THOT - Extraction de données et de schémas d'un SGBD

Pierre-Jean DOUSSET (France), **Benoît ALBAREIL** (France)
pj@miningdb.com, benoit@miningdb.com

Mots clefs :
Fouille d'information, base de données, système d'information, modèle de données, objet, collecte d'informations, analyse de données

Keywords:
Data mining, data base, information system, data modelling, object, information retrieval, data analysis

Palabras clave:
Datamining, base de datos, sistema de información, modelo de datos, objeto, recopilación de información, análisis datos

Résumé

Nous présentons une méthode utilisant un prototype (Thot) d'extraction de données dans les systèmes de gestion de base de données (S.G.B.D.) ayant pour but une modélisation du système d'information informatisé (S.I.I.) qui soit accessible et compréhensible par tous les acteurs utilisant la méthode.

Cette méthode permet de qualifier le S.I.I. en vue de l'améliorer et de le rendre exploitable par un système dédié à l'intelligence économique. En effet, l'un des problèmes des grands SII est qu'il arrive qu'on perde la maîtrise totale des bases de données : on ne connaît plus le schéma relationnel et on reste confiné au rôle d'utilisateur d'applicatifs sans pouvoir exploiter réellement les données. THOT permet aux décideurs de reprendre et/ou garder la maîtrise d'œuvre de leur S.I.I. Nous détaillons, ici, le processus d'extraction, de stockage et d'exploitation des schémas relationnels à partir des bases de données d'un S.I.I. Chaque étape de cette méthode donne lieu à des résultats intermédiaires exploitables suivant les objectifs des utilisateurs : récupération des schémas relationnels, exportation des données, analyse des données, amélioration et enrichissement des schémas relationnels rencontrés, fédération des données jusque là exploitées dans des applications métiers isolées, génération du reporting, … .

En bref, il faut :

- Intégrer toutes les informations dans un système d'intelligence économique.
- Améliorer le système d'information en terme homogénéité, de granularité, …

Cette méthode permet ainsi aux décideurs d'un organisme d'avoir une vue globale et un contrôle sur les applications métiers et les bases de données liées, qui constituent leur S.I.I. Ceci offre des informations supplémentaires afin d'optimiser la mise en place d'indicateurs sur la situation de leur entreprise ou organisation (en temps réel et a posteriori).
De nouvelles perspectives s'ouvrent alors : amélioration et vision globale de l'information endogène informatisée, rédaction des cahiers des charges en vue d'un refonte des modèles relationnels, mise en commun des données issues d'applications hétérogènes, mise en place d'indicateurs, conception et/ou enrichissement d'un système d'intelligence économique (conjointement à une politique de veille concurrentielle et technologique.).

1 Introduction

Cette méthode se décompose en plusieurs étapes successives formant des cycles possibles aboutissant à des améliorations du S.I.I.

