

Ecole Française d'Athènes
Andréa Iacovella
6, rue Didotou
10681 Athènes

Juin - Septembre 2001
Jocelyn Viallon

RAPPORT DE STAGE

Les Processus d'importation dans porphyre

INTRODUCTION	2
ORGANISATION DU STAGE	2
PRESENTATION DU SOUS PROJET	2
REPARTITION DES TACHES	2
RAPPORT TECHNIQUE	2
FORMAT DE DOCUMENT	2
<i>Représenter un graphe acyclique en XML</i>	3
ANALYSE DES DOCUMENTS	3
STOCKAGE DES CONTENUS	4
<i>Client ftp</i>	4
<i>Serveur ftp</i>	5
TRADUCTION DE FORMAT	5
CONCLUSION	6

Introduction

Ce document présente le travail effectué durant mon stage à l'Ecole Française d'Athènes du 5 juin au 3 septembre 2001, sous la direction de M. Andréa IACOVELLA et de M. Aurélien BENEL . Ce travail s'intègre dans le projet Porphyre relié à la numérisation de la chronique des fouilles. Nous présenterons d'un aspect technique les problèmes posés dans le cadre de ce stage, nous discuterons les solutions qui ont été retenues et les améliorations possibles pour le développement du projet.

Organisation du stage

Présentation du sous projet

Le projet Porphyre permet d'indexer des publications et documents scientifiques, et de les publier à la communauté par l'intermédiaire du réseau internet. Ces documents sont publiés sous une forme qui lui est propre et qui ne peut être analysée directement par le logiciel d'indexation. Certains documents sont disponibles sous forme papier et subissent un travail de numérisation, ou de reconnaissance de caractère. D'autres peuvent déjà se présenter sous forme numérique mais dans des formats variés, par exemple TEI . Un besoin d'exportation de documents et de structures pour la publication ou des traitements extérieur est aussi apparu. Il était donc nécessaire de disposer d'un outil permettant de partir des données existantes, afin de les importer dans l'arbre d'indexation et de l'intégrer au logiciel.

Répartition des tâches

Le travail s'est décomposé en trois parties distinctes permettant d'intégrer ce mécanisme d'importation dans le cadre déjà existant :

- La création d'une DTD XML permettant la définition d'un format de fichier unique qui pourrait être analysé sans problème et ajouté au graphe existant. Cette DTD servira aussi de structure pour l'exportation de documents ou de sous graphes.
- L'intégration dans le prototype existant d'un outil d'importation, afin d'ajouter des documents, fournis dans le format XML Porphyre, de manière efficace.
- La réalisation d'un outil de transformation de données brutes dans le format de porphyre. Dans ce cas là l'étude s'est limitée au cas de documents papiers numérisés par pages, structurés dans une base de données File Maker Pro.

Rapport technique

Format de document

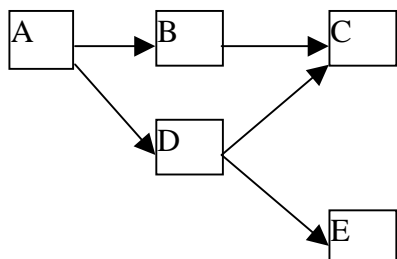
Face à la multitude de formats de documents existants devant être en mesure d'être importés dans Porphyre, il a été décidé de créer un format de référence permettant de représenter un graphe acyclique et par conséquent n'importe que graphe ou sous graphe de porphyre. Ce format devait répondre aux exigences de portabilité et d'ouverture. Le format retenu a été le XML qui permet de définir un sous format répondant à tous ces critères. De plus c'est un format ouvert, développé par un consortium ce qui lui vaut une certaine stabilité et maturité.

Ce format permet de créer une grammaire (DTD) qui va définir un type de document. La validité d'un document à cette DTD nous certifie que le document pourra être traité correctement sans erreur structurelle.

Représenter un graphe acyclique en XML

Construire une grammaire pouvant décrire un sous graphe de porphyre revient à construire une grammaire pour représenter un graphe acyclique. Les feuilles de ce graphe sont les différents types de données qui sont stockées i.e. des fichiers seuls ou des fragments de documents textes.

