

Titre du projet: Rédaction de requêtes SQL pour la récupération d'information au sein d'une base de données en écologie moléculaire

Personnes impliquées

Responsable: Florian Malard (UMR CNRS 5023 LEHNA UCBL1, Florian.Malard@univ-lyon1.fr)

Autres personnes impliquées: Philippe Grison (UMS 3468 BBEES, MNHN, Paris « Bases de données Biodiversité, Écologie, Environnement Sociétés) Lara Konency, Christophe Douady, Tristan Lefébure (UMR CNRS 5023 LEHNA UCBL 1)

Nature du projet:

Dans le cadre d'un développement informatique en toute autonomie de l'équipe d'étudiants, il s'agira de :

- 1) Rédiger des requêtes SQL permettant de récupérer des informations au sein d'une base de données en écologie moléculaire selon un format défini par les chercheurs
- 2) Eventuellement (fonction de l'expérience pratique des étudiants en bioinformatique), de développer une interface web proposant des formulaires de recherche pour les requêtes SQL rédigées.

La base de données en écologie moléculaire

Pour lire ce paragraphe, le lecteur s'appuiera sur les annexes 1 et 2 du présent document.

Au titre de leurs activités de recherche, les membres de l'UMR CNRS 5023 (LEHNA) collectent sur le terrain (en Europe et dans le monde) des organismes aquatiques et les identifient à partir de techniques relevant de la morpho-anatomie (dissection et montage des organismes sur lame) et de la biologie moléculaire (obtention de séquences d'ADN). Cette activité de recherche génère un ensemble de données / informations relatives aux :

- Lieu de prélèvement (la **STATION**)
- Méthode de prélèvement (la **COLLECTE**)
- Groupe d'individus issus d'une collecte (le **LOT MATERIEL interne**)
- Un individu issu d'un lot matériel et utilisé pour l'analyse moléculaire et/ou morpho-anatomique (l'**INDIVIDU**)
- l'ADN issu de cet individu (**ADN**)
- la PCR effectuée sur cet ADN (**PCR**)
- le chromatogramme issu de cette PCR (**CHROMATOGRAMME**)

- les séquences d'ADN issues d'un ou plusieurs chromatogramme(s) (**SEQUENCES internes**)
- les entités taxonomiques moléculaires opérationnelles identifiées à partir d'un grand nombre de séquences d'ADN (**MOTU**)

L'ensemble de ces données / informations est stockée dans une base de données dont la structure est présentée schématiquement en **annexés 1 (modèle simplifié) et 2 (modèle complet)**. Par ailleurs, la base de données contient également des informations relatives à des lots matériels et des séquences d'ADN collectés par d'autres laboratoires que le LEHNA (**LOT MATERIEL Externe** et **SEQUENCES externes**) qui ont été publiés dans la littérature (**SOURCES**).

Type de travail demandé et outils utilisés

1) Rédiger des requêtes SQL.

La base de données déposée sur le serveur de l'IN2P3 est accessible *via* l'outil PgAdmin 4 (**Annexe 3**) où se fera la rédaction des requêtes SQL. Trois exemples de formulation de requête par les utilisateurs et les résultats obtenus suite à la rédaction de ces requêtes sont fournis en **annexes 4 à 6**.

2) Développer une interface web proposant des formulaires de recherche pour les requêtes SQL.

Le cadre technologique du développement de cet interface web devra être compatible avec l'interface de gestion de la base de données, à savoir l'utilisation de :

- framework PHP Symfony3.
- ORM Doctrine (les requêtes devront être appelées via l'appel du manager de Doctrine)
- langage PHP 7
- bdd PostgreSQL
- les technologies pour le design: le framework Bootstrap et le langage de templates TWIG
- les librairies javascript : jquery, Ajax