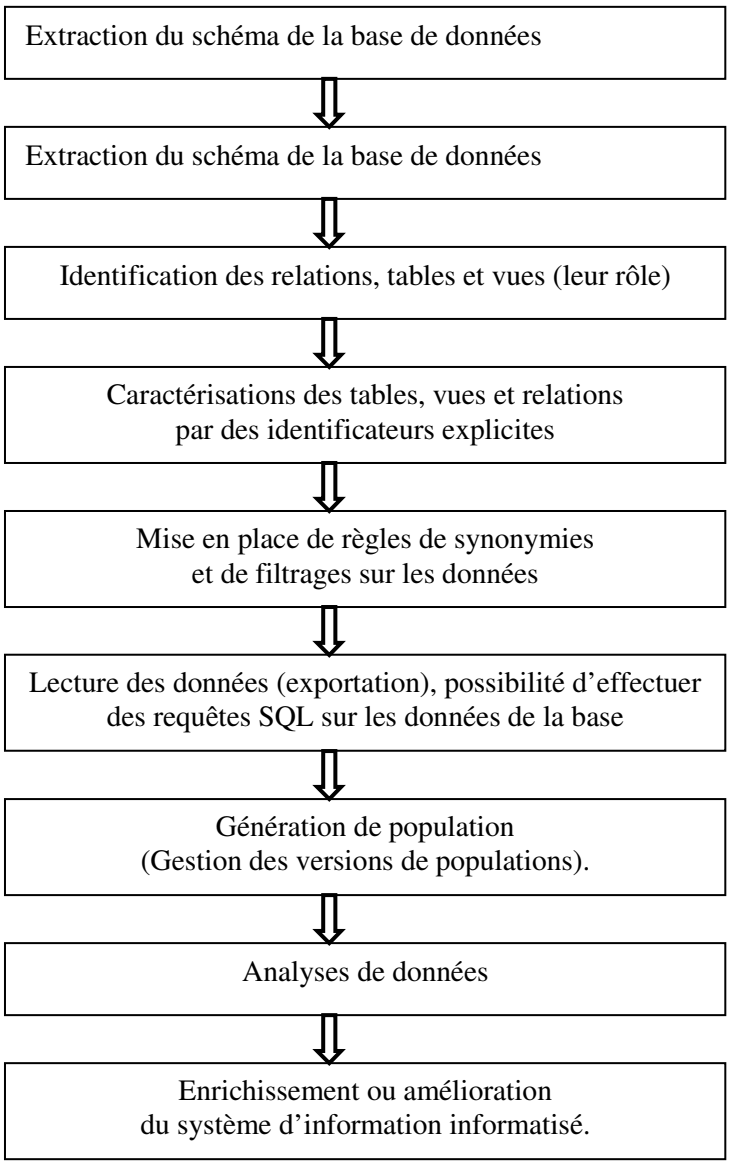


Figure 1. : Les étapes de la méthode

Nous examinerons dans un premier temps, la récupération et le stockage du schéma de la base de données, puis le traitement et l'interaction sur les données : génération et gestion des populations.
Nous développerons ensuite l'export et l'analyse de données avant d'aborder l'intégration de notre méthode dans un S.I.I. d'intelligence économique.

2 Stockage du schéma de la base de données

La première étape de cette méthode consiste à interroger le système de gestion de bases de données, afin de créer une couche applicative de la description des tables, des relations et des types de données présents dans la base. Ceci permet une compréhension du schéma relationnel réel rencontré.

2.1 Modèles relationnels type MERISE

2.1.1 Principe général

Les composants lus dans la base de données traitée sont mis en forme selon les normes des modèles physiques, logiques et conceptuels de la méthode MERISE.

Nous traduisons le schéma physique de la base de données (tables et relations) dans un méta modèle (modèle conceptuel) facilement accessible pour les acteurs de la méthode. Ce méta modèle, enrichi par l'intervention des experts métiers offre alors une lisibilité et une compréhension totale du schéma relationnel.
Le schéma proposé par THOT est le suivant :

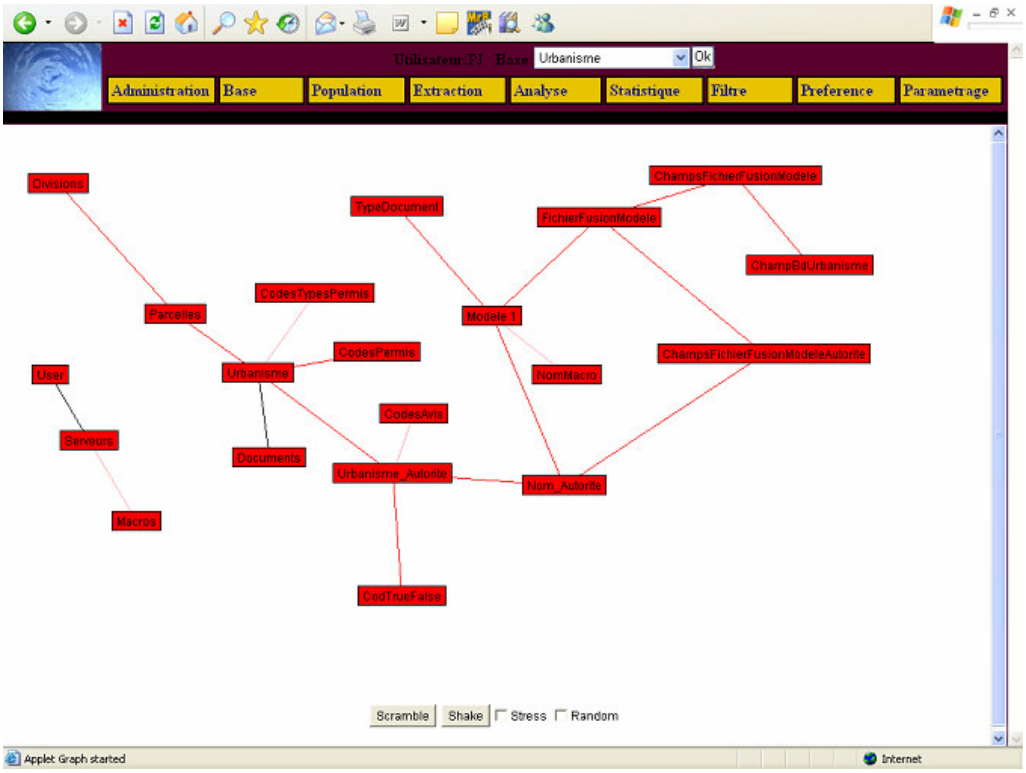


Figure 2. : Le schéma du modèle relationnel (THOT).