Le format XML nous permet de représenter des structures linéaires. Prenons l'exemple suivant :



L'élément 'C' se trouve pointé par deux parties du graphe, par la branche partant de 'B' et celle partant de 'D'. On peut trouver deux façons de modéliser cela de manière à le représenter linéairement :

```

<A>
  <B>
    <C/>
  </B>
  <D>
    <C/>
    <E/>
  </D>
</A>
  
```

solution 1

```

<A>
  <B>
    <C/>
  </B>
  <D>
    <@C/>
    <E/>
  </D>
</A>
  
```

solution 2

La première solution a l'avantage d'être complètement indépendante du sens de parcours du graphe, un document de ce type donnerait donc moins de contraintes à sa création. Mais la deuxième solution apparaît la meilleure car elle évite toute redondance d'information au niveau de 'C'. Seule l'adresse de 'C' est utilisée quand plus d'une branche pointe sur le même nœud. La notion de référence est présente dans la norme XML .

Analyse des documents

Partant d'un document valide qui respecte la grammaire définie précédemment, il s'agit alors de parcourir ce document pour en extraire ses données. La structure du logiciel porphyre nous fournit des fonctions de haut niveau permettant de créer une structure de graphe par des branches et des feuilles. Pour l'analyse d'un document XML, plusieurs API ayant des implémentations java sont disponibles. Deux sont normalisées et retiendront notre attention : DOM (Document Object Model) et SAX (Simple API for XML).

DOM est une API qui construit un arbre en mémoire à partir du document source. Cela offre l'avantage de pouvoir accéder de façon aléatoire à toutes les informations du document.

Au contraire SAX fonctionne sous la forme d'évènements qui sont déclenchés pendant le parcours unique et linéaire du document.

Dans notre cas, le parcours du document n'a besoin d'être effectué qu'une fois et dans l'ordre de création des balises XML. De plus il faut tenir compte que les documents analysés peuvent être assez volumineux, car ils peuvent regrouper des publications de plusieurs centaines de pages. Nous retiendrons donc l'implémentation SAX qui offre une meilleure rapidité et une occupation mémoire moins importante. Cela nous oblige par contre à stocker les informations sur les nœuds référencés. En effet, il faut noter que rien ne spécifie qu'un appel à un objet référencé implique que cet objet soit préalablement déclaré. Toutes les références seront donc résolues après avoir construit le graphes principal. Une contrainte persiste dans la création de documents valides car pour la vérification de la conformité à la DTD, il est indispensable que cette DTD soit référencée par une balise SGML dans l'entête du document, et que le fichier de DTD soit donc présent avec le fichier XML porphyre. La DTD étant fixe, il serait envisageable qu'elle soit stockée avec le logiciel.

Stockage des contenus

Différents types de contenus peuvent vouloir être stockés dans Porphyre : des bribes de texte brut, des fichiers en local, des fichiers externes, des références à des adresses Internet, etc. Dans tous ces cas pour obtenir une homogénéité il a été convenu de retrouver ces contenus quels qu'ils soient par l'intermédiaire d'une URL. Dans le cas d'une référence externe à une adresse Internet il ne se pose aucun problème. Mais dans le cas de contenu en local ou de fragments de textes, il a été retenu la solution de passer par un serveur WEB et une servlet qui renverra le contenu à partir d'une adresse Internet.

Un fragment de texte peut être représenté en XML soit par une section de texte de type CDATA incluant directement le fragment inclus dans une balise, soit par une référence à un fichier contenant ce même texte. Dans le cas de notre étude, et son application à la chronique des fouilles il se peut que nous soyons confrontés à l'utilisation de caractères requérant un encodage spécial. L'encodage UNICODE faisant parti intégrante du langage java, l'utilisation de texte inscrit dans du XML permet de définir facilement le type d'encodage utilisé, en le déclarant dans les en-têtes XML. Cela permet de déléguer le problème de l'encodage des caractères à un stade antérieur. Des fichiers pourront être recréés par la suite au passage du document.

A chaque fois que l'on se trouve confronté à un document présent en local sur l'ordinateur d'où se déroule le processus d'importation, ils doivent être transférés en local sur le système de fichiers du serveur de documents. Pour cela on utilisera un serveur ftp indépendant du serveur web existant. Les fichiers en local sur le serveur pourront ensuite être renvoyés par le serveur web et la servlet. La solution de transfert de données par RMI ne peut pas s'appliquer au cas de fichiers de taille importante.

Le transfert de documents par ftp nécessite donc un client ftp et un serveur ftp.

