Enrichissement sémantique de documents XML représentant des tableaux

Fatiha Saïs[1], Hélène Gagliardi[1]
Ollivier Haemmerlé[1,2], Nathalie Pernelle[1]

[1]LRI (UMR CNRS 8623 - Université Paris-Sud) / INRIA (Futurs), Bâtiment 490,
F-91405 Orsay Cedex, France
{prénom.nom}@lri.fr
http ://www.lri.fr/iasi
[2]UMR INAP-G/INRA BIA, 16 rue Claude Bernard,
F-75231 Paris Cedex 05, France
Ollivier.Haemmerle@inapg.fr
http ://www.inapg.inra.fr/ens_rech/mathinfo/index.html

Résumé. Ce travail a pour objectif la construction automatique d'un entrepôt thématique de données, à partir de documents de format divers provenant du Web. L'exploitation de cet entrepôt est assurée par un moteur d'interrogation fondé sur une ontologie. Notre attention porte plus précisément sur les tableaux extraits de ces documents et convertis au format XML, aux tags exclusivement syntaxiques. Cet article présente la transformation de ces tableaux, sous forme XML, en un formalisme enrichi sémantiquement dont la plupart des tags et des valeurs sont des termes construits à partir de l'ontologie.

Mots-clés : extraction de connaissances, entrepôt, ontologie, XML, Web.

1 Introduction

Le travail que nous présentons dans cet article est mené, en collaboration avec quatre partenaires[1], dans le cadre du projet e.dot (Entrepôt de Données Ouvert sur la Toile). Ce projet vise à permettre la construction automatique d'entrepôts thématiques de données stockées au format XML, alimentés par des données extraites du Web. Le domaine d'application choisi est la prévention du risque microbiologique (listeria, salmonelle, etc.) dans les aliments. Ce domaine présente un enjeu de santé publique mais également un enjeu industriel majeur. Notre entrepôt de données XML permet de compléter une base de données relationnelle et une base de graphes conceptuels préexistantes, contenant des données scientifiques et industrielles. Ces deux bases, interrogées de manière uniforme par le système MIEL (Buche et al.2004), ont été développées dans le cadre du projet Sym'Previus (Sym'Previus 1999) qui vise à construire un outil de prévision du comportement des germes pathogènes dans les aliments. La version étendue de MIEL interrogeant les deux bases existantes et l'entrepôt XML s'appelle MIEL++.

[1]IASI-Gemo (LRI), Verso-Gemo (INRIA-Futurs), INAP-G/INRA et la société Xyleme