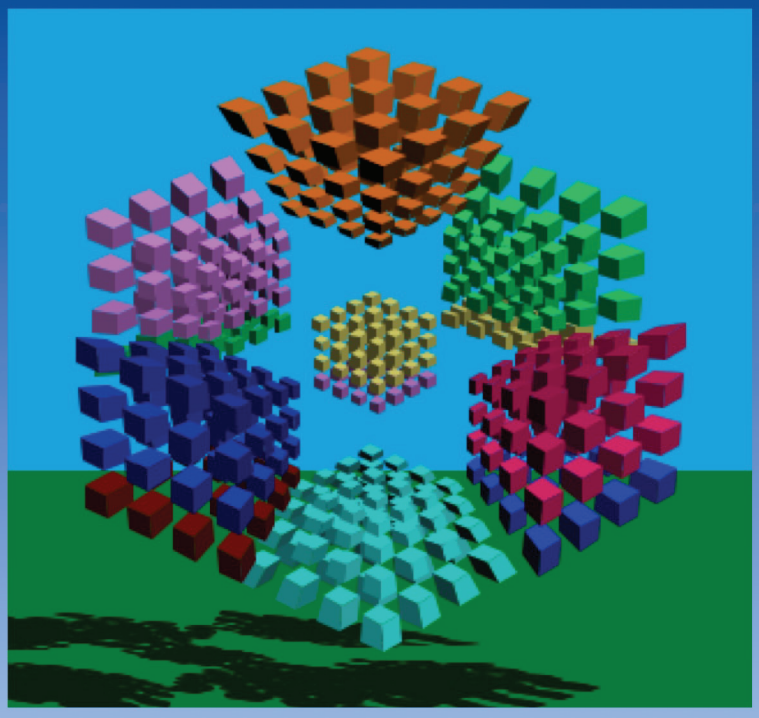
EGC'2013

13e Conférence Francophone sur l'Extraction et la Gestion de Connaissances

Université Paul Sabatier – IRIT – Toulouse

29 janvier 2013 – Journée Ateliers/Tutoriels



Aide à la Décision à tous les Etages (AIDE)

Université de Toulouse

Université Paul Sabatier TOULOUSE III

IRIT CNRS - INPT - UPS - UT1 - UTM

EGC

Atelier aide à la Décision à tous les Etages (AIDE)

Organisateurs : Frédéric AMBLARD (IRIT - Université Toulouse I Capitole), Cécile FAVRE (ERIC - Université Lumière Lyon 2), Franck RAVAT (IRIT - Université Toulouse I Capitole)

Préambule

Le terme "aide à la décision" est utilisé par différentes communautés qui ne communiquent pas toujours ensemble et qui y associent des acceptions différentes :

- En informatique décisionnelle, l'aide à la décision consiste à analyser des indicateurs en fonction d'axes d'analyses le plus souvent représentés dans un espace multidimensionnel. Ces données décisionnelles sont extraites généralement des sources de production pour être intégrées et synthétisées au sein d'un entrepôt de données ("datawarehouses").

- En fouille de données ("datamining"), l'aide à la décision consiste en l'extraction de connaissances dans de grands volumes de données pour en extraire des informations pertinentes. La fouille de données repose sur des algorithmes permettant de déterminer des corrélations ou des prédictions et peuvent facilement se coupler avec un entrepôt de données.

- En simulation, l'aide à la décision peut recourir à la construction (parfois en collaboration avec le décideur et/ou en incluant les utilisateurs de manière participative) de modèles concernant le système en jeu pour en prédire ou en anticiper les évolutions sous différents scénarios de gestion. En fonction du domaine considéré (physique, météorologique, social, È), ces simulations peuvent générer de grandes masses de données qui peuvent faire appel aux communautés précédemment citées pour en analyser le contenu.

- La théorie de la décision, quant à elle, vise à modéliser des situations en utilisant les outils des mathématiques appliquées. Parmi les outils, citons en systèmes interactifs d'aide à la décision, l'aide à la décision qui est associée à l'analyse multicritères. Une telle analyse vise à répondre à différents objectifs le plus souvent contradictoires afin d'aider à prendre une décision ou à évaluer plusieurs options dans des situations où aucune solution ne s'impose aux autres. Certaines méthodes reposent sur l'optimisation d'indicateurs ou sur une agrégation partielle, totale, locale ou itérative.

Après le succès de sa première édition qui s'était déroulée dans le cadre d'EGC 2012 à Bordeaux, cet atelier vise à recréer un espace de rencontres, d'échanges, de réflexions pour des chercheurs se positionnant selon les points de vue (domaines) précédemment cités, et ce avec les objectifs suivants :

1. Prendre connaissance des différentes définitions partagées ou non, des concepts associés et faire le point sur les avancées dans ces différents domaines ;

2. Tirer profit des retours d'expérience dans les différents domaines ;

3. Faire émerger des synergies entre ces différents domaines afin de permettre des collaborations futures pour proposer des solutions intégrées et innovantes dans le cadre de l'aide à la décision.

Les grandes thématiques d'intérêt pour la sollicitation de soumissions pour cet atelier incluaient :

- Aide à la décision
- Décision et entrepôts de données
- Prise de décision dans le processus de fouille de données
- Extraction de connaissances actionnables
- Représentation des modèles de décision
- Théorie de la décision
- Processus de décision et décideur
- Processus collaboratif de décision et décisions collectives
- Décision et réseaux sociaux
- Négociation et décision
- Simulation pour la décision
- Analyse multicritères pour la décision
- Passage à l'échelle et décision (Cloud, Big Data)
- Retours d'expériences sur le processus de décision
- Décisions dans tous les domaines

Processus de sélection

Lors de cette deuxième édition, 9 propositions ont été soumises et ont été chacune relue par trois évaluateurs. Parmi ces 9 propositions, 3 font l'objet d'une présentation longue et 2 d'une présentation courte.

Remerciements

Les responsables de l'atelier souhaitent remercier vivement toutes les personnes ayant contribué à la tenue de cet atelier. En particulier :

- les auteurs pour la qualité de leurs contributions constituant la base essentielle de discussions fructueuses,
- les membres du comité de lecture dont le travail d'évaluation était crucial pour assurer la qualité de l'atelier,
- les organisateurs d'EGC 2013 qui ont mis en place l'environnement et les moyens pour la réussite des ateliers.