2.1.2 Répartition des fonctionnalités par utilisateur

Notre application répertorie cinq types d'acteurs majeurs :

- informaticien
- expert métier
- linguiste
- statisticien, analyste de données
- décideur

L'informaticien responsable du service d'information est en charge de la mise en œuvre et de la mise à disposition des données.

L'expert métier identifie chaque objet de la base et ses relations avec les autres objets de la base en précisant son rôle et sa sémantique.

Le linguiste doit définir les règles de niveau de granularité, de synonymie et d'ontologie des données. Ces règles pourront par la suite être utilisées par l'éditeur et intégrées dans les applicatifs sous forme de règles de saisie.

Le statisticien / analyste de données établit des processus de traitement des données.

Le décideur valide la mise en avant d'indicateurs pertinents, résultants des étapes précédentes et soumet de nouvelles exigences à la vue de ces premiers résultats, afin de définir des indicateurs supplémentaires.

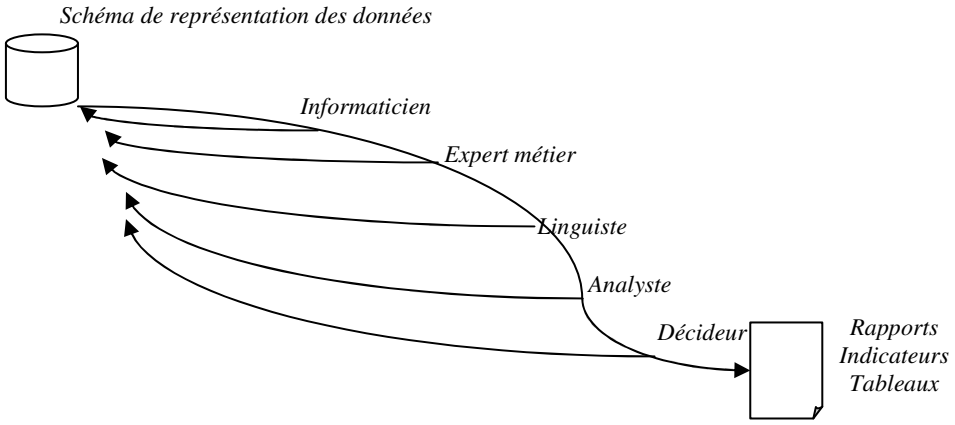


Figure 3. : Enchaînement de interventions des différents acteurs.

THOT se propose de stocker la description de chaque composant du modèle relationnel étudié dans un modèle objet de stockage type afin de faciliter les traitements liés à la représentation des bases de données relationnelles et à l'analyse des données.

2.2 Spécification des corrections et/ou améliorations

2.2.1 Les données

Les tâches inhérentes au linguiste, doivent être traitées selon des règles dans un ordre de prépondérance, elles-mêmes décomposées en opérations ordonnées. L'enchaînement cyclique des règles et opérations doit s'arrêter à un niveau de satisfaction sémantique défini au préalable. Parmi un ensemble de synonymes, le linguiste doit définir lequel choisir en fonction du métier et d'un grand nombre de paramètres exogènes au S.I.I. THOT va offrir au linguiste des containeurs de programmation réutilisables qui lui permettront de définir des règles qui l'aideront à établir les règles de synonymie.