Client ftp

Lors de l'analyse pour l'importation d'un document, la structure est importée par l'appel de méthodes RMI. Par souci de facilité pour l'utilisateur, un client ftp léger a été intégré au module d'importation. La seule implémentation libre d'un client ftp léger qui nous a été disponible et qui a par conséquent été utilisée, est un projet sourceforge appelé "finj", en licence LGPL, développé par jiglesias. Il permet les fonctions de base de transfert ftp :

connexion et envoi de fichier. Dans le cas où le réseau est protégé par un firewall, la connexion ftp dans un mode client normal est impossible, sans ouvrir un port spécial. Pour supprimer cette contrainte, le poste client pouvant se trouver ou non derrière une protection de type pare feu, il a été choisi par défaut d'utiliser le mode ftp passif. Par souci de simplicité, le choix n'est pas donné à l'utilisateur, le mode passif fonctionnant dans les deux cas.

Serveur ftp

Pour la configuration du côté serveur, il est nécessaire d'installer un démon ftp afin de recevoir les fichiers et de les stocker en local dans un répertoire. Il pourra être utilisé n'importe quel serveur ftp disponible, pourvu que celui ci soit configuré pour que l'utilisateur ftp qui s'y connecte dispose d'un accès en écriture.

Un nom unique est créé pour les fichiers à envoyer sur le serveur. Ceux-ci sont stockés directement dans le répertoire racine de l'utilisateur. On pourrait améliorer ce mécanisme et permettre une arborescence supplémentaire, soit en créant une hiérarchie par processus d'importation, soit en demandant à l'utilisateur d'organiser lui même ses documents. Le renommage des fichiers permet une transparence entre les fichiers créés et les fichiers fournis. L'extension des fichiers est gardée ainsi il est plus facile de déterminer le type de fichier présent.

De plus, les fichiers ainsi stockés doivent être accessibles par un serveur web, à partir d'une adresse de base et du nom du fichier. Par exemple si le fichier article13.html est transféré sur le serveur ftp. La configuration du serveur web doit être telle qu'à partir d'une adresse de base on puisse retrouver le fichier de la façon suivante : si

<http://serveur.efa.gr/documents/> est l'adresse de base, l'url du fichier est donc :

<http://serveur.efa.gr/documents/article13.html> .

Traduction de format

Les données brutes qui sont destinées à être importées dans porphyre peuvent être présentes sous différentes formes, numériques ou non.

Dans le cas de données numériques, on trouve dans de nombreux cas un formatage de données XML ou traductible en XML. Bien sûr le type de document peut être complètement différent de la DTD Porphyre. La transformation de ce type de document en XML porphyre peut se faire facilement grâce aux mécanisme de feuilles de style de transformation (XSLT). On peut définir pour chaque format XML une feuille de style qui donnera les mécanismes permettant de transformer tout document du premier format au deuxième format. Cela pourra se faire en perdant des informations inutiles au format de Porphyre.

Mais il existe un grand nombre d'archives qui ne sont présentes que sous forme papier. Un moyen d'intégrer ces documents est de numériser les fragments de documents, le plus souvent les pages, en utilisant une base de données qui décrit la structure des documents.

En ce qui concerne ce projet, nous avons étudié le cas de documents scientifiques, revues et monographies, qui ont été préalablement scannés page par page. Leur structure a été rangée dans une base de données File Maker Pro, de la façon suivante :

Table Documents						
Id	ISBN	Titre	Auteur	Type	Revue	N°

Table Pages				
Id	Id-doc	Type	N°	URL

Cela donne une idée de l'intégration de documents qui pourrait être appliquée dans le cas de nombreuses données existantes qui ne sont pas sous forme numérique.

En ce qui concerne notre exemple, il sera adaptable en fonction de la structure véritable de la base de données et de l'indexation de fichiers que l'on souhaite réaliser. Le script java permet de générer un fichier dans le format de porphyre qui indexe les fragments de documents dans une structure physique et une structure logique, classés par ISBN, par Auteur et par Titre.

Conclusion

Cette étude a permis de définir une démarche d'importation de documents, en proposant une ouverture du format de données interne du projet sur les nombreux formats disponibles et utilisés par la communauté scientifique. Cela est possible par une simple traduction des documents qui, la plus part du temps peut être facilement automatisée mais qui reste spécifique à chaque format. Il a aussi été rendu possible d'envisager des processus réciproques d'exportation de données. Des approfondissements pourront être envisagés par la suite, en suivant le déroulement global du projet, telle la gestion de documents multilingues et des jeux de caractères.