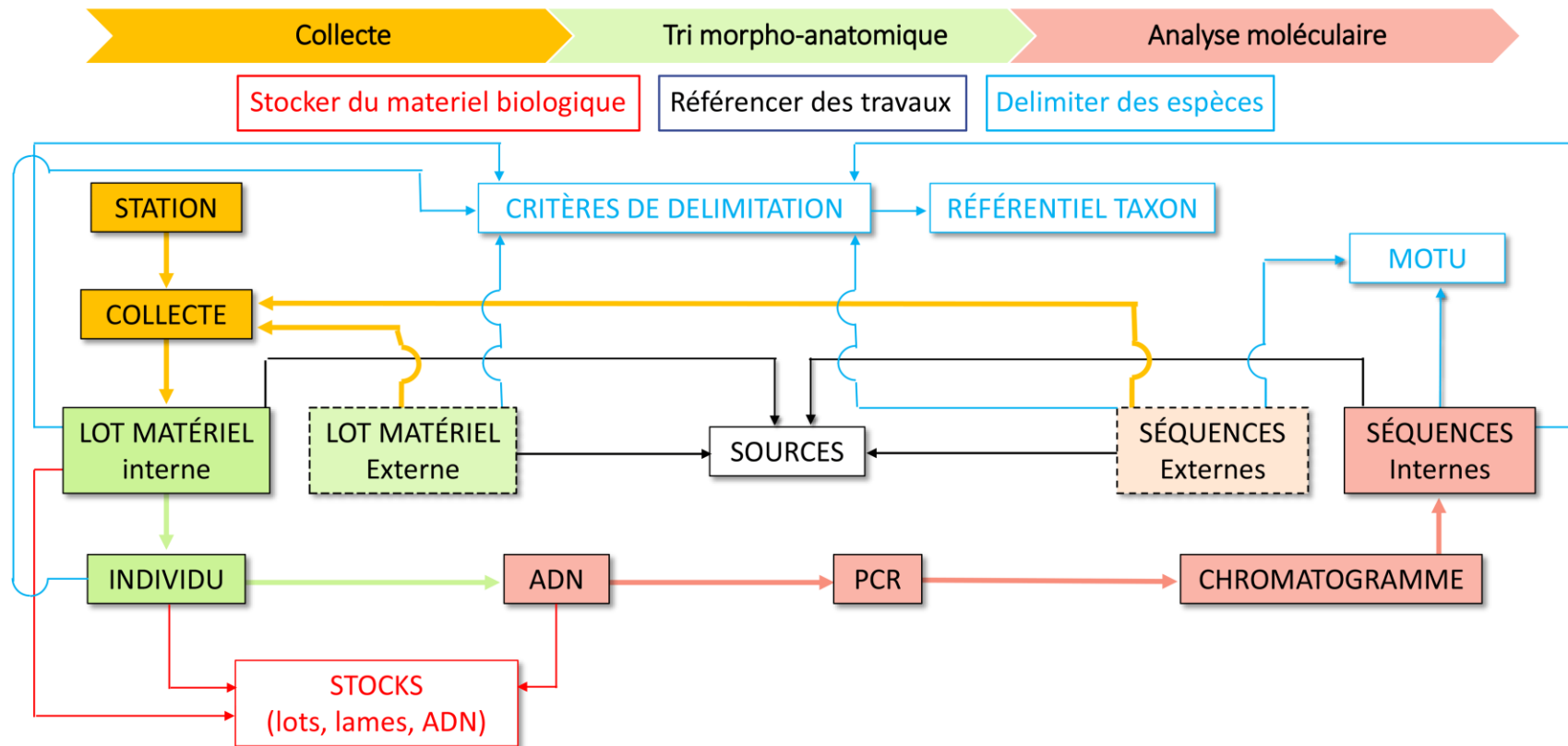
L'ensemble du projet pourra être interfacé directement avec la base de données hébergée au CC-In2p3 sous la condition d'un IP fixe préalablement fourni au CC-In2p3 (filtrage IP) ou éventuellement à partir d'un dump de la base de données

