence ici, c'est que l'on doit préciser à quelle clé primaire la clé étrangère fait référence (car toutes les clés étrangères font référence à une clé primaire d'une autre table). Dans cet exemple la clé étrangère de la table « Géographie » fait donc référence à la clé primaire de la table « Identité ». Cela implique que tous les « uid » insérés dans la table « Géographie » doivent obligatoirement être présents dans la table « Identité ». Cette contrainte nous force donc à devoir créer les clés primaires avant les clés étrangères.

```
--Clé relationnelle
    --Schéma Corps
ALTER TABLE ONLY corps.Identite
    ADD CONSTRAINT Identite_pkey PRIMARY KEY (uid);

ALTER TABLE ONLY corps.Geographie
    ADD CONSTRAINT Geographie_pkey PRIMARY KEY (uid);

ALTER TABLE ONLY corps.Geographie
    ADD CONSTRAINT Geographie_fkey FOREIGN KEY (uid) REFERENCES corps.Identite(uid);
```

Figure 10 : Exemple de script pour les clés relationnelles

b) Partie Données

Ce second script est uniquement composé des données des différentes tables. Il n'y a qu'un seul fichier pour toutes les données de la base et ce code est composé de 3040 lignes. Un élément très important à prendre en compte, comme dans le script précédent, il faut faire attention à l'ordre dans lequel les données sont insérées dans la base. La table « identité » est donc la première à être intégrée (car pratiquement toutes les tables ont une clé étrangère faisant référence à la table « identité »).

Pour intégrer les données dans la base, j'ai utilisé la syntaxe « INSERT INTO … VALUES … ».

La première partie de cette syntaxe, « INSERT INTO … », permet de choisir dans quel schéma et dans quelle table on veut insérer les données. A la suite de la syntaxe, on précise donc le nom du schéma suivi du nom de la table (ici « corps.Identite ») puis, on donne entre

parenthèses le nom des colonnes de la table (pour « Identite » on a donc uid, nom, type, Ecrin et position).

```
INSERT INTO corps.Identite(uid,nom,type,Ecrin,position) VALUES (
```

Figure 11 : Exemple de script pour l'insertion des données (INSERT INTO)

La deuxième partie de la syntaxe, « VALUES … » permet de mettre les données à intégrer pour chacune des colonnes précisées dans la partie « INSERT INTO ». Dans l'exemple de la figure 12, le code permet d'intégrer le refuge « 1451 » du nom de « La Loge », c'est un refuge de type « non gardé » et il ne fait pas partie du parc des Ecrins.

```
VALUES (1451,'La Loge','non gardé',false,ST_GeometryFromText('POINT(5.95427 46.29274)'));
```

Figure 12 : Exemple de script pour l'insertion des données (VALUES)

Toujours avec sa particularité, la table « identité » nécessite un script différent pour insérer les données géographiques (les données de la dernière colonne). Ce script est possible avec l'extension Postgis, il s'agît de la fonction « ST_GeometryFromText ». Cette syntaxe permet de récupérer des données géographiques (ici les coordonnées longitude et latitude du refuge choisi) et de les intégrer dans la base dans le format voulu (ici sous format point). Il est aussi possible avec cette fonction de préciser le srid mais il n'était pas nécessaire de le préciser ici.

```
ST_GeometryFromText('POINT(6.34515 44.96892)')
```

Figure 13 : Fonction ST_GeometryFromText

5) PostgreSQL et PgAdmin

Lorsque les scripts étaient prêts, ils étaient importés dans PgAdmin 4. PgAdmin est une interface graphique installée en même temps que PostegreSQL et qui permet d'administrer la base de données assez facilement. Pour importer les données dans la base, je faisais un copié / collé du code des fichiers NotePad++ vers la partie « Create Script » de PgAdmin. Lorsque cela était fait, j'exécutais le code. La figure 14 est un exemple de message renvoyé par PgAdmin pour nous dire que le code a correctement été exécuté.

