

Synthèse Système d'information I

Volet base de données

1. Introduction

Le volet base de données est transversal aux différentes études menées dans le cadre du projet GRIPAVI. Il s'articule donc autour des 6 observatoires. L'objectif principal est de disposer d'un outil commun de centralisation et restitution des données afin de pouvoir réaliser des études comparatives entre observatoires.

Le volet faune sauvage (et interactions avec le domestique au Zimbabwe et au Mali), le volet socio-économique (ou filière) au Vietnam, à Madagascar et prochainement en Éthiopie, ainsi que les volets sérologiques et virologiques doivent être intégrés dans la base. Chaque étude traite une problématique particulière, ce qui engendre des données hétérogènes qui doivent être mises en relation. De plus, la base de donnée doit être un outil collaboratif, c'est-à-dire permettant l'échange et la réutilisation du travail d'enquête fait sur le terrain.

Le travail de réflexion a commencé en janvier 2008 et a été mené par Véronique Chevalier et Raphaël Duboz en relation avec les différents partenaires des autres volets.

2. Moyens mis en place

Ces premiers mois de réflexions ont nécessité la mobilisation d'un licencié en informatique pour une durée de 5 mois et demi de mars à mi-août (C. Visseq). Il a été chargé de développer un prototype de base de donnée pour nous permettre, en interaction avec les partenaires sur le terrain, de définir un modèle conceptuel des données adapté à la problématique et d'implémenter une base de données relationnelle accessible par internet pour valider les choix techniques.

3. Activités 2008

Harmonisation des questionnaires

A partir de questionnaires filière et socio-eco existants, un questionnaire relativement générique a été réalisé. Ce questionnaire peut servir de base pour les nouvelles enquêtes. Le volet faune sauvage conserve son format de donnée qui est relativement stable et partagé.

Définition des besoins fonctionnels

Au début de l'année 2008, nous n'avions pas une visibilité claire des variables récoltées. Il s'avère qu'en fait l'ensemble des variables ne peut pas être totalement connu *a priori*. En effet, les variables reflètent les questions de recherches qui évoluent tout au long de l'étude.

Conception d'une base de données relationnelle accessible par internet

La base de données ne pouvait donc pas être une base de données classique où les différentes tables et les différents champs sont définis à partir des variables et de leur valeur. Il a fallu concevoir une base qui permette la création de nouveaux questionnaires et de nouvelles variables en cours d'étude. Afin de tester la faisabilité d'une telle base et de mieux cerner les besoins du projet, nous avons demandé à C. Visseq de définir un cahier des charges et de développer un prototype.

L'idée de base du site est de proposer une plate-forme de stockage des données collectées lors des enquêtes terrains effectuées par les membres de l'unité de recherche. La plate-forme doit à la fois proposer une interface de stockage et de restitution des données, une interface de création de questionnaires et de saisie des fiches. L'analyse des données étant spécifique, on propose des mécanismes de restitution des données permettant un post-traitement avec des logiciels plus appropriés.

Une des premières demandes a donc été de développer une interface permettant de faciliter la création de questionnaires. L'interface doit permettre de créer de nouveaux questionnaires, mais aussi de modifier et d'adapter des questionnaires existants. Elle doit proposer une liste de questions existantes, tout comme permettre la création de nouvelles. Chaque question renseigne une variable. Une liste de variables est proposée par défaut, mais de nouvelles doivent pouvoir être ajoutées si nécessaire. L'idée est que, arrivé à un certain temps d'utilisation, la grande majorité des variables nécessaires et des questions existantes auront été créées, si bien que la conception de nouveaux questionnaires consistera simplement à piocher parmi une liste de questions existantes celles que l'on souhaite utiliser. La création de questionnaire sera donc simplifiée au fur et à mesure que la base est utilisée.

Au niveau des données, deux aspects doivent être intégrés au site : la saisie de donnée à partir de fiches papiers et la restitution de données. La saisie se base sur un questionnaire précédemment créé, elle doit être la plus aisée possible mais aussi la plus exacte (contrôle de saisie pour éviter les erreurs). La restitution, elle, doit à la fois permettre de récupérer les données saisies, mais aussi de les sélectionner selon divers critères (géographiques, temporels...). Les données doivent être restituées de manière à pouvoir être utilisable dans d'autres logiciels, tel que Excel ou Access, que les chercheurs utilisent habituellement.

L'intérêt de la base est de permettre une centralisation des données, une saisie et un partage facilité mais aussi une analyse transversale plus aisée. De par l'utilisation d'un ensemble de variables limités, il devient aisé de recouper les enquêtes et d'étendre l'étude grâce à des données saisies lors d'enquêtes passées. De plus, toutes les fiches doivent être localisées dans l'espace (point GPS) et dans le temps, pour permettre un tri sur ces critères et pouvoir suivre à la fois les évolutions et les échanges entre les lieux étudiés.

Environnement d'utilisation

4. Résultats

Le modèle conceptuel des données correspond tout à fait aux attentes en permettant de disposer d'une base où de nouvelles variables (et donc de nouveaux type de données) peuvent être enregistrés. La notion de variables partagées permet de mettre en relation des enquêtes et missions de terrain différentes et donc d'assurer le rôle de transversalité entre observatoire et questions de recherche.

La base de donnée relationnelle est fonctionnelle (<http://avianbd.cirad.fr/>) pour ce qui est de la création de nouveaux questionnaires et de nouvelles variables, de la réutilisation de ceux-ci et de l'enregistrement des données. Ces fonctionnalités ont été testées avec succès à partir des enquêtes menées à Madagascar. Néanmoins, la notion classique de requête se complique sur un tel type de base du fait de l'apparition de nouvelles variables. Il reste donc un travail important à faire sur la restitution de données en vue des analyses.

L'interface web n'est pas assez souple pour le travail terrain et complique l'interopérabilité avec les outils standards de saisie et d'analyse des données. Cette limitation technologique n'a pas pu être anticipée avant le développement du prototype.

Les besoins fonctionnels pour aller vers un système d'information performant ont été clairement définis. L'interface web (utilisation d'un navigateur) limite les fonctionnalités de saisie et restitution. Il serait plus approprié de développer un logiciel dédié à installer sur chaque poste utilisateur. Ce logiciel devrait fonctionner avec et sans connexion internet pour permettre une saisie directe sur le terrain et une mise à jour de la base commune une fois connecté. Conceptuellement, il serait intéressant de se tourner vers les outils de fouilles de données (ou *data mining*), à la croisée des statistiques et des bases de données.

5. Commentaires et recommandations :

D'un point de vue pratique, la base de données n'étant pas 100% fonctionnelle, il a été décidé que chaque partenaire sur le terrain renseigne sa propre base. Une semaine de travail de mise en commun des données et d'identification des questions transversales qui peuvent être traitées au regard des données recueillies aura lieu à la suite des journées de septembre du CIRAD. Ainsi l'ensemble des participants expatriés et métropolitains sera présent. Ce travail en commun nous semble incontournable pour :

- éviter de mauvaises interprétations des bases construites par chacun
- obtenir des informations périphériques aux données et non enregistrées
- définir ensemble les questions transversales auxquelles nous pouvons répondre
- co-construire une méta-base commune renseignée par des données pré-traitées pour répondre aux questions transversales.

Le travail de cette première année a également permis de définir les verrous conceptuels et technologiques et permettrait à terme de développer un outil commun de fouille de données. Cette solution nécessite des collaborations avec des spécialistes du domaine sur une durée d'un an environs. Ce travail n'a pas été planifié pour 2009 et pourra l'être en 2010.