LAGPLON

Importation de données et interface hors-ligne

Action ZONECO

Guénolé BOUVET
GEOLITTO

22/02/06

Sommaire

Introduction

Le système LAGPLON actuel s'appuie sur les technologies web. L'ajout de données ou leur modification ne peuvent se faire qu'en ligne (depuis un poste connecté à internet).

Cependant, des procédures ont été développées afin d'intégrer des données historiques « en bloc » dans le système à partir d'un fichier Excel. Ces procédures ne sont qu'en partie automatisées et ne sont pas actuellement utilisables par l'IRD.

Afin de permettre une autonomie plus importante pour les plongeurs de l'IRD et plus précisément, afin de leur donner la possibilité de constituer hors-ligne des fichiers de données prêts à être intégrés dans la base de données, il convient d'améliorer les possibilités d'importation et d'exportation de données disponibles dans le système actuel.

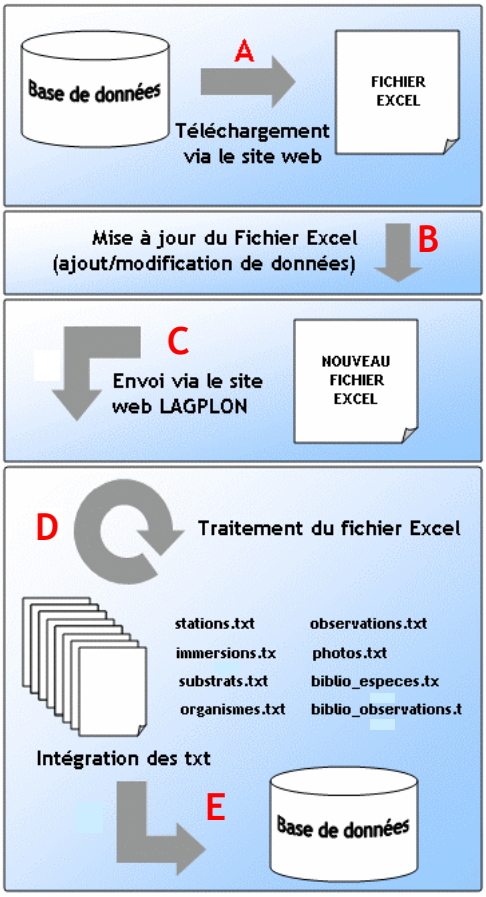
Ce document propose dans un premier temps une solution technique pour l'importation des données mises à jour puis, dans un deuxième temps, les développements possibles afin d'améliorer l'ergonomie du côté des producteurs de données.

1. Importation en bloc de données mises à jour

1.1. Principe général

La procédure est clairement définie mais son automatisation n'est pas complète. Elle est constituée de 4 grandes étapes (figure 1) :

A- téléchargement du fichier Excel contenant toutes les données de la base ;
B- Mise à jour du fichier Excel contenant les données ;
C- Renvoi du fichier vers le serveur ;
D- traitement du fichier Excel et constitution des 8 fichiers txt d'importation ;
E- intégration des données.



L'étape A permet aux plongeurs de l'IRD de récupérer l'intégralité des données de la base. Il dispose ainsi des informations nécessaires pour l'ajout de nouvelles données ; ils peuvent notamment connaître les identifiants déjà utilisés pour les stations, les espèces, les observations.

Les étapes B et C sont réalisées par les plongeurs de l'IRD : les fichiers de données sont mis à jour (modification d'enregistrements existants, ajout ou suppression). Les étapes D et E doivent être réalisées automatiquement sur les serveurs de la DTSI.

Figure 1 : chaîne d'importation de données en bloc

Les procédures pour les étapes C et D restent à être construites. Les procédures permettant la réalisation de l'étape E sont bien avancées (certains points pouvant encore être améliorés).

La transformation du fichier Excel en 8 fichiers txt n'est pas une nécessité mais les procédures actuellement développées pour l'étape E ne fonctionnent qu'avec des fichiers txt.

1.2. Description détaillée

1.2.1. Téléchargement du fichier Excel

Une page du site web doit permettre de télécharger un fichier Excel contenant toutes les données de la base. Ce fichier Excel doit impérativement être construit « à la volée » afin d'intégrer les dernières modifications qui auraient pu avoir lieu sur les données.

Etat du sytème actuel : le téléchargement proposé ne permet pas d'obtenir un fichier Excel contenant toutes les données, mais 8 fichiers txt. Les plongeurs de l'IRD doivent alors recomposer un fichier Excel à partir de ces fichiers.