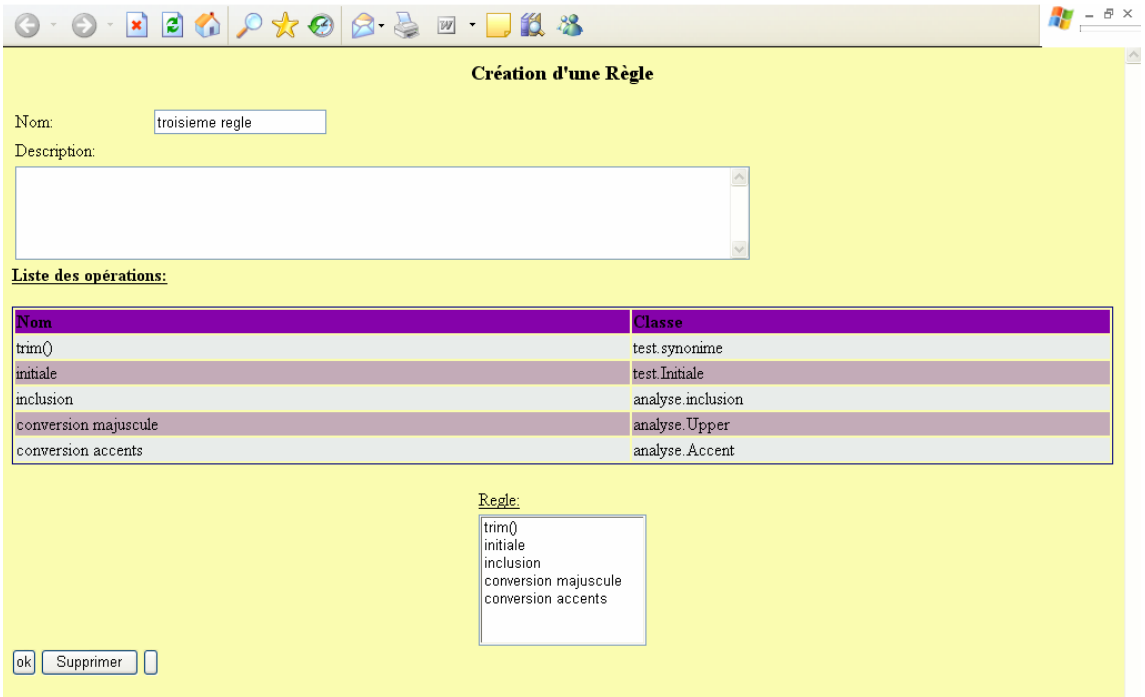


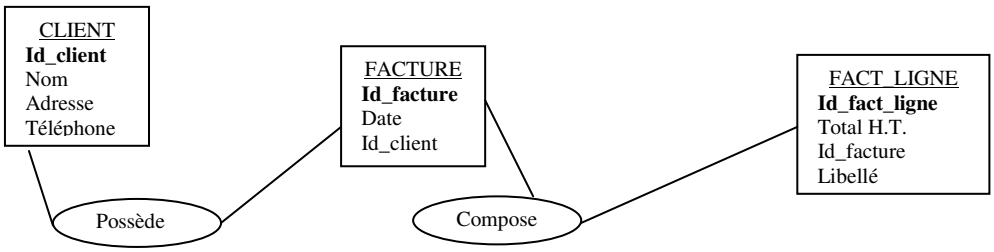
Figure 4. : Définition d'une règle.

2.2.2 Le modèle relationnel

Après analyse des données, il peut s'avérer que certains renseignements aient été oubliés voire mal définis. Il est alors possible pour le décideur de rajouter l'emplacement de cette information dans le schéma relationnel et d'en définir les caractéristiques.
=> Le décideur est maître d'œuvre car il va être en mesure d'évaluer les charges induites que ce soit en terme de coût ou en terme de charge de travail.

3 Les Populations

Un des buts de la méthode présentée ici est la génération de population en vue du traitement en analyse de données.



Une population possible:

Nom	Adresse	Téléphone	Date	Total H.T.	Libellé

Figure 5. : Exemple de population (l'individu central est le client) à partir d'un extrait de bases de données :

*Dans ce cas, il y aura autant de lignes que d'enregistrements dans la table **FACT_LIGNE**.*
*Les valeurs d'un **Id_facture** donné seront répétées autant de fois qu'il y a d'enregistrements de **FACT_LIGNE associés***

Cette génération est facilitée par l'organisation des métas modèles stockés dans des classes.
Une population est un ensemble de tables liées. Elle est définie soit automatiquement, soit via différentes relations choisies par l'utilisateur. Elle peut également être déjà présente dans la base de données.

Exemple

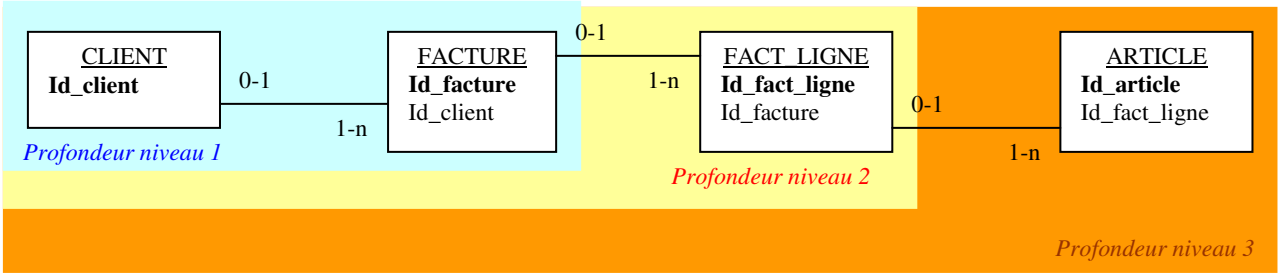


Figure 6. : Profondeur des liaisons.

THOT propose quatre environnements de définition de populations dans une base de données :

- Automatique
- Par niveau de profondeur
- Pas à pas
- SQL

Ils ont tous pour but une approche progressive de la génération de populations en vue d'une analyse de données. Ces méthodes permettent ensuite de mettre en place des outils d'intelligence économique.