L'objectif de ces scripts est de donner une version de la structure de la base et une version des données les plus « à jour » possible afin que les personnes devant reprendre ce travail puisse le faire dans de bonnes conditions. Etant donné qu'il y a encore beaucoup de discutions autour de la base de données, il n'est pas impossible que ces scripts changent encore. A terme, la

partie structure ne devrait plus être modifiée et la partie donnée ne sera mise à jour que lorsque de nouvelles données seront ajoutées (nouvelles enquêtes, mesures, etc).

```
NOTICE:  l'extension « postgis » existe déjà, poursuite du traitement
ALTER TABLE

Query returned successfully in 1 secs 22 msec.
```

Figure 14 : Exemple de message renvoyé par PgAdmin après exécution du code

6) Création du lien entre PostgreSQL et Qgis

Après avoir importé les données dans la base, j'ai créé le lien entre la base de données sur PostgreSQL et le logiciel de cartographie (dans mon cas Qgis). Pour cela, j'ai utilisé l'extension Postgis qui a permis de créer une connexion entre la base de données sur PostgreSQL et Qgis et donc de pouvoir utiliser les données de la base avec Qgis.

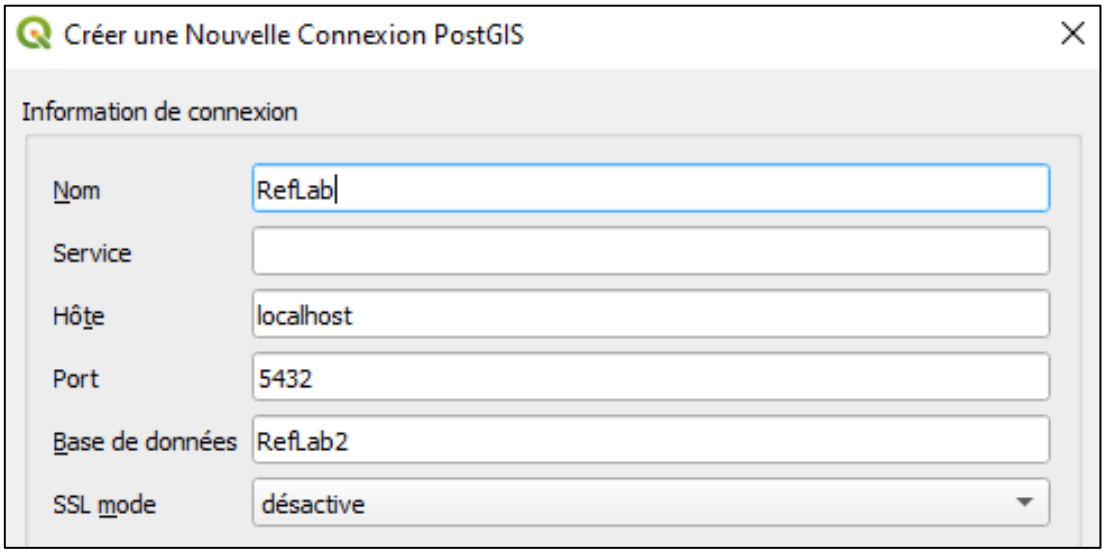


Figure 15 : Fenêtre de création de la connexion PostgreSQL / Qgis

Après avoir créé cette connexion, les données sont prêtes à être interrogées et étudiées dans le cadre d'analyses spatiales. La table « identité » étant liée directement ou indirectement à toutes les tables de la base, il est possible sur Qgis de récupérer toutes les informations concernant un refuge. La partie études spatiales sera commencée sur les deux dernières semaines de mon stage.

La figure 16 est la première carte « test » réalisée sur Qgis, où il est possible de voir l'ensemble des refuges présents dans la base. Le fond de carte est le fond de carte d'OpenStreetMap disponible par défaut dans Qgis.