Frédéric AMBLARD
IRIT, Université Toulouse I

Cécile FAVRE
ERIC, Université Lyon 2

Franck RAVAT
IRIT, Université Toulouse I

Membres du comité de lecture

Le Comité de Lecture est constitué de:

Fadila BENTAYEB, Université Lyon 2
Stéphane BONNEVAY, Université Lyon 1
Bernard ESPINASSE, Université Aix-Marseille 3
Nouria HARBI, Université Lyon 2
Jean-Paul JAMONT, Université Pierre Mendès France Grenoble, IUT de Valence
Bertrand JOUVE, Université Lyon 2
Christine LARGERON, Université Jean Monnet Saint Etienne
Thomas LOUAIL, Université Toulouse 1 (relecteur additionnel)
Sabine LOUDCHER, Université Lyon 2
Patrick MARCEL, Université François Rabelais de Tours
Elsa NEGRE, Université Paris Dauphine
François PINET, IRSTEA, Clermont-Ferrand
Yoann PITARCH, Université Lyon 1
Christophe PRIEUR, Université Paris Diderot
Olivier TESTE, Université Toulouse 3
Ronan TOURNIER, Université Toulouse 1
Alexis TSOUKIAS, Université Paris Dauphine
Laurent VERCOUTER, INSA Rouen
Pascal ZARATE, Université Toulouse 1
Esteban ZIMANYI, Université Libre de Bruxelles

TABLE DES MATIÈRES

Les entrepôts de données pour les nuls... ou pas !

Cécile Favre*, Fadila Bentayeb*, Omar Boussaid*, Jérôme Darmont*,
Gérald Gavin**, Nouria Harbi*, Nadia Kabachi**, Sabine Loudcher*

Université de Lyon
*ERIC - Lyon 2
{prenom.nom}@univ-lyon2.fr
**ERIC - Lyon 1
{prenom.nom}@univ-lyon1.fr

Résumé. Dans cet article, nous portons notre regard sur l'aide à la décision du point de vue des systèmes décisionnels au sens des entrepôts de données et de l'analyse en ligne. Après avoir défini les concepts qui sous-tendent ces systèmes, nous nous proposons d'aborder les problématiques de recherche qui leur sont liées selon quatre points de vue : les données, les environnements de stockage, les utilisateurs et la sécurité.

1 Introduction

Le processus décisionnel ou les systèmes décisionnels au sens des entrepôts de données sont nés d'un besoin exprimé par les entreprises qui n'était pas satisfait par les systèmes traditionnels de bases de données. En intégrant la technologie des entrepôts de données (*data warehouses*), le processus décisionnel apporte une réponse au problème de la croissance continuelle des données pouvant être de formats différents. De plus, il supporte efficacement les processus d'analyse en ligne (*On-Line Analytical Processing* - OLAP) (Chaudhuri et Dayal, 1997; Chaudhuri et al., 2011).

L'entreposage de données est donc né dans les entreprises. Ainsi, les "grands comptes" sont les principaux utilisateurs de ces technologies qui font partie intégrante de l'entreprise comme outil d'aide à la décision (le terme de *Business Intelligence* est aussi largement utilisé). Nous pouvons citer les secteurs de la grande distribution, des banques et des assurances, ainsi que ceux de l'automobile et des institutions médicales. Mais bien au-delà, l'entreposage de données suscite de plus en plus d'intérêt, avec une ouverture vers des entreprises plus petites mais qui peuvent tirer parti aujourd'hui de ces outils. Notons aussi que plusieurs domaines d'application ont vu le jour autour du Web, des systèmes d'informations géographiques, des flux de données, etc. Le Web est par ailleurs devenu une source de données à part entière.

Dans cet article, nous nous attachons à aborder la thématique de l'aide à la décision au travers du prisme de ces systèmes décisionnels en exposant leur fonctionnement, en faisant état des travaux de recherche réalisés. Mais il s'agit aussi de tenter de cerner les enjeux des recherches futures dans ce domaine par rapport à l'évolution du contexte actuel, et ce aux niveaux technologique et économique en particulier avec le succès de l'informatique dans le

nuage (*Cloud Computing*) et des outils libres (*Open Source*) entre autres. En effet la prolifération des outils libres et la possibilité de délocaliser les données dans le nuage ouvre un accès à ce processus décisionnel à un plus grand nombre d'utilisateurs et crée de nouveaux verrous scientifiques.

Cet article est organisé de la façon suivante. Dans un premier temps, nous définissons les concepts clés du domaine des entrepôts de données et de l'analyse en ligne dans la section 2. Nous abordons ensuite les quatre volets qui nous apparaissent cruciaux, à savoir les données (section 3), les environnements de stockage de ces données (section 4), les utilisateurs (section 5) et la sécurité (section 6), en détaillant pour chacun de ces volets les tendances qui se dessinent pour l'avenir. Nous concluons finalement dans la section 7.

2 L'informatique décisionnelle dans tous ses états

2.1 Préambule

Contrairement à certains processus fondés uniquement sur l'utilisation d'outils logiciels, un processus décisionnel est un projet qui se construit. Il doit s'insérer dans un cadre pouvant prendre en compte des données, des informations et des connaissances. L'approche d'entreposage de données ("data warehousing") constitue un champ de recherche important dans lequel de nombreux problèmes restent à résoudre. Les entrepôts de données sont généralement intégrés dans un système d'aide à la prise de décision où l'on distingue deux espaces de stockage : l'entrepôt de données et les magasins de données. Une architecture du processus décisionnel est représentée dans la Figure 1 (Bentayeb et al., 2009).