3.1 Génération automatique de population

Il s'agit d'identifier les tables pertinentes en vue d'une analyse de données suivant certains critères :

- Nombre de clés exportées
- Nombre de clés importées
- Nombre d'enregistrements
- Nombre de colonnes

Cette identification peut être automatisée, via des algorithmes de sélection faisant appel à des critères donnant ainsi la ou les table(s) les plus porteuses d'information. La génération automatique de population peut ne pas être la plus pertinente, mais elle permet d'aborder l'intelligence économique et de mettre en place les premières études

3.2 Génération manuelle de populations

3.2.1 Par niveau de profondeur

L'utilisateur définit au préalable un niveau de profondeur maximum ; ainsi tous les chemins induits par toutes les relations de la table de départ seront suivis, dans la limite de ce niveau de profondeur, afin de générer la population.

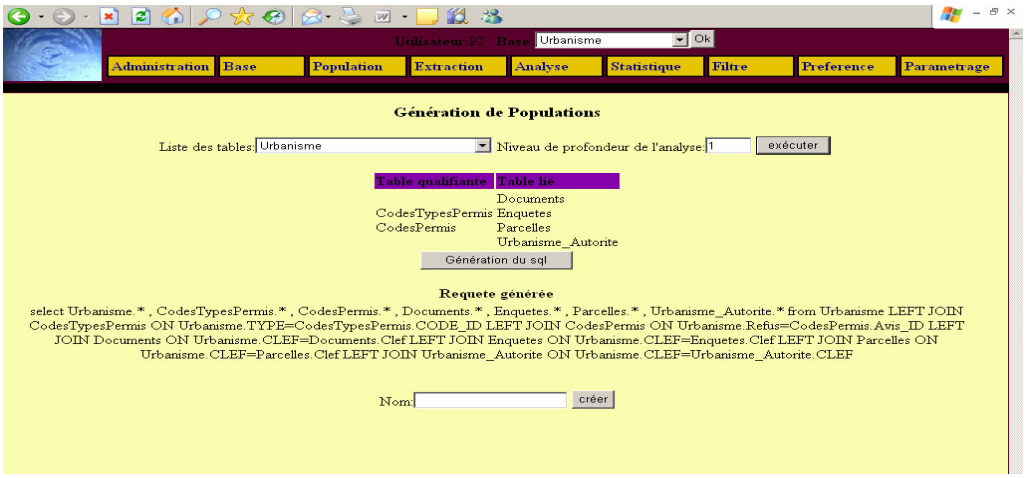


Figure 7. Génération par THOT.

C'est le premier pas d'une interaction entre l'utilisateur et sa (ou ses) base(s) de données puisque l'utilisateur choisit sa table. Néanmoins, cette méthode est limitée : un niveau de profondeur trop élevé pourrait entraîner un volume de données trop important qui perdrait en pertinence.

3.2.2 « Pas à pas »

En suivant des relations définies dans les méta modèle, d'une table à l'autre, l'utilisateur voit se construire table après table, relation après relation le chemin qui constituera sa population. Cette solution permet à l'utilisateur de construire les caractéristiques de sa population.

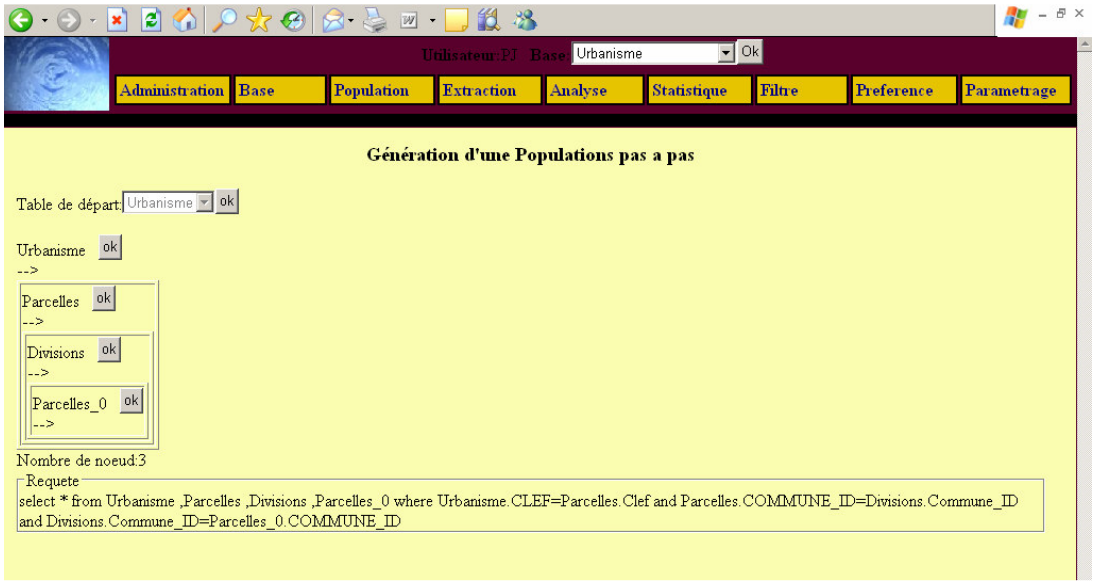


Figure 8. Génération par THOT.