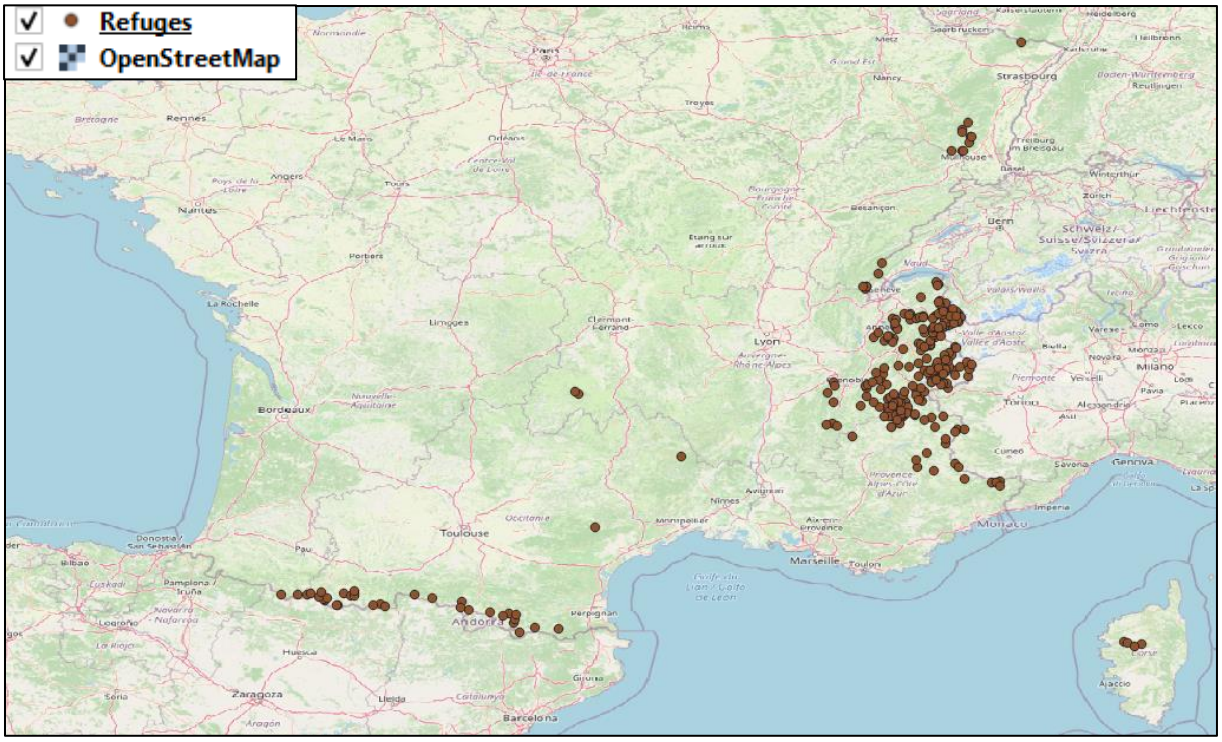


Figure 16 : Première carte Qgis réalisée avec les données de la base géographique

7) Autre réalisation

En plus de ma mission principale évoquée précédemment, j'ai également réalisé un travail sur la situation de la base de données et la qualité de la donnée. L'objectif de ce travail était de donner un indicateur sur la « qualité de la donnée » afin de savoir si les protocoles réalisés en amont lors de la récolte des données, étaient bien réalisés ou non. Pour cela, j'ai mis une petite description pour chacune des tables afin de recontextualiser, et j'ai calculé un ratio du nombre de valeur NULL sur le nombre de valeurs totales possibles.

Table « Identité » :

Nombre valeur NULL	0
Nombre valeur possible	560
Rapport	0%

Figure 17 : Exemple de ratio pour la table « Identité »

IV) Compétences mobilisées

Dans cette partie, je vais développer en termes de compétences, les apports de mon stage et du travail réalisé sur la problématique de départ.

Pour commencer, j'ai appris à utiliser un nouveau logiciel (SGBD) de gestion et de manipulation de données : PostgreSQL et PgAdmin 4. C'était pour moi une opportunité de pouvoir apprendre à utiliser ce nouvel outil non étudié à STID pour mes futures études mais également pour ma future carrière dans le monde professionnel.

Par ailleurs, j'ai aussi dû élaborer une structure appropriée pour la base de données géographique. Lors de mon stage, j'ai pu me rendre compte à plusieurs reprises qu'une mauvaise organisation / structure de départ pour la base (ou tout autre sujet) peut être source de problèmes et confusions multiples. Ce travail d'élaboration de structure est très intéressant dans le monde professionnel, puisque tout le monde échange, propose des idées sur ce point, et cela permet de limiter les oublis et les erreurs. L'outil principal utilisé dans ce cas est le schéma relationnel de la base, qui peut mettre rapidement en évidence les futurs problèmes que l'on pourrait rencontrer.