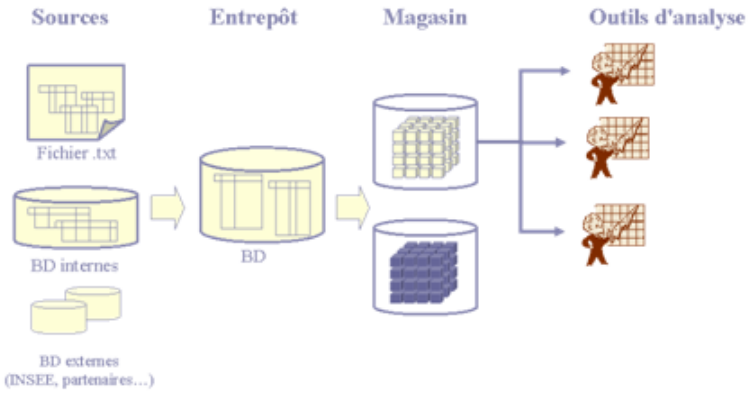


FIG. 1 – *Architecture générale d'un système décisionnel.*

Plusieurs auteurs ont défini le concept d'entrepôt de données. Selon Inmon (1996), c'est une collection de données orientée sujets, intégrée, non volatile et en mode de lecture seule, importée à partir de sources de données hétérogènes et stockée à différents niveaux de granularité dans un but de prise de décision. Ainsi, un entrepôt de données est généralement vu comme

un espace de stockage centralisé regroupant dans un format homogène les données issues de différentes sources, qui peuvent faire l'objet de transformations et d'historisation, à des fins d'analyse pour la prise de décision. Un magasin de données peut constituer un extrait de l'entrepôt, où les données sont préparées de manière spécifique pour faciliter leur analyse et leur exploitation par un groupe d'utilisateurs, en fonction par exemple d'une orientation métier.

Finalement, les possibilités d'analyse des données sélectionnées sont très variées. Elles dépendent des besoins des utilisateurs et font appel à des techniques différentes :

- le *reporting* avec la construction de tableaux de bord, d'indicateurs, de graphiques ;
- la navigation multidimensionnelle dans les données avec la technologie OLAP ;
- la fouille dans les données à l'aide des méthodes de *Data Mining*.

2.2 Modélisation et alimentation de l'entrepôt

2.2.1 Modélisation multidimensionnelle

Les modèles multidimensionnels ont pour objectif de proposer un accès aux données intuitif et très performant. Pour cela, les données sont organisées autour des faits que l'on cherche à analyser, caractérisés à l'aide d'indicateurs (appelés mesures) qui sont des données normalement numériques et additives, permettant de mesurer l'activité modélisée. Ces faits sont décrits par un ensemble d'axes d'analyse, ou dimensions, d'où le terme de modèle multidimensionnel.

Ce modèle de base correspond au modèle en étoile (Kimball et al., 2000; Chaudhuri et Dayal, 1997). Citons l'exemple classique de faits concernant des ventes, dont les mesures sont la quantité commandée et le prix correspondant. Les dimensions (clients, produits concernés, dates, etc.) sont des descripteurs des faits de vente. Ainsi, pour un client donné, un produit, une date, nous disposons de la quantité commandée et du prix correspondant.

Si l'on considère une implémentation en relationnel (ROLAP), les faits seront dans une table (table de faits) et chacune des dimensions sera dans une table respectivement (tables de dimension), chacune étant reliée à la table des faits. Les avantages de ce modèle sont la facilité de navigation, grâce à la table de faits centrale, de bonnes performances en raison du faible nombre de jointures à effectuer pour l'analyse sur une dimension donnée et des agrégations faciles des mesures.

La modélisation en flocons est une première variante du modèle en étoile. Il consiste à décomposer les dimensions d'un modèle en étoile en des hiérarchies explicites, chacun des niveaux de la hiérarchie correspondant à une table dans une implémentation ROLAP. Cette modélisation permet de réduire le volume de stockage et autorise des analyses par paliers sur la dimension hiérarchisée. En revanche, les jointures nécessaires pour accéder aux données entraînent une dégradation des performances.

Finalement, la modélisation en constellation consiste à fusionner plusieurs modèles en flocons, permettant le partage de certaines dimensions par plusieurs ensemble de faits.

2.2.2 Alimentation

L'alimentation d'un entrepôt de données est une phase essentielle dans le processus d'entreposage. Elle se déroule en plusieurs étapes : extraction, transformation, chargement et rafraîchissement des données, qui sont prises en charge par le processus d'ETL (*Extracting, Transforming and Loading*). Ce processus constitue la phase de migration des données de production

dans le système décisionnel après qu'elles ont subi des opérations de sélection, de nettoyage et de reformatage dans le but de les homogénéiser. Cette phase constitue une étape importante et très chronophage dans la mesure où on l'estime à environ 80% du temps de mise en place de la solution décisionnelle. Ainsi cette phase fait l'objet de nombreux travaux de recherche, en terme de modélisation, d'automatisation du processus (Simitsis et al., 2010; Jovanovic et al., 2012; Papastefanatos et al., 2012; Akkaoui et al., 2011; Muñoz et al., 2009).

2.3 Analyse en ligne

L'analyse en ligne constitue un autre aspect du processus d'entreposage des données. Codd (1993) a défini l'OLAP comme "l'analyse dynamique d'une entreprise qui est requise pour créer, manipuler, animer et synthétiser l'information des modèles d'analyse de données. Cela inclut la capacité à discerner des relations nouvelles ou non anticipées entre les variables, la capacité à identifier les paramètres nécessaires pour traiter des grosses quantités de données, la création d'un nombre illimité de dimensions". Un système OLAP est un dispositif muni d'opérateurs spécifiques permettant l'analyse en ligne des données. Il est également considéré comme un serveur d'applications pouvant traiter directement les données d'un entrepôt ou pouvant être utilisé comme un outil d'exploration de données grâce à une navigation interactive. Les applications OLAP permettent entre autres de travailler sur des données historiques pour étudier les tendances ou les prévisions d'une activité, ou de travailler sur des données récapitulatives pour créer de l'information stratégique pour la prise de décision. L'analyse en ligne peut aussi bien s'appliquer aux données de l'entrepôt qu'à celles d'un magasin de données. Généralement, elle est plutôt effectuée sur une collection de données encore plus fine appelée cube de données.

2.3.1 Cubes de données

Le modèle multidimensionnel permet d'organiser les données selon des axes représentant des éléments essentiels de l'activité d'une entreprise. Trois niveaux de représentation des données sont définis dans le processus décisionnel : l'entrepôt qui regroupe des données transversales à l'ensemble des métiers de l'entreprise, le magasin de données qui est une représentation verticale des données portant sur un métier particulier et enfin le cube de données (ou hypercube). Le cube correspond à une vue métier où l'analyste choisit les mesures à observer selon certaines dimensions. Un cube est une collection de données agrégées et consolidées pour résumer l'information et expliquer la pertinence d'une observation. Le cube de données est exploré à l'aide de nombreuses opérations qui permettent sa manipulation.