3.2.3 SQL

L'utilisateur rédige directement dans un éditeur de texte la requête en langage SQL qui sera à l'origine de sa population. En outre, ceci permet de mettre à la disposition des utilisateurs un « requêteur » SQL, même si ses applicatifs n'en disposent pas.

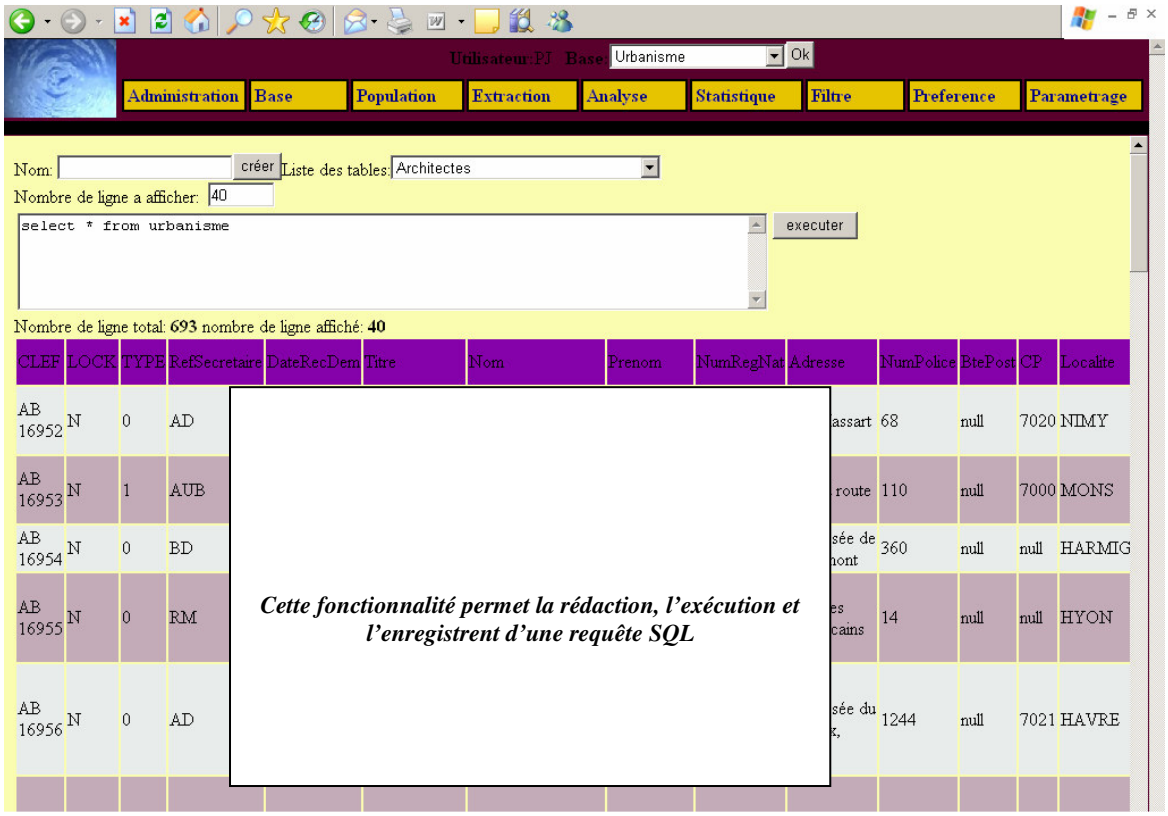


Figure 9. : Module SQL de THOT.

Thot permet ainsi un apprentissage progressif de l'analyse de données. Par les différentes façons de créer une population, l'utilisateur va de la plus automatique, et donc la plus générale, vers les formules

les plus complexes. Cela permet aussi aux décideurs de découvrir le monde complexe des bases de données et d'avoir une vision plus précise des coûts engendrés par une modification sur la base.

3.3 Les vues

Les vues extraites sont elles mêmes des populations potentielles. Ceux sont des ensembles de tables liées existant dans la base de données étudiée. Elles ont en général été créées pour optimiser et simplifier le traitement des applicatifs.

THOT permet de sélectionner tout ou partie de ces vues comme populations pertinentes en vue d'une analyse de données

4 Systèmes multi bases de données

Les systèmes de gestion de bases de données sont souvent hétérogènes au sein d'un SII. De plus, ils nécessiteraient d'être enrichis par des données externes. Parce que THOT va stocker la description de chaque base de données, on peut envisager au sein de THOT, soit de décrire des relations inter bases, soit d'importer des données externes dans la base étudiée et intégrer leur description dans THOT.

