Comparaison de deux modes de représentation de données faiblement structurées en sciences du vivant

Rallou Thomopoulos*, Patrice Buche**, Ollivier Haemmerlé***, Frédéric Mabille* et Nongyao Mueangdee*

*INRA, UMR IATE, 2 place Viala, 34060 Montpellier cedex 1
{rallou, mabille, nongyao}@ensam.inra.fr
**INRA, UMR Mét@risk, 16 rue Claude Bernard, 75231 Paris cedex 5
Patrice.Buche@inapg.fr
***GRIMM-ISYCOM, Univ. Toulouse le Mirail, Dépt. Mathématiques-Informatique
5 allées Antonio Machado, 31058 Toulouse cedex
Ollivier.Haemmerle@univ-tlse2.fr

Résumé. Cet article présente deux modes de représentation de l'information dans le cadre d'une problématique en sciences du vivant. Le premier, appliqué à la microbiologie prévisionnelle, s'appuie sur deux formalismes, le modèle relationnel et les graphes conceptuels, interrogés uniformément via une même interface. Le second, appliqué aux technologies des céréales, utilise le seul modèle relationnel. Cet article décrit les caractéristiques des données et compare les solutions de représentation adoptées dans les deux systèmes.

1 Introduction

L'étude de la représentation de données faiblement structurées (ou semi-structurées) a connu une explosion récente avec l'émergence de l'internet et la popularité du standard XML. Abiteboul (1997) recense les principaux aspects pouvant caractériser ces données : une structure irrégulière, implicite ou partielle ; un schéma qui se veut indicatif plutôt qu'impératif, souvent construit *a posteriori*, de grande taille, évoluant rapidement ; des types de données éclectiques et une difficulté à établir la distinction entre schéma et données.

De telles données sont courantes dans les sciences du vivant, où l'on trouve également d'autres "verrous" liés à la complexité des phénomènes étudiés (Keet, 2003) : des données dont la précision est limitée par les techniques de mesure, des données variables, non répétables, voire contradictoires, des paramètres nombreux et imbriqués, des données ne pouvant couvrir tous les cas d'expérimentation possibles. C'est notamment en génomique que les bases de données biologiques ont été le plus tôt et le plus abondamment étudiées (Cherry et al., 1998). On trouve cependant des bases de données dans de nombreux autres domaines (environnement, botanique, etc.) (Keet, 2004), avec des modèles de représentation différents fondés notamment sur le modèle relationnel (Bukhman et Skolnick, 2001), le modèle objet (Raguenaud et al., 2002) ou les graphes (Zhong et al., 1999).