2.3.2 Opérateurs OLAP

De manière générale, il existe deux classes d'opérations. La première, liée à la structure des données, permet de la manipuler pour mettre en relief la pertinence de certaines informations. Les opérations de manipulation des données multidimensionnelles permettent de réorienter la vue multidimensionnelle ou d'en changer l'agencement en agissant sur la position des membres des dimensions et des mesures : rotation (*rotate*), permutation (*switch*) , division (*split*), emboitement (*nest*), enfoncement (*push*) et retrait (*pull*) . La deuxième classe d'opérations est liée à la granularité des données. Ces opérations agrègent et résument les données

ou les détaillent et permettent une analyse par paliers : agrégation (*roll up*), forage vers le bas (*drill down*). Dans ce cas, on a recours à une opération d'agrégation qui est appliquée sur la (ou les) mesure(s) étudiée(s) (somme, moyenne, max, min, etc.). Ces deux derniers opérateurs sont largement évoqués dans les travaux de recherche contrairement à ceux de la première catégorie. En effet, ils se basent sur les hiérarchies et soulèvent donc les problèmes de complexité des hiérarchies à modéliser (Malinowski et Zimányi, 2004) et d'additivité des données (Mazón et al., 2009).

2.4 Un point sur le positionnement par rapport aux bases de données classiques

Généralement, le processus décisionnel est basé sur un entrepôt de données qui constitue son élément central. Il est alors intéressant de comprendre ce qu'est ce concept de stockage des données et de le positionner par rapport aux bases de données classiques.

La règle-clef du développement d'une base de données traditionnelle est d'optimiser le traitement efficace d'un ensemble de transactions. En effet, les bases de données classiques sont dites transactionnelles car elles sont conçues pour des opérations quotidiennes. Ces transactions nécessitent des données détaillées et actualisées. Elles lisent ou mettent à jour des enregistrements accessibles par leur identifiant. Elles sont conçues pour refléter une sémantique plutôt opérationnelle en minimisant les conflits et en garantissant la persistance des données avec un minimum de redondance et un maximum de contrôle d'intégrité. Les requêtes visent un nombre relativement peu important d'enregistrements. Le but est de mettre à jour les données pour garder une trace des événements de l'entreprise. Ces bases de données sont qualifiées alors de production. Elles sont orientées vers des applications de type OLTP (*On-Line Transactional Processing*).

OLAP, autrement dit l'analyse en ligne, est une démarche orientée "aide à la décision". Les données sont stockées dans un entrepôt de données, où elles sont historisées, résumées et consolidées. Le volume de données des entrepôts est important et va de centaines de gigaoctets à des téraoctets, voire même encore davantage de nos jours. Les entrepôts de données stockent des données collectées au cours du temps, en provenance de plusieurs bases de données opérationnelles. Le temps de réponse doit être court. Pour cela, il est nécessaire d'agréger les données afin d'apporter des réponses rapides à des requêtes pouvant être posées à de multiples niveaux. Il est nécessaire d'optimiser les requêtes les plus fréquemment utilisées afin d'améliorer les temps de réponse. Divers travaux se sont intéressés à cette question de l'optimisation de performances qui est cruciale dans ce contexte d'analyse en ligne. Un entrepôt de données vise à répondre à un utilisateur en termes d'informations et non en termes d'applications (Franco, 1997). Ainsi les systèmes transactionnels et les systèmes d'analyse en ligne ne peuvent coexister dans un même environnement de base de données de par leurs caractéristiques différentes (Codd, 1993), même si un entrepôt de données peut être stocké de manière relationnelle.

2.5 Outils

Le domaine des entrepôts de données est né dans les entreprises. Et c'est aujourd'hui un secteur en pleine expansion avec de nombreux projets décisionnels qui se construisent. La question de la mesure du retour sur investissement se pose alors. Le recours à des technologies de type "*Open Source*" peut constituer une alternative au coût de mise en place de tels projets.

Les outils proposés actuellement sont de plus en plus nombreux également et il est souvent difficile de s'y retrouver. L'objectif n'est pas ici d'en faire une liste exhaustive. Notons d'ailleurs, l'intérêt d'un éventuel travail qui viserait à recenser et positionner (cartographier) tous ces outils, un tel travail étant pour le moment inexistant à notre connaissance, malgré son intérêt indéniable.

Nous pouvons distinguer les outils selon ce qu'ils couvrent comme fonctionnalités. Citons par exemple les deux ETL *Open Source* les plus connus : Kettle (Pentaho Data Integration) et Talend.

Mentionnons également les moteurs OLAP tels que Mondrian (Open Source) qui permettent, à partir d'en entrepôt stocké dans un système de gestion de bases de données relationnelles, de pouvoir construire les cubes de données, qui peuvent être ensuite interfacés avec des applications de visualisation (telles que JPivot, Pentaho Analyzer, Pentaho Analysis Tool, Geo Analysis Tool, etc.)

Nous pouvons également citer d'autres outils connus qui sont dédiés au reporting tels que JasperSoft (Open Source), QlikView, BusinessObject.

3 Des données à tous les niveaux

3.1 Complexité des données

Les entrepôts de données et l'OLAP sont des technologies relativement bien maîtrisées quand il s'agit de données "simples". Cependant, la communauté scientifique s'accorde pour dire que, avec l'avènement du Web et la profusion des données multimédias (son, image, vidéo, texte...), les données sont de plus en plus hétérogènes, diverses et qu'elles sont devenues complexes. L'avènement des données complexes a remis en cause le processus d'entreposage et d'analyse des données ; il a induit l'émergence de nouveaux problèmes de recherche comme l'intégration des données complexes dans les entrepôts, le stockage, la représentation ou la modélisation, l'analyse en ligne et la fouille de telles données.