1.2.2. Préparation du fichier Excel

Le fichier Excel doit contenir les 8 feuilles suivantes :

- stations : descriptions des stations ;
- immersions : descriptions des plongées (lieu, date, but, etc.) ;
- substrats : descriptions des substrats ;
- organismes : description générale des espèces ;
- observations : description des observations et prélèvements ;
- photos * : liste des correspondances entre les photos et les espèces ;
- biblio_organismes* : liste des correspondances entre les références bibliographiques et les espèces ;
- biblio_observations* : liste des correspondances entre les références bibliographiques et les observations/spécimens ;

* les feuilles photos, biblio_organismes et biblio_observations ne doivent pas être mises à jour hors-ligne. Mais elles sont nécessaires pour toute nouvelle intégration de données afin de conserver les relations entre les organismes et les URL des photos (voir plus bas).

Dans chacune des feuilles :

- tous les champs vides doivent être remplacés par 99999999 ;
- tous les "retour charriot" qui pourraient figurer dans les champs doivent être supprimés ;
- les latitudes et longitudes en degrés ne doivent pas avoir de décimales (champs latdeg/longdeg) ;
- les ' doivent être remplacés par des ` ;

Etat du sytème actuel : ces opérations sont effectuées « à la main » avant l'intégration des données.

1.2.3. Envoi du fichier Excel

Le site web doit permettre aux producteurs de données d'envoyer le fichier Excel de données mises à jour vers le serveur de traitement qui intégrera ensuite automatiquement les données dans la base.

L'opération se fait par le biais d'un formulaire sur le site web :

Etat du sytème actuel : le fichier Excel est envoyé *par email* à l'administrateur de la base de données.

1.2.4. Constitution des fichiers txt

Après réception du fichier Excel de données mises à jour, les huits fichiers txt suivants doivent être constitués à partir du fichier Excel (séparateur : tabulation) :

stations.txt
immersions.txt
substrats.txt
organismes.txt
observations.txt
photos.txt
biblio_especes.txt
biblio_specimens.txt

Pour chacun des fichiers, les opérations suivantes doivent être réalisées :

- suppression de la première ligne (intitulé de colonne) et remplacement par une ligne vide ;
- remplacement des tabulations par des ### ;
- suppression des espaces inutiles (espaces après ### ou avant ###) – méthode : remplacement en boucle des « espace### » par des « ### » et des « ### espace » par des « ### » ;
- conversion des différents fichiers en UTF-8 - DOS ;

Etat du sytème actuel : ces opérations sont effectuées « à la main » par l'administrateur de la base de données.

1.3. Intégration des données dans la base

1.3.1. Sauvegarde du contenu de la base

Dans la mesure où l'intégralité du contenu de la base est écrasé par les nouvelles données importées, la première étape est de réaliser une sauvegarde de la base avant l'importation. Cette sauvegarde automatique permettra à l'utilisateur de récupérer les données en cas de mauvaise manipulation (importation d'un fichier de données ne contenant qu'une partie des données par exemple).

Remarque : les différentes sauvegardes doivent être téléchargeables sur le site web.

Etat du sytème actuel : ces fonctionnalités sont opérationnelles. La sauvegarde consiste en l'écriture des 8 fichiers txt. Elle pourrait donc être améliorée pour aboutir à un fichier Excel.

1.3.2. Effacement des tables

Afin d'éviter de nombreux problèmes (perte de lien entre tables, notamment), tout ajout de données effectué en utilisant la chaîne de traitement implique un effacement du contenu de la base de données avant une réécriture à partir des fichiers fournis pour la mise à jour.

Etat du sytème actuel : cette fonctionnalité est opérationnelle.

1.3.3. Mise à jour des tables

En respectant les contraintes d'intégrité de la base, les données sont importées et distribuées dans chacune des tables. Les numéros de séquences sont créés et reportés dans

les différentes tables à chaque fois que c'est nécessaire (afin d'établir les relations entre les différentes enregistrements de tables différentes).

A ce niveau, de nombreuses vérifications sur le formatage des données dans les fichiers Excel peuvent être réalisées avant même le lancement des requêtes SQL pour l'insertion des données.

Toute erreur doit être signalée en fin de traitement afin d'informer l'utilisateur sur le déroulement de l'importation.

Les étapes à réaliser sont les suivantes :

