Chapitre 9

Conclusion générale

9.1 Bilan et contributions

Cette thèse a été réalisée dans le cadre d'une convention CIFRE entre le laboratoire ERIC et l'établissement bancaire LCL. Au-delà de l'aspect d'ingénierie que revêtait la mission réalisée en entreprise, LCL a su suscité des problématiques scientifiques bien réelles. En effet, le souci de la personnalisation des outils est permanent chez LCL. Ceci est sans doute dû au fait que les employés exercent des métiers différents tout en faisant partie de la même banque, ils ont donc des attentes et des besoins différents, etc.

Ainsi, nous avons proposé une solution de personnalisation qui consiste à recueillir les connaissances des utilisateurs pour créer de nouveaux axes d'analyse répondant à leurs propres besoins. Cette solution se base sur une évolution du schéma de l'entrepôt guidée par les utilisateurs qui vise plus précisément à mettre à jour les hiérarchies de dimension en créant de nouveaux niveaux de granularité. Nous avons alors tenté de redonner tout son sens à l'expression «technologie centrée utilisateur» en le plaçant au cœur du processus d'évolution, en plus du fait qu'il soit maître de l'analyse au sens de la navigation.

Notre approche de personnalisation se basant sur une évolution de schéma, elle permet de rendre les nouvelles possibilités d'analyse pérennes et partageables avec d'autres utilisateurs, ce qui n'est pas le cas avec la proposition de création de variables offerte par certains éditeurs de logiciels décisionnels.

Vis-à-vis de cette évolution de schéma, notre approche de personnalisation s'inscrit dans une alternative de mise à jour de schéma. Un des avantages est alors de pouvoir la mettre en œuvre sur n'importe quel entrepôt de données existant sans complexité particulière.

Du point de vue de cette mise à jour de schéma, le principe que nous utilisons se rapproche des travaux proposés dans [BSH99], dans lesquels un ensemble d'opérateurs de mise à jour de schéma étaient proposés. Dans notre travail, nous nous sommes concentrés sur une partie des évolutions proposées et ce, pas seulement au niveau structurel. En effet, dans notre approche, nous nous sommes intéressés également à fournir les données nécessaires pour l'évolution, en exploitant la connaissance des utilisateurs.

Notre objectif d'enrichissement des hiérarchies de dimension est commun avec celui de Mazon et al. [MT06]. Les deux approches diffèrent néanmoins sur le moyen d'y parvenir. En effet, leur approche vise à enrichir les hiérarchies de dimension de façon automatique, en exploitant WordNet. Mais dans notre cas, cet enrichissement est réalisé grâce aux utilisateurs, donnant ainsi une place centrale à l'utilisateur dans le processus d'évolution.

Cet enrichissement permet alors de créer de nouveaux chemins d'agrégation, allant au-delà de la proposition faite dans [EV01], dans laquelle les utilisateurs pouvaient seulement modifier les chemins d'agrégation existants, en exprimant des exceptions dans le processus d'agrégation au niveau des instances.

Concernant les travaux traitant de la personnalisation dans les entrepôts de données eux-mêmes, il s'avère que notre approche s'inscrit dans une perspective différente des travaux émergents dans le domaine. En effet, les principaux travaux se basent sur l'expression de préférences pour personnaliser le processus d'analyse en diminuant les réponses aux requêtes [BGMM06, BGM+05] ou en diminuant le nombre d'opérations à réaliser lors de la navigation [RTZ07]. Dans notre travail, la personnalisation n'est pas fondée sur une expression de préférences pour gérer des possibilités existantes. Il s'agit au contraire d'étendre ces possibilités en permettant la réalisation de nouvelles analyses qui soient personnalisées par rapport aux besoins des utilisateurs, en prenant en compte leurs propres connaissances.

Notre démarche se base alors sur une architecture globale comprenant quatre modules :

- un module d'acquisition des connaissances utilisateurs sous forme de règles d'agrégation de type «si-alors» ;
- un module d'intégration des règles d'agrégation dans l'entrepôt de données ;
- un module d'évolution de schéma permettant la mise à jour des hiérarchies de dimension ;
- un module d'analyse permettant à l'utilisateur d'avoir de nouvelles analyses OLAP basées sur le nouveau schéma.