L'informatique décisionnelle tente de s'adapter aux données complexes depuis plusieurs années. De nombreuses adaptations ou évolutions pourraient être citées. Par exemple, les opérateurs OLAP, comme celui d'agrégation (souvent basé sur la somme ou la moyenne), sont définis pour des données classiques (numériques) et ils deviennent inadaptés quand il s'agit de données complexes, par exemple composées de textes, d'images, de sons ou de vidéos. Plusieurs équipes de recherche travaillent sur ce problème clef d'agrégation des données complexes, par exemple textuelles (Ravat et al., 2008), ou images (Jin et al., 2010). D'autres équipes travaillent sur l'association des Systèmes d'Information Géographique, des entrepôts de données et de l'analyse OLAP pour créer le SOLAP (*Spatial OLAP*) (Bédard et Han, 2009). Les données spatiales sont une forme de données complexes. En effet, dans un cube de données spatiales, les dimensions et les mesures peuvent contenir des composantes spatiales ou géométriques. Un autre exemple de données complexes est celui des flux de données (*data stream*). Dans ces flux, les analystes souhaitent détecter des changements dynamiques par une analyse en ligne. On parle de fouille de flots de données multidimensionnelles, d'*OLAPing Stream Data* ou de *Stream cube* (Han et al., 2005). Enfin, le *XOLAP* (ou *XML OLAP*) cherche à faire des analyses OLAP sur des documents XML tout en tenant compte de leurs spécificités (hiérarchies multiples, imbriquées, incomplètes...) (Wang et al., 2005).

Ces déclinaisons de l'OLAP sont des exemples d'adaptation des entrepôts de données et de l'OLAP aux différents types de données, mais elles ne portent souvent que sur la structure des données et non pas sur leur contenu. Une autre spécificité des données complexes réside dans la sémantique qu'elles véhiculent. Par conséquent, un nouveau problème émerge : comment prendre en compte la sémantique contenue dans les données complexes pour la modélisation et l'analyse ? Le recours à des solutions telles que les ontologies constitue une issue prometteuse explorée dans différents travaux (Cao et al., 2006; Selma et al., 2012).

3.2 Volume des données

Parallèlement à cette problématique de la sémantique des données, la question du volume de ces données peut également poser problème au niveau de leur requêtage en terme de performance. En effet, les requêtes décisionnelles s'appliquent sur de très grandes quantités de données. Elles nécessitent pourtant des temps de réponse ne dépassant pas quelques secondes ou quelques minutes. Il existe plusieurs techniques traitant le problème de l'amélioration des performances des requêtes avec un souci constant de l'optimisation en utilisant des techniques issues des bases de données : la matérialisation des vues, l'indexation, la fragmentation, etc. (Aouiche et Darmont, 2009; Benkrid et Bellatreche, 2011) (se basant souvent sur l'exploitation d'algorithmes de fouille de données)

La production croissante de données, le partage des informations entre utilisateurs, la diffusion des données via les réseaux engendrent de très gros volumes de données disponibles et intéressantes à analyser. L'expression anglaise *Big Data* est utilisée pour désigner des données dont le volume est tel qu'il devient difficile de les stocker, de les interroger, de les modéliser, de les analyser et de les visualiser avec les outils et architectures informatiques existants, du fait également de leur manque de structure.

En effet, la prolifération de très grandes quantités de données, produites principalement par le Web, notamment par les grands acteurs d'Internet et les réseaux sociaux, engendre des évolutions technologiques qui posent de réels problèmes scientifiques. Les volumes de données à très grandes échelles nécessitent des moyens de stockage appropriés (Agrawal et al., 2011). L'utilisation de nouvelles unités de mesures de stockage, telles que les peta-octets voire les zeta-octets sont aujourd'hui des réalités. Outre le stockage, l'exploitation de telles données soulève également de nouveaux challenges scientifiques. De nombreux travaux de recherche proposent aujourd'hui des solutions de gestion de données à très grande échelle. Disposer en ligne de plus en plus de données historisées pour l'analyse est un besoin réel pour les grands acteurs d'Internet ainsi que pour d'autres entreprises, entrainant une expansion des bases de données orientées analyse, tels que les entrepôts de données. L'informatique dans le nuage tente d'apporter des réponses à ces problèmes.

4 Environnement de stockage

L'informatique décisionnelle (*Business Intelligence*) a beaucoup évolué depuis une trentaine d'années passant d'une discipline exclusivement réservée à un groupe d'utilisateurs, les décideurs, pour se démocratiser en délocalisant la prise de décision du haut de la pyramide au plus proche du terrain pour une meilleure réactivité. L'enjeu est de disposer de la bonne

information afin de délivrer la bonne connaissance à la bonne personne. Cela passe par le déploiement d'un environnement de stockage qui doit permettre de rendre accessible, de mettre en forme et de présenter les informations clés aux différents utilisateurs concernés afin de faciliter la prise de décision.

4.1 Au-delà du relationnel, les entrepôts continuent

Comme nous l'avons vu, l'architecture d'un système décisionnel est généralement vue comme une architecture à trois niveaux :

- les sources d'information qui correspondent à l'ensemble des bases de données de production et sites dont sont extraites les informations ;
- l'entrepôt qui contient l'ensemble des données extraites de ces sources ;
- les magasins extraits de l'entrepôt et dédiés aux différentes classes de décideurs.

Les sources d'informations utiles aux décideurs peuvent être stockées sur des sites de nature diverse (sites Web, bases de données...). Cependant, avec l'avènement des données très volumineuses, peu ou pas structurées (*Big Data*), le monde traditionnel des bases de données relationnelles, support des entrepôts de données, n'est plus adapté pour gérer et traiter ces grandes masses de données de type texte, image, etc. provenant du Web, des publications sur les média sociaux, les logs des serveurs Web et des applications, etc. Pour faire face à ces énormes volumes de données, de nouvelles technologies sont apparues comme Hadoop, MapReduce ou les bases de données NoSQL (*Not only SQL*) (Cattell, 2011; Leavitt, 2010). Pour autant, est-ce que l'émergence de ces nouvelles technologies *Big Data* signe la fin des entrepôts de données ? Nous pensons que les bases de données NoSQL n'ont pas la vocation de remplacer les bases de données relationnelles, mais de les compléter selon les besoins des entreprises en proposant une alternative pour adapter le fonctionnement des bases de données à des besoins spécifiques.

