Conception et construction d'entrepôts en XML

Omar Boussaid*, Riadh Ben Messaoud‡, Rémy Choquet†, Stéphane Anthoard†

Laboratoire ERIC – Université Lumière Lyon 2
5 avenue Pierre Mendès-France
69676 Bron Cedex
*omar.boussaid@univ-lyon2.fr, ‡rbenmessaoud@eric.univ-lyon2.fr
†{ remy.choquet | stephanea }@gmail.com
http://eric.univ-lyon2.fr

Résumé. Les entreprises sont de plus en plus concernées par des données dites complexes, se présentant sous une forme autre que numérique ou symbolique, issues de sources différentes et ayant des formats hétérogènes. Pour une exploitation à des fins décisionnelles, ces données complexes nécessitent un travail préparatoire pour les structurer et les homogénéiser.
La prolifération des données sous forme de documents XML incite à une solution d'entreposage. Nous proposons dans ce papier une approche basée sur XML, d'entreposage de données complexes contenues dans des documents XML, appelée *X-Warehousing*. Celle-ci définit une méthodologie pour concevoir des entrepôts de données complexes à l'aide du formalisme XML. Pour valider notre approche, nous avons implémenté une application *Java* et nous avons réalisé une étude de cas sur des données complexes concernant les régions suspectes sur des mammographies.

1 Introduction

Avec l'avènement des nouvelles technologies de communication et plus précisément Internet, les entreprises recueillent des masses de données de plus en plus importantes. Ces données étant généralement hétérogènes car elles proviennent de différentes sources. Elles sont dites complexes car elles sont de formats différents et sont sur des supports différents. En médecine, le dossier d'un patient contient des informations générales sur le patient (age, sexe, etc) ; ainsi que des images de scanner, des interrogatoires sous forme d'enregistrements sonores ou des compte-rendus manuscrits de médecins. Pour exploiter de telles données à des fins décisionnelles, il est nécessaire de les structurer et de les homogénéiser. Le langage XML (eXtensible Markup Language) s'avère comme une solution appropriée à ce travail préparatoire sur les données complexes. XML est une norme de W3C[1] et est considéré comme un standard dans la description et l'échange des données. Il représente les données de façon semi-structurée. Sa capacité d'auto-description et sa structure arborescente donne à ce formalisme une grande

[1] http ://www.w3.org/