De plus, afin d'être sûr du bon fonctionnement de la base, j'ai exploité la compétence de savoir constituer, consolider et tester la base de données. Lors de mon travail, les données de la base ont été modifiées plusieurs fois. Ces modifications provenaient de discussions avec des professionnels, de changements de structure dans la base ou encore d'erreurs dans les données qui ont été repérées après avoir testé la base avec des requêtes SQL types.

J'ai également utilisé les requêtes SQL pour interroger la base de données. Ces requêtes étaient utiles pour s'assurer que toutes les données été « accessibles » comme nous l'avions imaginé au moment où nous avions créé le schéma relationnel de la base.

J'ai aussi dû remettre en question mon travail et apprendre des erreurs. Lors de mon stage, nous avons contacté le gestionnaire de la base de données du parc des Ecrins, M. MONCHICOURT pour avoir un point de vue extérieur sur notre base. Une des erreurs principales que j'ai efefctuées durant ce stage est d'avoir conservé la structure « Excel » pour certaines tables. Par la suite, j'ai su ne pas reproduire cette erreur et créer des tables propres et adéquates à la base.

J'ai aussi pu observer comment M. MONCHICOURT a validé ou non la base de données et ces différents éléments. Aujourd'hui, j'essaie d'utiliser sa méthode de travail et de réflexion lorsque je travaille sur la base de données, toujours dans le but de réduire les erreurs avant vérifications. Cela complète mon travail sur le compte-rendu de la base (qui était focalisé sur la qualité de la donnée).

Lors de la création des scripts SQL, j'ai dû mettre en place des commentaires pour que mon travail puisse être repris sans problèmes par un autre stagiaire, et ceci sans qu'il ait besoin d'une formation importante sur les bases de données. Ces commentaires sont placés directement dans le script ou dans le compte-rendu sur la qualité de la donnée de la base.

Ce travail m'a également demandé de comprendre des données issues d'un milieu très spécifique. Cette compréhension des données est un objectif fondamental puisque tout le travail réalisé par la suite sur le schéma relationnel, les liens entre les tables, la hiérarchie des schémas Postgre, découle de cette compréhension. Elle permet aussi de pouvoir communiquer et comprendre les autres personnes qui travaillent avec nous sur le même sujet et nous permet plus facilement de commencer à imaginer des études ou des analyses sur les données.

V) Conclusion

Pour répondre à la problématique de départ, on peut donc voir qu'il est important de créer, structurer et organiser une base de données géographiques pour gérer des données multi-sources et multi-format. Une base de données adaptée, comme celle que j'ai utilisée, permet de traiter toutes les données conjointement en limitant un maximum d'erreur. Cette base permet aussi d'utiliser les données sur différents logiciels comme dans mon cas Qgis qui sera utilisé dans un contexte d'études et d'analyses. La maîtrise du langage SQL est à mon sens un des éléments le plus important, d'une part pour une partie test et vérification du bon fonctionnement de la base mais aussi pour pouvoir interroger conjointement toutes les données et donc réaliser des études multivariées.

En perspectives, je pense qu'il serait très important de réaliser rapidement des études sur Qgis afin de se rendre compte si la base est bien structurée ou si d'autres changements sont nécessaires avant de commencer des gros projets d'études. Aussi, il serait important de reprendre l'intégralité de la base en revue dans le cadre des refuges nationaux. Ces refuges ont peut-être un fonctionnement différent de ceux présents dans le parc des Ecrins et pourraient donc ne pas correspondre à la structure actuelle de la base en termes de données.

Ce travail a été pour moi une expérience très enrichissante sur le plan professionnel. Malgré un contexte compliqué avec l'épidémie du Covid-19 et la mise en place du télétravail, je pense avoir réussi à m'adapter à la situation assez rapidement, à comprendre les objectifs et missions qui m'ont été confiés et à produire des résultats cohérents par rapport à ce qui m'était demandé.

Cette expérience de télétravail m'a permis de découvrir ce qu'étaient les visioconférences. Elles m'ont permis de présenter mon travail et mes résultats et d'en discuter dans un cadre type réunion avec une ou plusieurs personnes, où il m'était demandé de présenter et également d'animer la présentation.