Le terme NoSQL fait en fait référence à une diversité d'approches, classées en quatre catégories de bases de données : les bases de données orientées colonnes (comme MonetDB [1]) , les bases de données orientées graphes (comme Neo4J [2]), les bases de données orientées clé/valeur (comme Riak [3]) et les bases de données orientées documents (comme MongoDB [4]). Les différents systèmes de gestion de bases de données qui supportent les bases de données NoSQL sont destinés à manipuler de gigantesques bases de données pour des sites Web tels que Google, Amazon, ou Facebook. En abandonnant les propriétés ACID (Atomicité, Cohérence, Isolation et Durabilité) des bases de données relationnelles, les bases de données NoSQL permettent une montée en charge élevée et assurent une grande performance.

L'architecture décisionnelle "traditionnelle" avec sa base de données centralisée n'est donc plus l'unique architecture de référence. En effet, nous pensons que dans un contexte *Big Data*, il est important de mettre en place d'autres architectures décisionnelles, notamment pour la prise en compte à la fois de données peu ou pas structurées et le passage à l'échelle.

1. http ://www.monetdb.org/
2. http ://neo4j.org/
3. http ://basho.com/
4. http ://www.mongodb.org/

4.2 Jusque dans les nuages

Les solutions de bases de données orientées analyses doivent vérifier les mêmes propriétés que celles des environnements dans le nuage, à savoir : fiabilité, évolutivité, sécurité, bonne performance, tolérance aux pannes, capacité de fonctionner dans un environnement hétérogène, flexibilité de requêtes...

Les problèmes liés aux entrepôts de données et à l'analyse en ligne (OLAP) sont à réétudier dans le cadre des environnements de *Cloud Computing* et cela augure des perspectives prometteuses de combinaisons de ces deux technologies. Entreposer des données à très grande échelle suppose des moyens de traitements à grande échelle également. Le *Cloud Computing* offre ces moyens grâce à une association de plusieurs clusters regroupant un très grand nombre d'ordinateurs. Cependant, le recours à une telle infrastructure n'est pas gratuit. Il fonctionne selon un modèle de facturation à l'utilisation. Ceci engendre un ensemble de problèmes scientifiques à étudier pour mettre au point des approches techniquement et économiquement viables.

Par ailleurs, l'un des points cruciaux à prendre en charge porte sur la virtualisation des données. C'est un problème ouvert. La répartition, la réplication et la distribution des données à travers les nœuds des clusters nécessitent des modèles de données appropriés aux environnements du *Cloud*. Ceux-ci doivent permettre à l'utilisateur de ne voir que ses données.

Un autre point crucial à considérer porte sur les traitements des données. Il existe déjà des travaux dans la littérature, dont certains préconisent des approches basées sur les bases de données parallèles privilégiant les performances (Abouzeid et al., 2009). D'autres sont plus favorables à des solutions utilisant le paradigme de *MapReduce*, mettant en avant son adéquation avec des traitements répartis sur des données distribuées (Stonebraker et al., 2010). Cependant, *MapReduce* est plutôt adapté pour les données non structurées et s'illustre par sa congruence à des environnements tels le *Cloud*. Cependant, le traitement de requêtes réparties sur plusieurs nœuds ainsi que l'équilibrage des charges (requêtes) et des données sur les différents nœuds sont de réels challenges. L'apparition de nouveaux nœuds peut impliquer des changements de stratégie de répartition, de réplication et de distribution des données et des traitements. Ceci demeure un problème ouvert. La conception de démarches, utilisant les deux techniques de parallélisation et de partitionnement des données, constitue certainement une perspective prometteuse pour les entrepôts de données dans le *Cloud*.

Construire des entrepôts de données sur le *Cloud* devrait tenir compte des contraintes de ce dernier et plus particulièrement de la tarification de l'usage des ressources. Il s'agit de la notion d'élasticité qui constitue un argument financier convaincant. L'utilisateur peut demander des ressources selon ses préférences. Il peut avoir besoin soit de hautes performances avec des prix élevés, soit de basses performances avec un prix moindre. Du fait de l'hétérogénéité des ressources, il faut lui laisser la possibilité de louer des ressources sur mesure. Pour cela, il faut définir des métriques pour mieux évaluer et décider des performances des ressources à utiliser. Ces objectifs sont également des challenges à relever même si le déploiement d'un entrepôt sur le *Cloud* doit être totalement automatisé.

La construction de modèle de coûts est également un objectif important du fait que la construction d'entrepôts de données sur le *Cloud* ne porte pas seulement sur des aspects techniques, la dimension économique représente un point crucial. Dans les environnements de *Cloud*, les vitesses de communication (via LAN) peuvent être irrégulières selon la proximité des nœuds les uns des autres et l'architecture des réseaux. Ceci peut avoir un impact sur les

transferts de très grands volumes de données qui peuvent s'exprimer en tera-octets, voire en peta-octets. Ceci nécessite alors des techniques de compression de données.

Une partie des problèmes de recherche classiques qui se posent encore dans le domaine des entrepôts de données trouve une nouvelle expression lorsque l'on se situe dans le nuage. Faut-il continuer de dénormaliser les modèles physiques dans un cadre NoSQL pour bénéficier de meilleures performances, demeurer dans un environnement SQL qui garantit l'intégrité des données, ou encore tenter de travailler intégralement en mémoire vive ? L'élasticité est-elle la réponse à tous les problèmes de performance, ou ne vaut-il pas mieux adapter des techniques d'optimisation bien connues (index, vues matérialisées...) pour minimiser le coût en ressources (et donc, monétaire) des requêtes dans le nuage (Nguyen et al., 2012) ? Doit-on inclure dans la notion d'élasticité la prise en compte des données situationnelles (Pedersen, 2010) et les problèmes d'intégration des données qui en découlent ? De plus, travailler au moins en partie à partir de données situationnelles impose d'accepter une perte de contrôle sur les données du système décisionnel, notamment sur leur fiabilité et leur pérénité, et donc de se contenter d'analyses de tendances plutôt que d'historiques avérés (Middelfart, 2012). Evaluer ce degré de contrôle est donc important. L'étude de Kandel et al. (2012) constitue un point de départ tout à fait intéressant pour ces réflexions.

D'autres problèmes sont davantage liés au paradigme de l'informatique dans le nuage et aux usages décisionnels plus personnels et collaboratifs qu'il permet. Par exemple, classiquement, l'investissement (en général très important) dans un système décisionnel doit être effectué a priori par les entreprises. En revanche, dans le nuage, la construction d'un système décisionnel peut être incrémentale, collaborative et exploiter au mieux le paiement à la demande (Darmont et al., 2012). Il est tout à fait possible de "partir petit", voire de "rester petit", d'adaptant à la cible des utilisateurs. Nous abordons alors à présent ce volet utilisateurs si crucial pour des systèmes qui, par définition, sont centrés utilisateurs.