4.1 Relations inter bases

THOT permet d'enrichir virtuellement le schéma d'un SGBD existant. THOT étant un environnement multi base de données, on peut envisager de définir un lien inter bases de données, ce qui aurait pour effet de relier plusieurs schémas. Ceci nécessiterait la création d'un interpréteur complexe d'analyse de population. Il devrait déterminer les tables des différents schémas afin de simuler les relations décrites par l'utilisateur.

4.2 Intégration de données externes

THOT étant une zone de transit de données, il peut très bien intégrer des données d'une base vers une autre. Cette fonction est facilitée par le fait que THOT connaît la description des bases qui lui sont déclarées. Cette méthode a l'avantage d'enrichir une base existante par des données qui lui manquaient. Si cet ajout se révèle utile et pertinent, cela peut faire l'objet d'un cahier des charges en vue du regroupement de ces données dans une seule de ces applications. Cette méthode est de plus la plus facile à mettre en œuvre car des mécanismes d'intégration peuvent mettre à jour les données

Ces deux méthodes aux principes différents permettent une analyse de systèmes hétérogènes en matières de bases de données. Cependant l'intégration de données externes a l'avantage de faire évoluer le SII vers un système homogène.

5 Export et analyse de données

La fonction 'Analyse' de THOT permet de générer des analyses de base sur les populations stockées. Il est de plus possible de générer des formats d'export des populations et des analyses. Les évolutions futures de THOT seront axées sur des analyses plus complexes mais surtout sur la diversité des formes de leur représentation graphique.
Pour la pertinence de chacune des analyses et de toutes dans leur ensemble, elles doivent être stockées et « versionnées » ; permettant par ailleurs une analyse temporelle.

5.1 Versions des analyses

Chaque analyse générée est enregistrée et donne lieu à une nouvelle version.
La mise en version de chaque analyse permet un niveau supplémentaire d'analyse de données : une analyse de l'évolution dans le temps.

5.2 Export

Les données obtenues par le biais des requêtes ou de la génération de population sont exportables dans les formats les plus répandus pour permettre une exploitation de nos résultats (à toutes les étapes) par d'autres outils de reporting et d'analyse de données.
Nous permettons notamment des exports TETRALOGIE, csv, xml, MS Excel, texte.
Selon les besoins, nous adaptons un export à la demande en peu d'effort ; par simple lecture, exploitation de nos objets.

6 Intégration de la méthode dans le système d'intelligence économique (S.I.E.)

6.1 Complémentaire des outils et méthodes actuels

Notre méthode se positionne en complément des outils et méthodes existants en intelligence économique. En effet, la diversité des formats d'exportation de THOT permettent l'intégration des données dans les outils de reporting d'un organisme.
De plus, son originalité réside dans le versionnement de ses fichiers d'exports, qui assure la lisibilité de l'évolution des données dans le temps.
.

6.2 Complémentaire du S.I.I.

L'originalité de THOT est la mise en place d'un requêteur « universel » de base de données.
Il s'intègre au SII en proposant des représentations virtuelles sans modifier les systèmes existants.
Par l'ajout de nouvelles règles et de données, il peut être un soutien à la conception de nouvelles évolutions du système.

