Projet international GBIF 2004-2008

Etat de situation au 20 février 2006

F. Burri, M. Bouzelboudjen

Service Informatique et télématique
Université de Neuchâtel

1. Introduction

L'Université de Neuchâtel réalise le noeud informatique de GBIF Suisse et met à disposition des chercheurs, étudiants et du public un réseau de compétences et de connaissances dans le domaine de la biodiversité. Ce noeud informatique permettra de développer des bases de données accessibles via Internet dans les domaines de la faune principalement et de la flore (spécimens des collections scientifiques des Musées de Suisse).

2. Travaux réalisés durant l'année 2004

L'architecture informatique proposée est décrite dans l'exposé du 28 janvier 2004 présenté à Berne. Le noeud informatique du « participant node » suisse est déployé sur une architecture multi-tiers. Cette architecture est composée d'un serveur d'application dans lequel sont installés les logiciels spécifiques au projet GBIF. Le serveur d'application constitue l'interlocuteur immédiat des internautes. Le serveur de base de données s'appuie sur une architecture multi-processeurs connecté à une unité de stockage de masse de type SAN. Les données seront gérées par le SGBD Oracle. Le serveur de base de données concrétise le concept de « data nodes » centralisés. Il contient les données de toutes les institutions qui ont choisi cette solution pour leur intégration dans le cadre du projet GBIF Suisse. C'est ainsi que deux alternatives ont été proposées aux institutions qui souhaitent adhérer à la logique du projet International GBIF:

- Intégration de leurs données par le noeud centralisé (serveurs GBIF-CH) ;
- Intégration de leurs données dans un serveur propre à l'institution.

Dans sa version de test, la base de données Oracle 9i et l'application cohabitent sur le même serveur. Ce dernier est équipé d'un processeur de type Xeon sous Linux/Redhat. Il dispose d'une capacité de stockage de 74 Gb et d'une mémoire RAM de 4 Gb.

En septembre 2004, la solution décrite a été réalisée et a été présentée à Berne. Cette architecture s'articulait autour des logiciels standards fournis par GBIF International permettant l'intégration du format Darwin Core.

En résumé, les tâches suivantes ont ainsi été réalisées :

- Achat et installation d'un serveur sous Linux ;
- Installation d'une base Oracle 9i ;
- Intégration de données de test issues du musée d'histoire naturelle de Neuchâtel ;
- Installation des packages standards GBIF (Digir Provider, GBIF Portal Toolkit, GBIF Data repository Tools, DiGIR Portal) ;
- Intégration de données test au format texte et MS Excel ®.

Depuis décembre 2004, le comité scientifique GBIF Suisse a décidé d'intégrer dans la base de donnéées, des données de type image. Cette décision est importante. Ainsi, le format de données initial XML Darwin-Core est abandonné au profit du format ABCD (projet BIOCASE)

Les conséquences de cette décision sont :

- Abandon des outils DiGiR ;
- Abandon des applications liées à DIGIR réalisées par le noeud informatique GBIF-CH à l'Université de Neuchâtel ;
- Installation des nouveaux outils BioCAse dans le serveur de test Norma.

3. Travaux réalisés durant l'année 2005

Le portail permettant la publication des informations en rapport avec GBIF Suisse (http://www.gbif.ch) a été ouvert le 31 mai 2005. Ce site a été développé avec l'environnent Jahia. (http://www.jahia.org). Jahia est un CMS (Content Management System) ou système de gestion des contenus.

Pour la mise à disposition des données suisses de biodiversité accessibles via Internet, les actions suivantes ont été réalisées :

- Consolidation de l'architecture de la base de données du noeud Suisse
- Intégration de données tests
- Installation et configuration des logiciels nécessaires (PyWrapper) à la publication de ces données au format ABCD de BioCASE
- Enregistrement des sources de données auprès de GBIF International

La solution informatique basée sur BioCASE a été présentée à la commission scientifique, le 31 août 2005 à Berne.

En septembre 2005, un concept de backup Oracle ® de la base de données test GBIF-CH a été élaboré et mis en œuvre. Un document présentant la solution a été mis à disposition.

En octobre 2005, un document décrivant la procédure de transmission des flux d'information au nœud informatique central suisse (GBIF-CH) a été rédigé et mis à disposition. Ce document détaille les trois volets suivants :

- Information descriptive de l'institution ;
- Information descriptive de chacune des bases de données fournies par l'institution ;
- Données de collection aux formats MS Access ® ou MS Excel ®.

De plus, ce document précise le rôle du coordinateur GBIF. Ce dernier est le correspondant entre les institutions et le nœud informatique GBIF Suisse. Il doit être informé de l'existence des données et assure la « viabilisation » de ces données pour les nœuds GBIF Suisses.

4. Travaux en cours et prévus pour l'année 2006

L'analyse résultante de la problématique de l'intégration de données de plusieurs projets nous a conduits à développer des applications informatiques. De plus, les données provenant de diverses régions linguistiques nécessitent un « traitement » préalable pour faciliter la communication.

Ces applications informatiques sont basées sur un système de normalisation et de codification multi langues. Ainsi, sur les 46 champs retenus pour stocker les données de collection, 14 font l'objet de codes. Les fournisseurs de données n'ont pas de surcharge de travail lié à cette fonctionnalité. La codification des données est automatique et s'appuie sur un thésaurus. Pour les codes faisant partie d'une hiérarchie, des algorithmes complexes ont été mis en œuvre et permettent de compléter automatiquement les valeurs lacunaires de l'ensemble des niveaux supérieurs. C'est le cas par exemple pour la systématique, la stratigraphie, la géographie, etc. Un autre avantage du système consiste à permettre d'attribuer une valeur définie dans une langue de référence à toutes les données qui sont

codées. Compte tenu du caractère international du projet, cette fonctionnalité est essentielle pour la diffusion de l'information via Internet.

5. Travaux prévus en 2007-2008

Dans le contexte suisse où quatre langues nationales cohabitent, la mise à disposition de l'information à l'ensemble de la population passe par une offre d'information dans la langue du citoyen. Le portail GBIF- International ne fourni pas ce genre de fonctionnalités. Les informations n'y sont interrogeables que dans une seule langue.

Aussi, nous proposons de réaliser un portail GBIF- Suisse offrant aux Internautes Suisse la possibilité de réaliser des interrogations dans leur langue maternelle. Il est important de relever que la base de données GBIF-CH est déjà construite sur une architecture multi-langues. Un tel développement ne nécessite ainsi aucune modification de la structure de la base de données. Il s'inscrirait dans un contexte visant à tirer mieux parti de la qualité de l'information mise à disposition par le nœud informatique Suisse.

Durant ces deux prochaines années, les données des projets seront intégrées dans la structure mis en œuvre. Avec l'intégration de ces informations, des vérifications des modèles proposés par le nœud informatique suisse seront réalisées et permettront d'améliorer la gestion et la visualisation des données via Internet.

Nous pouvons citer succinctement quelques aspects liés à cette problématique de vérification automatique des règles prédéfinies dans les modèles :

- Vérifications automatiques de règles prédéfinies ;
- Colonne obligatoirement renseignée ;
- Valeur appartenant à un thésaurus ;
- Unicité ;
- Normalisation automatique de valeurs dans les colonnes de codes ;
- Production d'un journal qui peut être utilisé par l'institution qui a transmis les données.

Neuchâtel, le 20 février 2006

NB :
Les documents cités dans le texte sont disponibles à l'adresse : http://www.gbif.ch/