5 Des entrepôts pour tous : utilisateurs à tous les étages

L'informatique décisionnelle, en raison des architectures matérielles, logicielles et des compétences requises, n'a longtemps été accessible qu'aux grandes entreprises. Pourtant, les besoins en décisionnel existent dans de plus petites structures, que ce soient des PME (Petites et Moyennes Entreprises) ou TPE (Très Petites Entreprises), des ONG (Organisations Non Gouvernementales), des associations, des communautés en ligne ou même de simples citoyens (les indignés espagnols ont, par exemple, exprimé une forte demande de données publiques ouvertes). Pour ce type d'utilisateurs, des solutions bon marché, légères, faciles à utiliser, flexibles et rapides, sont nécessaires (Grabova et al., 2010). Avec l'avènement de l'informatique dans le nuage, le décisionnel à la demande (*cloud BI*, *personal BI*, *self-service BI*, *on-demand BI* ou encore *collaborative BI*, dans la terminologie anglo-saxonne encore non standardisée, avec *BI* pour *Business Intelligence*) sous forme de service est devenu possible et accessible avec un simple navigateur Web depuis une tablette ou un smartphone. Ce nouveau type de services en ligne doit permettre à des utilisateurs non-experts de prendre des décisions éclairées en enrichissant le processus décisionnel par des données situationnelles, c'est-à-dire très ciblées, de portée limitée dans le temps et pertinentes pour un petit groupe d'utilisateurs (Abello et al., 2013), soit typiquement des données glânées sur le Web.

A l'heure où l'informatisation tend à diminuer les relations inter-personnes, dans la mesure où beaucoup de ces relations se transforment en relations homme-machine, le besoin d'"humaniser" les systèmes se fait ressentir pour permettre le processus d'aide à la décision. Cette humanisation nécessite de rendre l'interaction système-utilisateur plus personnelle, afin d'assurer l'adaptation de l'informatique aux utilisateurs, avec pour objectif de répondre à leurs propres besoins. Ceci passe donc initialement par une conception de l'entrepôt de données où les utilisateurs finaux sont considérés.

5.1 Implication de l'utilisateur dans le processus décisionnel

L'un des points clés de l'entreposage de données réside dans la conception du schéma de l'entrepôt. En effet, les possibilités d'analyse sont conditionnées par ce dernier. Il est donc important que les utilisateurs soient impliqués dans la conception de l'entrepôt pour une bonne prise en compte de leurs besoins d'analyse.

Dans un second temps, pour permettre un processus décisionnel centré utilisateurs, la prise en compte de leurs préférences et de leurs caractéristiques à travers un profil constitue une piste intéressante. Dans l'exploitation des données, il s'agit alors de proposer la personnalisation du système (visualisation des données, par exemple), et la recommandation, par rapport à une aide à la navigation dans les données (Aligon et al., 2011), qui permet à terme une aide à la décision. En effet, par rapport au volume considérable de données, l'accès à une information pertinente devient un enjeu crucial pour l'utilisateur. Mais au-delà de cet aspect, il s'agit aussi pour l'utilisateur d'avoir l'impression que le système informatique ait été fait pour lui et qu'il s'adresse à lui "personnellement".

Par ailleurs, si les outils méthodologiques et technologiques permettant de mettre en œuvre des solutions décisionnelles à la demande existent depuis quelques années (entrepôts de données Web, de documents, de données XML, logiciels ETL et OLAP libres, systèmes de gestion de bases de données en mémoire vive... (Grabova et al., 2010)), le tout premier service a été le prototype Google Fusion Tables (Gonzalez et al., 2010). Ce dernier permet d'intégrer des données privées et situationnelles dans un tableur simple, de les visualiser, de les analyser et de les partager de façon très intuitive. Les applications en ligne de nombreux éditeurs de solutions décisionnelles proposent désormais également ce type de fonctionnalités. De plus, il a été proposé d'étendre le principe de fusion de tables à des cubes de données (*fusion cubes* dont le schéma et les instances peuvent être modifiés à la volée et qui intègrent des données situationnelles ainsi que les métadonnées décrivant leur provenance et leur qualité (Abello et al., 2013).

D'un point de vue technique, Essaidi (2010) a proposé une plateforme décisionnelle à la demande. Toutefois, si cette plateforme est bien disponible en tant que service dans le nuage (en mode SaaS : *Software as a Service*), l'intégration dynamique de données situationnelles n'est pas mentionnée. Thiele et Lehner (2011) proposent une solution à ce problème en combinant des données existantes chez l'utilisateur à des services Web qui créent de nouveaux contenus à partir de sources externes. Ainsi, le processus habituel d'ETL est conduit par l'utilisateur lui-même, de façon interactive. Toutefois, il n'y a aucune garantie quant à la qualité et à l'intégrité des données recueillies. Pour cela, il est toutefois possible d'utiliser les travaux de Jörg et Dessloch (2009), qui garantissent l'intégrité d'un entrepôt quand les données sources sont fournies avec une faible latence, comme c'est le cas pour des données situationnelles.

Enfin, l'aspect collaboratif du décisionnel est apparu dès 2007, avec l'annotation de cubes pour modéliser et permettre le partage de l'expertise des utilisateurs d'OLAP (Cabanac et al., 2007). Une architecture décisionnelle collaborative a ensuite été proposée par Berthold et al. (2010), qui inclut des fonctionnalités dites sociales afin d'enrichir le processus de décision grâce aux opinions d'experts. Une dernière approche répartit des magasins de données dans une architecture pair à pair (Golfarelli et al., 2012). Bien que le processus de décision soit amélioré dynamiquement grâce au partage de connaissances dans toutes ces approches, l'intégration de données situationnelles à la volée n'y est pas envisagée.

5.2 La visualisation pour aider l'utilisateur à décider

La phase d'analyse de données est bien évidemment cruciale par rapport à l'aide à la décision et au pilotage. Ainsi, la production de tableaux de bord et la visualisation interactive de l'information constituent des étapes phares, d'autant plus que l'exploration de données massives est un problème difficile, en particulier pour l'œil humain.