Pour soutenir cette architecture, nous avons défini un modèle d'entrepôt de données évolutif à base de règles d'agrégation *R-DW*. Ce modèle est composé d'une partie «fixe» et d'une partie «évolutive». La partie fixe est constituée de la table des faits et des tables de dimension qui lui sont directement reliées. Elle constitue une réponse à des besoins d'analyse initiaux, définis lors de la conception de l'entrepôt. Ces besoins initiaux peuvent être considérés comme communs à l'ensemble des utilisateurs. La partie évolutive est composée de l'ensemble des hiérarchies de dimension pouvant être mises à jour par les utilisateurs. Nous nous assurons que ces mises à jour n'introduisent pas d'incohérences dans les analyses en les propageant au niveau du schéma. Cette partie évolutive apporte ainsi une flexibilité pour la prise en compte de nouveaux besoins d'analyse.

En outre, nous avons proposé un méta-modèle de l'entrepôt de données évolutif qui nous permet d'appliquer notre démarche sur n'importe quel entrepôt. En effet, ce méta-modèle permet de décrire n'importe quel entrepôt de données évoluant selon notre approche. Nous assurons ainsi la généricité de notre modèle évolutif.

Nous avons déployé notre approche dans un contexte relationnel en proposant un modèle d'exécution qui a pour but de gérer l'ensemble des processus liés à l'architecture, de l'acquisition des règles à l'évolution du schéma.

Par ailleurs, nous nous sommes intéressés à l'évaluation de la performance de notre modèle évolutif. Étant donné que l'évaluation de la performance dans les entrepôts de données est généralement basée sur une charge, nous avons proposé une méthode de mise à jour incrémentale de la charge en fonction des modifications subies par le schéma de l'entrepôt, afin que la charge reste cohérente vis-à-vis de ces modifications.

LCL a non seulement suscité notre problématique de personnalisation, mais a constitué par la suite un terrain d'application pour la mise en œuvre de nos propositions en matière de personnalisation. Ainsi, l'ensemble de nos propositions a fait l'objet d'une implémentation au travers de la plateforme WEDriK qui permet donc d'impliquer les utilisateurs dans l'évolution du schéma de l'entrepôt. Nous avons donc appliqué notre démarche sur l'entrepôt de données test LCL-DW construit à partir de données réelles de LCL.

Cette expérience de thèse en collaboration avec LCL a été enrichissante tant sur le plan humain que professionnel. J'ai eu l'occasion d'utiliser mes compétences, mais j'ai également beaucoup appris. Par exemple, j'ai pu acquérir de nouvelles connaissances sur les structures organisationnelles, les produits, le marketing, etc.

LCL est une entreprise nationale qui présente l'avantage d'être fortement struc-

turée tout en permettant un dynamisme dans les différents niveaux hiérarchiques. En effet, j'ai pu réellement apprécier le fait que la direction d'exploitation Rhône-Alpes Auvergne soit très active en matière de développement d'outils en local pour faciliter le travail de ses employés. D'ailleurs, ces outils sont parfois repris au niveau national pour en faire profiter les autres directions d'exploitation. Un tel fonctionnement me paraît pertinent dans la mesure où les personnes développant ces outils au niveau local sont plus à même de déterminer les besoins réels des employés de part leur proximité. Partir des besoins est en effet essentiel.

C'est ce que nous avons fait dans le cadre de la conception et du développement de la plateforme MARKLOC. En effet, nous avons réalisé une importante étude préalable des besoins, de l'existant, etc. Dans le cadre de cette étude pour le développement de MARKLOC et d'autres outils, j'ai pu constater la difficulté pour des opérationnels à définir un cahier des charges. En effet, il s'agit de travailler avec des personnes qui ne sont pas informaticiennes, qui ne peuvent donc mesurer le travail nécessaire au développement et n'ont pas conscience de l'importance d'être précis dans l'expression des besoins.

J'ai pu alors développer ma capacité à être une véritable interface entre les utilisateurs et la définition d'un cahier des charges. Il s'agit réellement de passer du langage naturel à un langage technique, une formalisation des problèmes et des solutions que l'on peut apporter au niveau informatique.

Ce travail est d'autant moins évident, qu'il est très difficile pour les utilisateurs d'exprimer de façon exhaustive les besoins, les procédures existantes, etc. Cela implique, lors de la récolte des informations, d'être capable de creuser par des questions, pour être sûr qu'aucun détail ne nous a échappé. Dans ce contexte, les validations successives sont nécessaires. Elles permettent de faire émerger des points qui n'avaient pas été abordés, des précisions qui ont leur importance.