Implication des chercheurs dans le projet

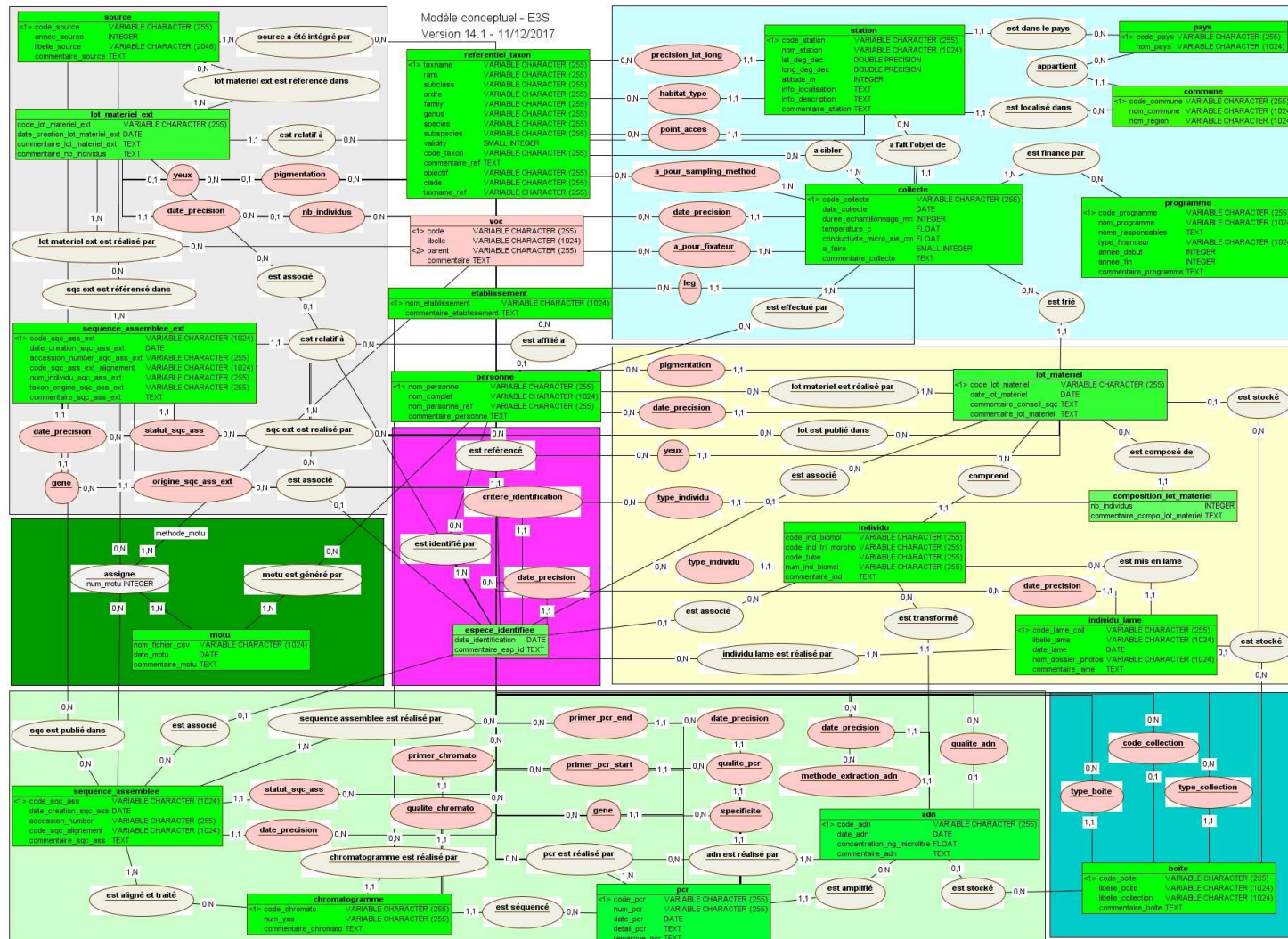
Les chercheurs pourront accompagner les étudiants dans l'expression du cahier des charges, la compréhension de la structure de la base de données nécessaire à la rédaction des requêtes et la fourniture de liens vers des documentations utiles.