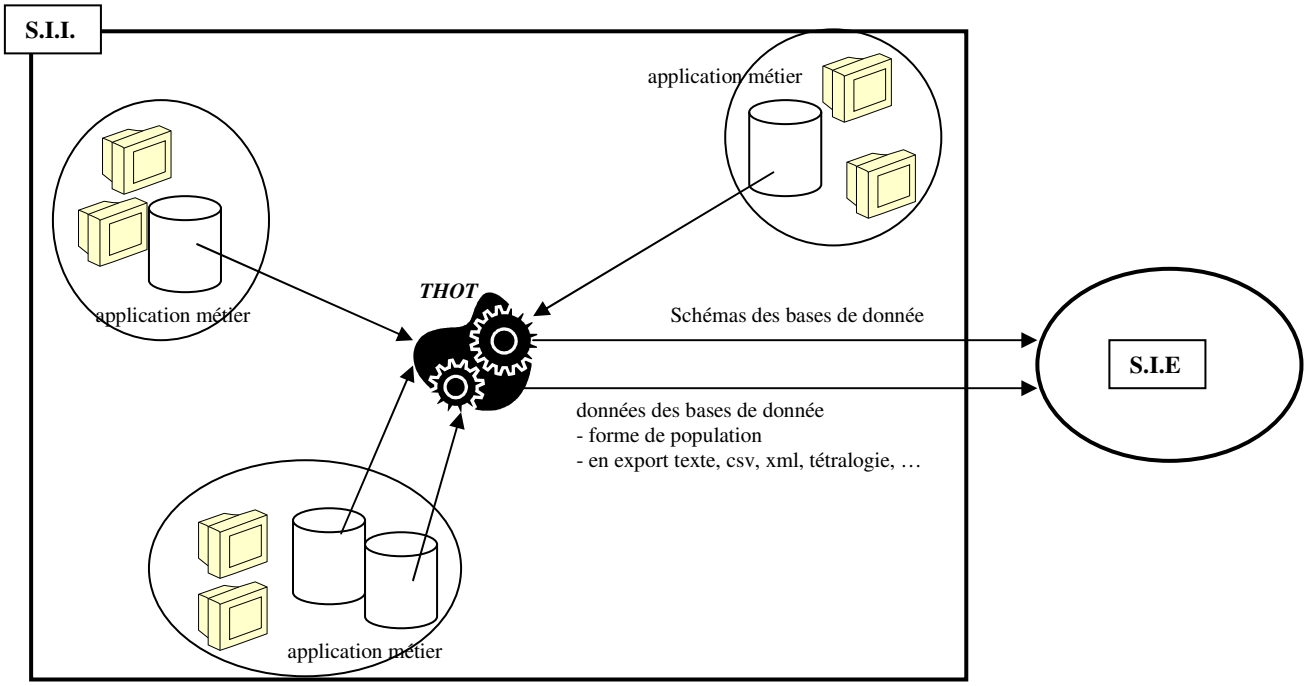


Figure 10. Intégration de THOT dans un S.I.

6.3 Architecture autonome et indépendante

Au niveau applicatif nous avons fait le choix d'un conception purement orientée web.
Nous offrons ainsi une grande maniabilité et intégration dans des systèmes d'information existants : par exemple la mise en place via un portail web.

Sur un autre plan, notre applicatif s'exécute sur quasiment n'importe quel système d'opération (Microsoft, Linux, Unix, Apple, …).
Nous interagissons sur presque n'importe quel SGBD. Nous avons pu valider nos solutions sur des SGBD différents : MySQL (innoDB, ISAM, MyISAM, …), Oracle (7x, 8x, 9x), DB2, Paradox, Access ; quelque soit la l'architecture du serveur de base de données (Windows, IBM AS400, Linux, UNIX).
Il suffit d'une connexion réseau avec le serveur de base de données : via un LAN, une ligne dédiée / spécialisée, le réseau Internet, etc. Les résultats (extraction de données, population, …) sont récupérables de la même manière, indépendamment de l'architecture système et réseau.

7 Conclusion

Les travaux effectués à ce jour, en partenariat avec Tétralogie, montrent l'efficacité et la complémentarité de notre méthode Les données endogènes d'un organisme sont une composante essentielle dans un système d'intelligence économique. L'analyse des données des bases de données de l'organisme, ajoutée aux analyses faites sur les informations exogènes (veille, etc.) permet d'enrichir tout le système d'intelligence économique, d'affiner les indicateurs existants et d'en obtenir de nouveaux.

8 Bibliographie

[1] BRIARD G., avec la collaboration de la société Digora, *Oracle9i sous Windows*, 2e tirage 2004, éditions Eyrolles

[2] CABANE R. & LEBOEUF C. *Algèbre linéaire*, Ellipses

[3] DKAKI T., DOUSSET B., MOTHE J., *Recherche de l'Information Stratégique dans les Bases de Données - Veille Scientifique et Technique*. Dans : *15ème Conférence Informatique des Organisations, des Systèmes d'Information et de Décision*, -, juin 1997

[4] GEORGIN J.P., *Analyse interactive des données (ACP, AFC) avec Excel 2000 Théorie et pratique*, Collection « Didact Statistique » Presses Universitaires de Rennes 2002

[5] Hinz S. Dubois P., Pedersen C., *MYSQL 5 Guide officiel*, éditions Campus Press

[6] HOLZ H. SCHMITT B., TIKHART A., *Internet et intranet sous Linux – Guide pratique pour l'entreprise*, édition Eyrolles

[7] LESIEUR L. & LEFEBVRE J. *Mathématiques, Tome 3 : Compléments d'analyse, statistiques et probabilités*, éditions Armand Colin

[8] SALLES M., *L'importation des méthodes des systèmes d'information vers l'intelligence économique*, lors de VSST 2004 (octobre) à Toulouse

[9] FRANCO J.M., *Le Data Warehouse, le Data Mining*, Eyrolles, 1996

[10] LEFEBURE R. et VENTURI G., *Le Data Mining*, Eyrolles, 1998

[11] http://altlas.irit.fr

[12] http://ieut1.irit.fr/