ESIEA Datalab
Logiciel de Nettoyage et Préparation de Données

Christopher Corsia*

*ESIEA pôle ECD, 9 rue vésale, 75005 Paris
christopher.corsia@esiea.fr

1 Introduction

Il est communément admis que le temps de préparation des données peut occuper jusqu'à 80% du temps lors d'un projet industriel de fouille de données (Pyle, 1999). L'hétérogénéité des sources, la présence de valeurs manquantes, les erreurs de saisie ou de calcul, les pannes de capteurs, une mauvaise fusion de données sont autant de causes qui peuvent introduire erreurs et incohérences dans une table de données. ESIEA Datalab est une plateforme évolutive programmée en Java qui met à disposition de nombreux outils pour aider à la détection d'incohérences, la correction d'erreurs, la transformation ou la contrainte de variables, etc.

2 Le concept du logiciel

Le nettoyage et la préparation de données peuvent être vus sous la forme d'un processus représenté par la figure 1.

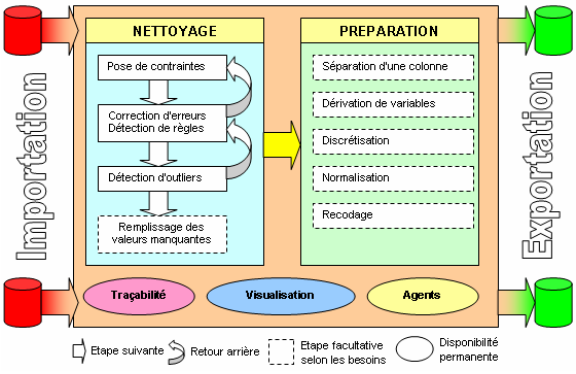

FIG. 1 – *Le nettoyage et la préparation de données vus comme un processus.*

Le logiciel n'impose pas ce processus à l'utilisateur, mais fournit tous les outils nécessaires à sa réalisation. En parallèle, le nettoyage et la préparation des données sont tracés dans