Dans le cadre du développement de la plateforme MARKLOC, la phase de conception et la phase de tests ont été relativement longues. Mais je suis convaincue aujourd'hui qu'il s'agit de la meilleure façon de faire. En effet, il y a une forte tendance dans les entreprises à développer des outils dans l'urgence sans forcément prendre du temps pour réfléchir de façon posée au problème, à la conception de l'outil. Or, j'ai pu constater, que ce temps est bénéfique non seulement vis-à-vis du développement lui-même, mais également pour l'évolution de l'outil. En effet, en prenant le temps de bien concevoir un outil, il est généralement plus facile de le faire évoluer. De plus, le fait de mettre à disposition l'outil sans qu'aucun problème technique ne soit relevé est une réelle satisfaction. Pour permettre ce résultat, une longue phase de tests a été mise en œuvre. Nous avons dû définir un protocole qui

envisage tous les scénarios possibles : en fonction des différents utilisateurs (différents profils), des différentes interventions qu'ils peuvent réaliser (demandes validées ou refusées par exemple), etc.

La plus grande satisfaction que l'on peut avoir par rapport à un projet de développement de ce type est bien sûr de constater que le produit est réellement utilisé, d'avoir des retours sur le fait que les personnes qui l'utilisent apprécient l'outil. En effet, cette satisfaction n'est pas forcément obtenue dans le cadre du développement réalisé dans un contexte scientifique pour tester et valider des propositions qui ne seraient pas des réponses à des besoins réels (d'une entreprise par exemple).

Ainsi, je suis convaincue qu'il est réellement intéressant de pouvoir développer des collaborations entre les laboratoires de recherche et les entreprises. Cela permet, pour les premiers, de s'imprégner de la réalité des problèmes des entreprises et de voir émerger de réelles problématiques scientifiques et, pour les seconds, de bénéficier de forces de proposition pour résoudre ces problématiques. En effet, les objectifs des uns et des autres ne sont pas incompatibles, il finissent par converger.

C'est ce qui s'est produit dans mon cas. La spécificité de ma thèse était d'arriver à mener un projet selon deux points de vue différents mais complémentaires : ingénierie d'une part, et scientifique d'autre part. Le fait d'être imprégnée de problèmes réels, de se nourrir de lectures, etc. a permis de proposer des solutions à des problèmes concrets, en les mettant en œuvre.

L'objectif de LCL était de disposer d'un outil qui fonctionne, facile à utiliser et répondant à certains besoins. L'objectif pour le laboratoire se situe davantage au niveau de la proposition de solutions à des problèmes et à la publication de contributions expliquant ces solutions, solutions pouvant être validées par un travail de développement. Les objectifs convergent alors lorsque le laboratoire va être à même de proposer des solutions scientifiques à des problèmes réels basées sur des outils utilisables.

La plus grande satisfaction que l'on peut obtenir par rapport à une thèse de ce type est alors d'obtenir une satisfaction finale vis-à-vis des attentes de l'entreprise et vis-à-vis des attentes du laboratoire, ce qui est mon cas aujourd'hui. En effet, la plateforme MARKLOC est aujourd'hui régulièrement utilisée et est appréciée de ses utilisateurs. Sur le plan scientifique, nous avons pu faire des propositions qui ont été validées par des publications.

9.2 Perspectives de recherche

Concernant notre proposition de personnalisation, les discussions que nous avons menées sur notre travail tout au long de ce mémoire ont fait émerger de nombreuses perspectives, directement liées à notre travail sur la personnalisation, ou de façon plus large sur l'évolution de schéma et la performance.

Tout d'abord, un des points cruciaux qu'il nous reste à explorer dans le cadre de notre proposition est la gestion de l'évolution des règles. Si ce problème peut être abordé sous l'angle de l'évolution de schéma de façon générale avec les alternatives que l'on connaît de mise à jour ou de versionnement, il n'en demeure pas moins que les particularités de notre approche doivent être prises en compte. L'une des particularités les plus notables est l'implication de l'utilisateur dans le processus de mise à jour des hiérarchies de dimension. Il s'agit alors de connaître quels sont les besoins réels au niveau de l'historisation des dimensions. Mais il s'agit également de bien prendre en compte une certaine facilité d'utilisation.