La principale difficulté que j'ai rencontrée est la gestion du temps de travail et des délais de restitutions des résultats. Cette idée de délais a beaucoup été travaillé à STID, mais dans un contexte de milieu professionnel, cette gestion est assez différente. Cela m'a permis de sortir de ma « zone de confort » et donc de commencer à réfléchir à une nouvelle méthode de travail adapté au milieu professionnel. Cette méthode pourra être utilisée et améliorée pour ma future alternance et future carrière professionnelle.

Bibliographie

PostgreSQL : https://www.postgresql.org/

Postgis : https://postgis.net/

PgAdmin 4 : https://www.pgadmin.org/

Qgis : https://www.qgis.org/fr/site/

NotePad++ : https://notepad-plus-plus.org/downloads/

Excel : https://www.microsoft.com/fr-fr/microsoft-365/excel

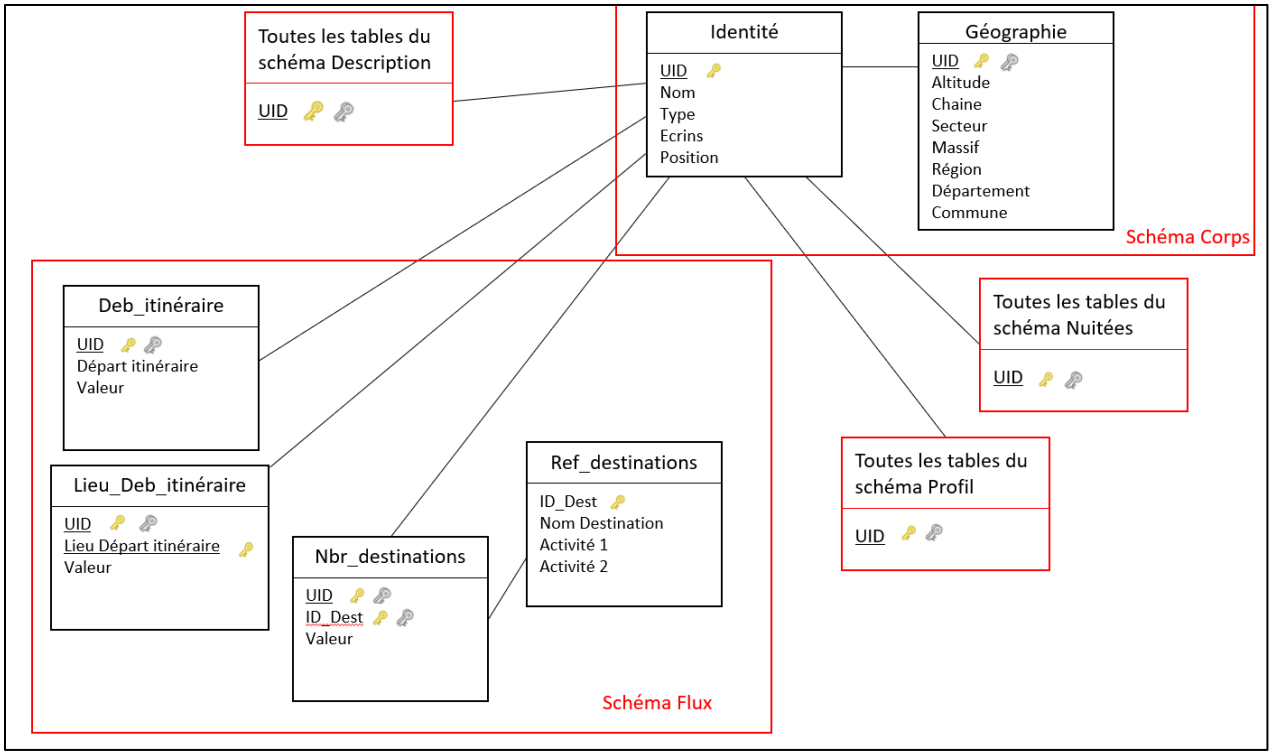
Laboratoire PACTE : https://www.pacte-grenoble.fr/

Programme Refuge Sentinelle : http://www.ecrins-parcnational.fr/actualite/refuges-sentinelles-observatoire-haute-montagne

RefLab : https://reflab.hypotheses.org/

Annexes

Schéma relationnel simplifié



Modèle de données relatif au volet « Fréquentations et pratiques »