Annexe 1 : La structure de la base de données (modèle simplifié)

La structure de la base de données



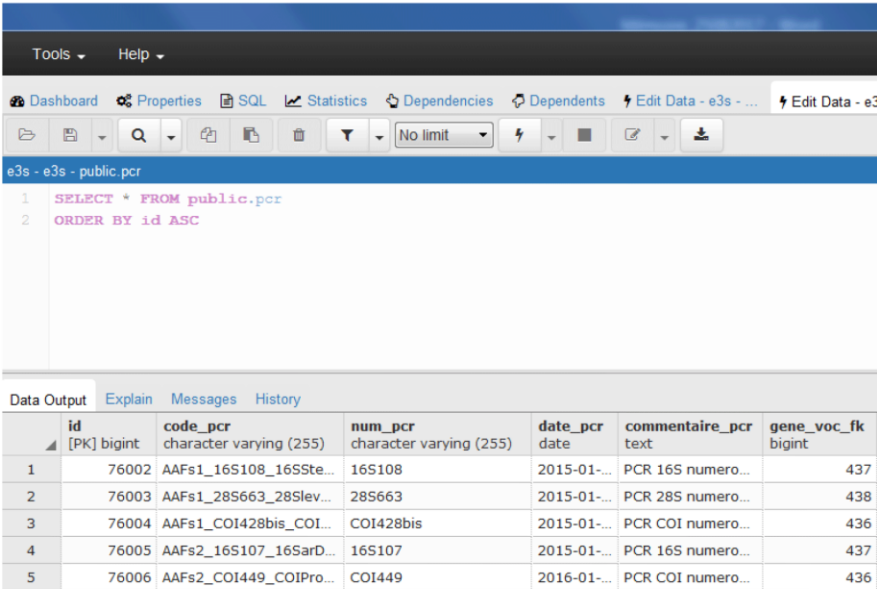
Annexe 2: La structure de la base de données (modèle complet)



Annexe 3 : Serveur et outil pour la rédaction des requêtes

- Base de données déposés sur le serveur de l'IN2P3 (Institut National de Physique nucléaire et de Physique des Particules)

- Base de données accessible *via* PgAdmin 4



The screenshot shows the PgAdmin 4 web interface. At the top, there's a navigation bar with 'Tools' and 'Help' menus. Below it, a toolbar contains icons for Dashboard, Properties, SQL, Statistics, Dependencies, and Dependents. The main area displays a SQL query in a text editor:

```
1 SELECT * FROM public.pcr
2 ORDER BY id ASC
```

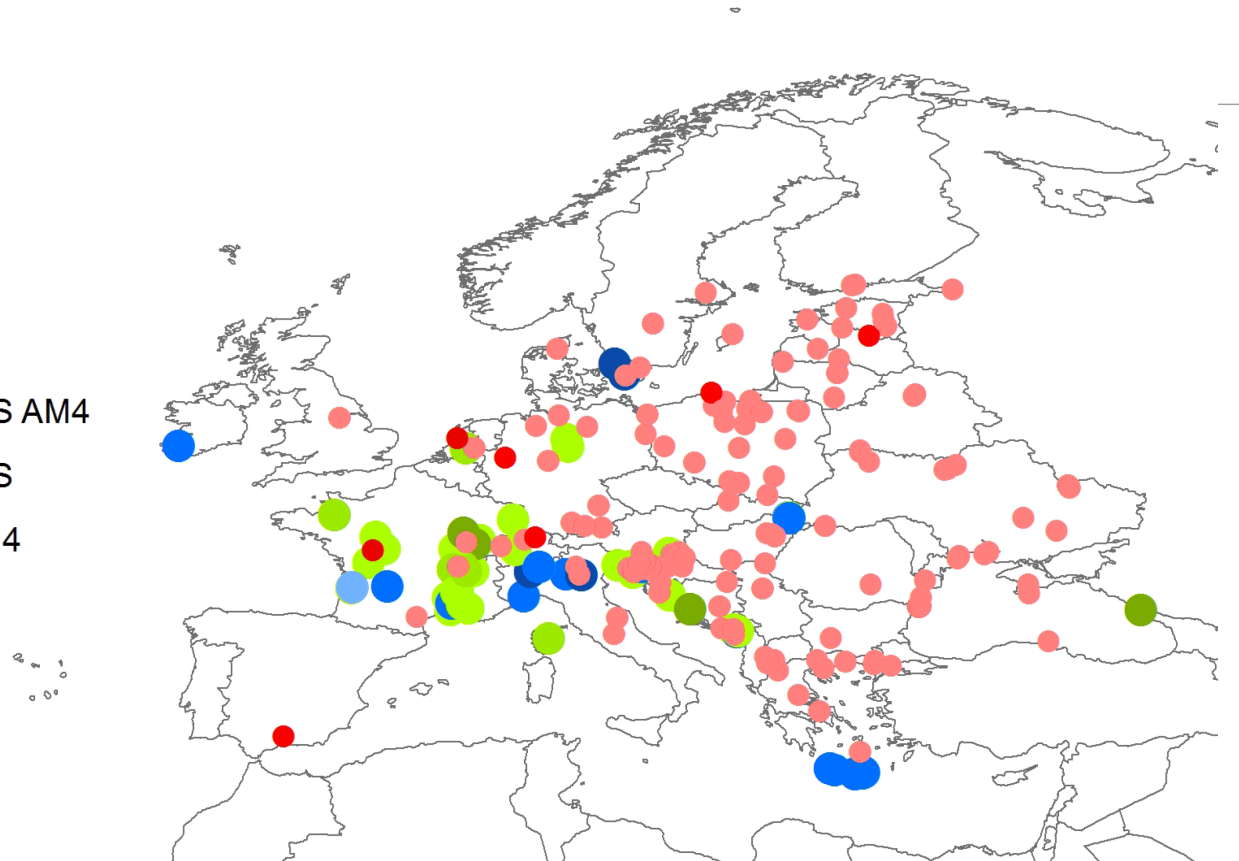
Below the query editor, there's a tabbed interface with 'Data Output', 'Explain', 'Messages', and 'History'. The 'Data Output' tab is active, showing a table with 7 columns: id, code_pcr, num_pcr, date_pcr, commentaire_pcr, and gene_voc_fk. The table contains 5 rows of data.