Lorsque le nombre d'instances à identifier lors du regroupement devient trop important, donc la tâche trop fastidieuse pour l'utilisateur, une méthode d'apprentissage permettant un regroupement automatique des instances paraît pertinente. Cette méthode permettrait également de découvrir des regroupements intéressants pour l'analyse auxquels l'utilisateur n'aurait pas pensé. Par exemple, dans [RB07], les auteurs proposent d'utiliser la méthode des K-means pour construire les classes de regroupement d'instances pour représenter le nouveau niveau de hiérarchie à créer. Notre approche étant basée sur des règles d'agrégation de type «si-alors», il nous paraît intéressant de pouvoir générer ces règles de façon automatique par l'application d'une méthode d'apprentissage non supervisée. Ainsi nous pensons que l'algorithme KEROUAC [JN03] peut répondre à notre objectif. KEROUAC correspond à l'acronyme des termes anglais Knowledge Explicit Rapid Off-beat User-centered. Ces termes font référence aux caractéristiques de la méthode : caractérisation explicite des classes (Knowledge Explicit), coût calculatoire relativement faible (Rapid), bonne utilisabilité (User-centered). L'un des aspects qui nous intéresse particulièrement est que chacune des classes de la partition résultat est caractérisée par une règle logique ; alors que très souvent le résultat des algorithmes fournit la composition des classes uniquement. Dans un contexte relationnel, nous pensons qu'il serait intéressant d'intégrer cet algorithme au sein du SGBDR, en particulier si le nombre d'instances à classer est important. En effet, comme nous l'avons montré dans [BDFU07], il est possible d'intégrer les méthodes de fouille au cœur des SGBD, optimisant ainsi les temps de réponses face à de larges volumétries de données à

traiter, en exploitant les outils du SGBD tels que les index bitmap par exemple comme nous l'avons proposé dans [FB05b, FB05a].

Dans notre travail, la personnalisation a pour objectif de répondre à des besoins d'analyse spécifiques donnant la possibilité aux utilisateurs de créer de nouveaux niveaux de hiérarchie. Ces niveaux créés peuvent intéresser d'autres utilisateurs. Il est donc crucial qu'un utilisateur qui réalise une analyse en fonction d'un niveau créé par un autre utilisateur connaisse exactement la sémantique de ce niveau. Pour ce faire, nous pensons que le recours à un processus d'annotations, comme il a pu être proposé dans [CCRT07], peut être pertinent. En effet, dans ce travail les auteurs traitent du concept de mémoire d'expertises décisionnelles. Un des objectifs de cette mémoire est d'éviter la perte de connaissances lors du départ d'un collaborateur et de faciliter le transfert de ces connaissances entre les collaborateurs. Deux aspects ont retenu plus particulièrement notre attention dans cette proposition. Il s'agit d'une part de l'idée de préciser la sémantique au niveau des concepts dans le schéma. D'autre part, il s'agit de l'idée d'usage collectif, de partage d'expertises. Ainsi, notre approche présente ces deux idées : il est en effet crucial de pouvoir préciser la sémantique du niveau créé dans le schéma afin de pouvoir partager cette possibilité d'analyse supplémentaire avec d'autres utilisateurs. Le créateur du nouveau niveau de hiérarchie pourrait annoter celui-ci afin de lui donner une bonne description, pour que la compréhension soit facilitée pour les autres utilisateurs. C'est ce qui permettra un réel partage des nouvelles possibilités d'analyse en assurant la bonne interprétation de ces analyses. Rappelons que cette nécessité est accentuée dans le cas de versions utilisateurs différentes qui consistent à représenter un même niveau avec des règles de construction différentes (cas des classes d'âge par exemple).

Dans le contexte des entrepôts de données évolutifs, nous pensons qu'une «simple» maintenance des structures d'optimisation peut s'avérer insuffisante. En effet, l'évolution du schéma et des données de l'entrepôt peut nécessiter une évolution de la configuration même des structures d'optimisation. Nous pensons alors que le travail réalisé pour l'évolution de charge pourrait être exploité de façon pertinente dans ce contexte. Les algorithmes de sélection d'index et de vues à matérialiser se basent généralement sur l'utilisation d'une charge. Ainsi, répercuter l'évolution du schéma et des données sur la charge pourrait permettre une gestion des performances de façon pro-active. Nous entendons par pro-active le fait de mettre à jour la charge pour sélectionner et donc éventuellement modifier la configuration des index et des vues à matérialiser. Dans ce contexte, afin d'évaluer la performance de ces entrepôts de données évolutifs, nous pensons qu'un recours au benchmark (banc d'essai) peut être pertinent. Dans ce contexte, il serait nécessaire de doter les bancs d'essais

d'opérateurs permettant de faire évoluer les schémas et les données des entrepôts.