- effacement des tables ;
- ouverture du fichier stations.txt et intégration des données dans la table TSTATION ;
- ouverture du fichier immersions.txt et intégration des données dans la table TIMMERSION ; pour cette étape, l'identifiant de station doit être récupéré dans la table TSTATION pour chacune des plongées à insérer ;
- ouverture du fichier substrats.txt et intégration des données dans la table TSUBSTRAT ; pour cette étape, l'identifiant de la plongée doit être récupéré dans la table TIMMERSION puis, l'identifiant de station correspondant doit être récupéré dans la table TSTATION pour chacune des descriptions de substrat à insérer ;
- ouverture du fichier organismes.txt et intégration des données dans la table TORGANISME ;
- ouverture du fichier observations.txt et intégration des données dans la table TOBSERVATION ; pour cette étape, l'identifiant de la plongée doit être récupéré dans la table TIMMERSION puis, l'identifiant de station correspondant doit être récupéré dans la table TSTATION pour chacune des observations à insérer ;
- ouverture du fichier photos.txt et intégration des données dans la table TPHOTO ; pour cette étape, l'identifiant de l'espèce concernée doit être comparé avec l'ancien identifiant de l'espèce (ie l'identifiant avant effacement de la base) ; le nouvel identifiant d'espèce doit être récupéré dans la table TORGANISME et inséré pour chaque chemin de photo à insérer ;
- ouverture du fichier biblio_especes.txt et intégration des données dans la table TPHOTO ; pour cette étape, l'identifiant de l'espèce concernée doit être comparé avec l'ancien identifiant de l'espèce (ie l'identifiant avant effacement de la base) ; le nouvel identifiant d'espèce doit être récupéré dans la table TORGANISME et inséré pour chaque référence bibliographique à insérer et liée à une espèce ;
- ouverture du fichier biblio_observations.txt et intégration des données dans la table TPHOTO ; pour cette étape, l'identifiant de l'observation concernée doit être comparé avec l'ancien identifiant de l'observation (ie l'identifiant avant effacement de la base) ; le nouvel identifiant d'observation doit être récupéré dans

la table TOBSERVATION ; l'identifiant de l'espèce concernée doit être comparé avec l'ancien identifiant de l'espèce (ie l'identifiant avant effacement de la base) ; le nouvel identifiant d'espèce doit être récupéré dans la table TORGANISME et inséré pour chaque référence bibliographique à insérer et liée à une observation ;

- le nombre d'enregistrements dans chacune des tables doit être récupéré.

Etat du sytème actuel : cette fonctionnalité est opérationnelle. Le traitement des erreurs peut être amélioré.

1.3.4. Affichage des statistiques

A la fin du processus d'intégration, les enregistrements de chaque table sont comptés automatiquement et affichés sur la page web. L'utilisateur peut ainsi rapidement détecté un problème ayant eu lieu pendant l'intégration.

Etat du sytème actuel : cette fonctionnalité est opérationnelle.

1.4. Précautions à prendre

Dans le cas où la mise à jour des données est réalisée hors-ligne à partir du fichier Excel, il est capital qu'aucune modification du contenu de la base ne soit faite en ligne. En effet, puisque l'intégralité du contenu de la base est effacé lors d'une intégration à partir d'un fichier Excel de données mises à jour, toute modification qui aurait été réalisée en ligne serait perdue.

Ainsi, il serait intéressant de pouvoir bloquer les menus permettant d'ajouter, de modifier ou de supprimer des données en ligne.

2. Amélioration de l'ergonomie pour la mise à jour du fichier Excel

2.1. Etat actuel

Dans l'état actuel des choses, l'importation de données mises à jour nécessite pour l'utilisateur la constitution des 8 fichiers txt cités plus haut. A l'intérieur de ses fichiers, les identifiants sont parfois des triplets : station + date + heure. C'est le cas notamment pour tous les enregistrements d'observations ; chaque observation fait référence à une station et à une date, c'est à dire à une plongée donnée.

La gestion de ces triplets est délicate et toute erreur se répercutera immédiatement sur l'intégration des données.

Il est évident qu'un solution plus optimale est envisageable. L'enjeu est donc le suivant : construire un système hors-ligne permettant d'une part l'ajout, la modification et la suppression de données et d'autre part la constitution d'un fichier Excel d'importation de données.

2.2. Proposition

Une solution envisageable est la constitution d'une base de données hors-ligne qui pourrait en sortie produire le fichier Excel nécessaire à l'intégration dans la base ORACLE (ou tout au moins les fichiers txt cités plus haut)(figure 2).

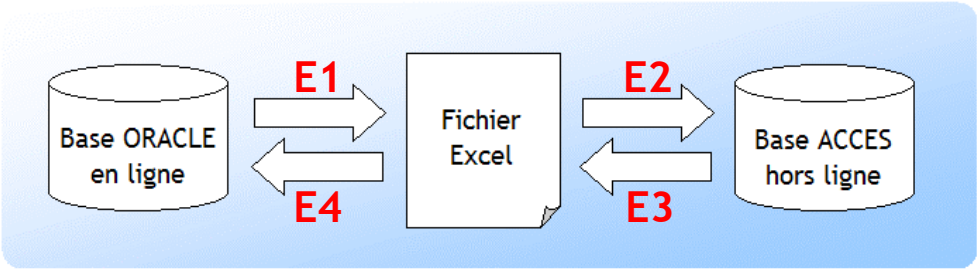


Figure 2 : échanges entre la base en ligne et la base hors-ligne

2.2.1. Avantages

Le gros avantage de cette solution est le suivant : un confort non négligeable pour les producteurs de données. En effet, la gestion des identifiants d'enregistrements devient transparente pour l'utilisateur.

2.2.2. Inconvénients

Toute évolution de la structure de la base en ligne implique, en plus des modifications nécessaires dans les procédures d'importation/exportation, des transformations similaires pour la base hors-ligne.