	id [PK] bigint	code_pcr character varying (255)	num_pcr character varying (255)	date_pcr date	commentaire_pcr text	gene_voc_fk bigint
1	76002	AAFs1_16S108_16SSte...	16S108	2015-01-...	PCR 16S numero...	437
2	76003	AAFs1_28S663_28Slev...	28S663	2015-01-...	PCR 28S numero...	438
3	76004	AAFs1_COI428bis_COI...	COI428bis	2015-01-...	PCR COI numero...	436
4	76005	AAFs2_16S107_16SarD...	16S107	2015-01-...	PCR 16S numero...	437
5	76006	AAFs2_COI449_COIPro...	COI449	2016-01-...	PCR COI numero...	436

Annexe 4 : Exemple 1 de requête. Un chercheur souhaitant lancé un projet de recherche utilisant l'espèce *Asellus aquaticus* comme modèle biologique souhaite connaître le matériel biologique dont il dispose (lot matériel, individus, séquences, etc...). Les données extraites par requête SQL sont ensuite cartographiées (hors requête).

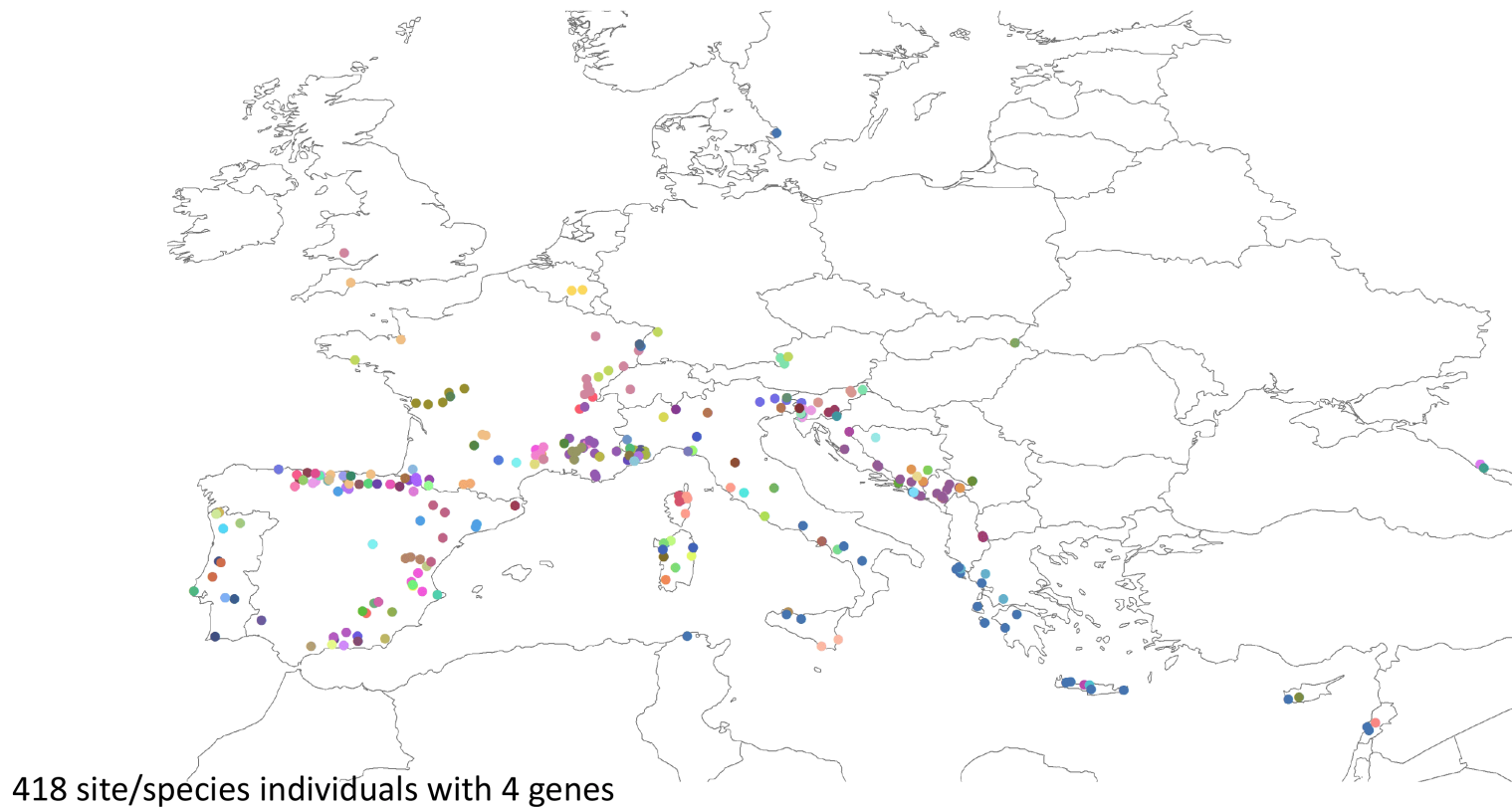
Exemple 1: que dispose-t-on sur l'espèce *A. aquaticus*?

- SQ_ASS_EXT, 16S
- SQ_ASS_EXT, 28S
- SQ_ASS_EXT, COI
- IND_BIOMOL,
- IND_TRI_MORPHO,
- LOT_MATERIEL,
- SQC_ASS, 16S COI 28S AM4
- SQC_ASS, 16S COI 28S
- SQC_ASS, 16S COI AM4
- SQC_ASS, 16S COI
- SQC_ASS, 16S AM4
- SQC_ASS, 16S
- SQC_ASS, AM4



Annexe 5 : Exemple 2 de requête. Un chercheur souhaitant réaliser une phylogénie avec 4 gènes souhaite connaître les espèces et leurs répartitions (stations) pour lesquels il dispose de quatre gènes sur un même individu. Les données extraites par requête SQL sont ensuite cartographiées (hors requête).

Exemple 2: combien de stations / espèces avec 4 genes différents ?



Annexe 6 : Exemple 3 de requête. Un chercheur souhaitant séquencer un individu de l'espèce *Proasellus walteri* veut sélectionner le primer qui a le plus de chance de fonctionner (la base sert également d'outil métier). Les données extraites par requêtage SQL sont ensuite exprimées sous forme graphique (hors requête).

Exemple 3: quels primers pour *P. walteri*?

Quels sont les couples de primers qui donnent des PCR dont la qualité PCR est TB (très bien), pour les individus associés à l'espèce *Proasellus walteri*

