Projet Papillon : Architecture du serveur Web

Mathieu MANGEOT-LEREBOURS, Gilles SÉRASSET
GETA-CLIPS-IMAG BP53, Campus universitaire,
385 avenue de la bibliothèque 38041 Grenoble, France
Tel. : +33 4 76 51 43 80
Fax. : +33 4 76 51 44 86
E-mail: Mathieu.Mangeot@imag.fr

Résumé
Le projet Papillon est un projet de développement coopératif à travers Internet d'une base lexicale multilingue. Ce projet a été lancé en janvier 2000 par une coopération entre le GETA-CLIPS et le National Institute of Informatics (NII) japonais avec le support actif de l'Ambassade de France à Tokyo. Les langues actuellement traitées sont le français, le japonais, l'anglais, le thaï et le lao. Ce projet a pour but d'appliquer le paradigme de développement de LINUX à la construction d'une base lexicale : n'importe qui peut contribuer bénévolement selon ses compétences et en retour, toutes les données de la base sont disponibles avec une licence de logiciel libre (OpenSource).
Cet article montre plus précisément l'architecture générale du serveur Web utilisé dans le projet puis l'interface de consultation. Les articles sont encodés en XML.
Les outils utilisés dans ce projet sont tous disponibles avec une licence de logiciel libre (OpenSource). L'ouverture au public du serveur est prévue pour la fin de l'année 2001.

Abstract
The Papillon project is a cooperative building project on the Internet of a multilingual lexical database. This project started in January 2000 as a cooperation between the GETA-CLIPS laboratory in Grenoble, France and the Japanese National Institute of Informatics (NII) with the active support of the French embassy in Tokyo. The languages involved in the project are French, Japanese, English, Thai and Lao. This project aims to apply the LINUX development paradigm to the building of a lexical database: anyone can contribute freely according to his/her competences and all the data are available with an OpenSource license.
This article shows more precisely the general architecture of the webserver used in the project and the interface for consultation. The entries are encoded in XML.
All the tools used in this project are available with an Open Source license. We plan to open te server to the public by the end of the year 2001.

Introduction
Le projet Papillon [3,b4] est un projet de développement coopératif à travers Internet d'une base lexicale multilingue. Ce projet a été lancé en janvier 2000 par une coopération entre le GETA-CLIPS et le National Institute of Informatics (NII) japonais avec le support actif de l'Ambassade de France à Tokyo. Depuis, des partenaires travaillant sur le thaï (Kasetsart University & NECTEC à Bangkok) et le lao se sont joints au projet. Les langues actuellement traitées sont le français, le japonais, l'anglais, le thaï et le lao. À court terme, il est prévu d'inclure le vietnamien et le malais.
Ce projet a fait suite à un constat de manque de ressources lexicales gratuites (par exemple entre le français et le japonais) et du manque d'informations importantes dans les ressources existantes (difficile de trouver à la fois le kanji, les kanas et le romaji pour un article japonais).
Il a pour but d'appliquer le paradigme de développement de LINUX à la construction d'une base lexicale : n'importe qui peut contribuer bénévolement selon ses compétences et en retour, toutes les données de la base sont disponibles avec une licence de logiciel libre (OpenSource).
L'intérêt scientifique principal du projet est de construire une base lexicale fondée sur une architecture pivot : les articles des dictionnaires monolingues sont reliés entre eux par des liens interlingues ou acceptions formant le dictionnaire pivot.
Dans cet article, nous nous intéressons plus précisément à l'architecture du serveur Web.
Nous montrons tout d'abord l'architecture générale du serveur avec sa partie consultation puis nous détaillons la structure d'un article monolingue avec différentes vues qu'un utilisateur peut obtenir en consultant la base.

1. Architecture générale du serveur

Voici la page d'accueil du serveur Papillon :

Le menu situé dans la partie gauche de la page ainsi que la barre horizontale des sections située en haut de la page constituent la partie fixe du serveur. Lorsque les utilisateurs se connectent pour la première fois, ils doivent s'enregistrer en cliquant sur le menu de gauche. Lors des sessions suivantes, ils s'identifient de la même manière. Ensuite, les utilisateurs choisissent une section dans la barre horizontale.

La section "Informations" contient l'archivage des listes de distribution du projet Papillon ainsi que tous les documents entreposés par les utilisateurs concernant le projet. Les utilisateurs peuvent, à l'aide d'une interface, télécharger sur le serveur leurs documents.

Le projet Papillon s'adresse à des utilisateurs parlant différentes langues et donc susceptibles d'utiliser différents encodages. Pour que tous les utilisateurs puissent accéder aux informations, celles-ci sont convertis de leur encodage d'origine vers l'UTF-8 puis stockées sur le serveur.

La section "Consultation" permet de consulter les dictionnaires de la base Papillon.

La section "Édition" permet de rédiger de nouveaux articles ou de contribuer localement sur des articles.

2. Gestion des données lexicales

Le seul codage couvrant toutes les langues du projet Papillon est Unicode. Nous avons donc choisi ce codage avec l'encodage UTF-8 pour stocker toutes les données de la base Papillon. Les données lexicales sont stockées en UTF-8 au format XML pour faciliter la lecture, la manipulation et l'export vers d'autres formats cibles.

Lors de la consultation, les utilisateurs peuvent annoter ou corriger les données. Ces annotations et ces corrections sont stockées dans la base sous forme de feuilles de style XSL. Elles sont ensuite appliquées dynamiquement sur les données d'origine lors des requêtes ultérieures.

3. Interface de consultation

L'utilisateur peut effectuer des requêtes dans la base lexicale en accédant à la section "Consultation". Il peut effectuer des requêtes simples comme la recherche de l'article "abat-jour" mais aussi des requêtes beaucoup plus complexes en fonction de critères variés dépendant des informations contenues dans la base et aussi de l'architecture pivot multilingue de la base lexicale [xxxJST00]. Il peut par exemple chercher les traductions japonaises des articles dont la catégorie grammaticale est un nom et dont la fonction lexicale "S1" contient "auteur".

4 Outils utilisés pour l'implémentation`

Les données de la base Papillon sont toutes stockées dans une base de données relationelle classique. Nous avons choisi PosgreSQL [b3] pour sa capacité à stocker des documents encodés en Unicode UTF-8. Les listes de distributions sont archivées avec l'outil MHonArc [b2] modifié pour l'occasion afin de convertir les documents en UTF-8. Le serveur Web utilisé est le serveur dynamique Enhydra [b1] écrit en Java. Il s'interface facilement avec la base de données via un pilote JDBC. Les outils implémentant le DOM (Document Object Model) utilisés pour manipuler les données XML et XSL sont l'analyseur Xerces [b6] et le moteur XSLT Xalan [b7] en Java dont les fonctions sont directement utilisables avec Enhydra. Tous les outils utilisés dans le projet sont disponibles avec une licence de logiciel libre (OpenSource).

Conclusion

La première année du projet Papillon a servi principalement à la définition de l'architecture et des structures de données utilisées dans le projet ainsi qu'à la politique d'ouverture et de contribution. Nous avons malgré tout avancé les développements portant sur le serveur. Ceux-ci sont complexifiés par la décision de créer un site totalement multilingue et multiutilisateurs. Nous envisageons cependant l'ouverture officielle du serveur avant la fin de l'année 2001.

Références

[1] **Igor A. Mel'tchuk (1997)** Vers une linguistique Sens-Texte. Leçon inaugurale, Collège de France, Chaire internationale, 43 pages. http://www.fas.umontreal.ca/LING/olst/FrEng/melcukColldeFr.pdf

[2] **Alain Polguère (1998)** La théorie Sens-Texte. Dialangue, Vol. 8-9, Université du Québec à Chicoutimi, pp. 9-30. http://www.fas.umontreal.ca/LING/olst/FrEng/PolgIntroTST.pdf

[3] **Mutsuko Tomokiyo, Mathieu Mangeot & Emmanuel Planas (2000)** *Papillon : a Project of Lexical Database for English, French and Japanese, using Interlingual Links.* Journées Science et Technologie de l'ambassade de France au Japon, 13 Novembre 2000, Tokyo, Japon, 3 p.

Signets

[b1] **Enhydra** Serveur Web dynamique java : http://www.enhydra.org/
[b2] **MHonArc** convertisseur mel vers HTML : http://www.mhonarc.org/
[b3] **PostgreSQL** SGBD : http://www.postgresql.org/
[b4] **Projet Papillon** : http://vulab.ias.unu.edu/papillon/
[b5] **Xalan** moteur XSLT : http://xml.apache.org/xalan-j/
[b6] **Xerces** analyseur DOM/XML : http://www.apache.org/xerces-j/