Nous pouvons distinguer deux types de travaux de recherche dans ce domaine : les travaux sur la visualisation elle-même et les besoins émergents par rappport aux nouveaux supports de communication.

Le premier porte sur l'amélioration de la visualisation par des algorithmes. Ainsi, on a vu se développer la combinaison de l'analyse en ligne avec des techniques de fouille de données (Messaoud et al., 2006). Et les chercheurs spécialisés en visualisation commencent à s'intéresser au domaine de l'OLAP. Citons en particulier la possibilité de navigation OLAP en 3D (Sureau et al., 2009).

Parallèlement, l'évolution technologique en matière de support modifie considérablement le rapport des utilisateurs à la visualisation de données. Il est nécessaire de considérer l'adaptation d'outils d'analyse aux nouveaux supports de diffusion. Selon K. Bornauw [5], "si nous parvenons à relever ce défi de la visualisation des données et à la rendre conviviale et accessible depuis n'importe quel appareil (ordinateur, smartphone, tablette,..), elle sera non seulement économiquement utile, mais également agréable à l'utilisateur de systèmes d'information décisionnels, qui se verra soulagé du fardeau des modèles complexes et douloureux d'exploration des données".

Si les outils traditionnels tels que Cognos ou BusinessObjects sont encore d'actualité, on a vu émerger de nouveaux outils comme Spotfire, QlikView, Tableau 7. Et l'usage de nouveaux supports a nécessité le développement d'applications spéciales par les terminaux mobiles (Roambi-ESX pour Ipad, Yellowfin, etc.). En effet, le déploiement des applications de Business Intelligence sur des terminaux mobiles complique la problématique de visualisation.

Dès l'avènement des premiers téléphones intelligents, la question de l'accès au système d'information de l'entreprise depuis tout lieu et à tout instant s'est posée. D'ici 2014, les accès internet seront majoritairement mobiles. Les applications mobiles transforment la communication et donc l'organisation même des entreprises, commente Benoit Herr [6], l'auteur de l'étude. L'essor des téléphones intelligents et autres tablettes modifient les usages de la *Business Intelligence* puisque l'accès à distance au système d'information depuis son terminal portable est devenu réel. Ces possibilités d'accès à distance de données à fort potentiel stratégique reposent

5. Kris Bornauw , BI Expert, EoZen, Groupe SQLI,- www.eozen.com, 2012.

6. Proginov, "Cloud, SaaS et mobilité : nouveaux outils, nouveaux usages". Mars 2012, Journal Solutions & Logiciels, N28.

bien évidemment la question de la sécurité, que nous nous proposons d'aborder dans la section suivante.

6 Sécurité

Chaque jour de nouvelles vulnérabilités sont découvertes sur tous les types de composants d'un système d'information classique, et aussi décisionnel a fortiori. Lorsqu'elles sont exploitées par des individus malveillants, elles risquent de perturber gravement le système d'information décisionnel : indisponibilité (partielle ou totale, temporaire ou prolongée), pertes de données, vol d'informations confidentielles, pertes d'exploitation, la liste n'est malheureusement pas exhaustive... La protection du système d'information décisionnel est une lutte incessante. Elle exige des administrateurs système et réseau en charge de la maintenance informatique de s'astreindre à :

- surveiller les menaces qui pèsent sur les systèmes d'information
- mettre en œuvre rapidement les parades permettant de réduire les possibilités d'attaque

Pour cela, il faut définir le périmètre de surveillance : systèmes d'exploitation ou applications, et ceci pour les équipements réseaux, serveurs, postes de travail, et périphériques. Nous constatons que la veille technologique, dans le domaine de la sécurité, concerne jusqu'à présent le suivi des nouvelles technologies disponibles sur le marché, mais concerne également le suivi des alertes de sécurité ou plus précisément des nouvelles vulnérabilités découvertes sur les systèmes informatiques.

Renforcer la sécurité des systèmes d'information décisionnelle consiste pour la plupart des acteurs à ajouter des équipements supplémentaires : serveurs, pare-feux... ou à complexifier et à sophistiquer la gestion des accès ... Nous sommes persuadés que la sécurité doit aussi être intégrée dans la phase de conception, dans les mécanismes d'architecture des entrepôts de données pour imposer des méthodes et des outils. Cette démarche permet de pallier à d'éventuelles défaillances des dispositifs mis en place au niveau des infrastructures et des systèmes de détection d'intrusions (IDS).

Les systèmes d'information décisionnels sont souvent stockés sur des machines virtuelles différentes pour des raisons de volumétrie et d'optimisation. La communication entre les différentes machines est très vulnérable. Cette faille doit être supprimée par des moyens de communication naturellement sécurisés. En considérant ces machines virtuelles comme des parties indépendantes, des primitives cryptographiques peuvent permettre de sécuriser les communications. Basée sur la cryptographie asymétrique, la signature numérique (parfois appelée signature électronique) est un mécanisme permettant de garantir l'intégrité d'un document électronique et d'en authentifier l'auteur, par analogie avec la signature manuscrite d'un document papier. Un mécanisme de signature numérique doit permettre au lecteur d'un document (une couche) d'identifier l'expéditeur (une couche) qui a apposé sa signature. Il doit garantir que le document n'a pas été altéré entre l'instant où l'auteur l'a signé et le moment où le lecteur le consulte. La confidentialité des données peut être assurée classiquement par des cryptosystèmes symétriques. Le problème qui se pose est le stockage de la clé privée. Il s'agit en effet de prémunir les systèmes contre des attaques visant à la recouvrer. Aucune solution ne permet de se prémunir totalement contre ce risque. Cependant, on assiste depuis quelques années à l'émergence de cryptosystèmes complètement homomorphiques. Ces cryptosystèmes permettent de faire des calculs sur des valeurs encryptées sans avoir à les décrypter. Ils peuvent

donc grandement limiter l'usage de la clé privée. Toutefois, ces cryptosystèmes nécessitent de grosses ressources et ne sont pas encore opérationnels en pratique.

Les questions relatives à la sécurité et à la confidentialité des données sur le *Cloud* ont été les premières préoccupations des fournisseurs et des usagers du *Cloud*. Il en est de même dans le cas des entrepôts de données dans le *Cloud*. Différents scénarios peuvent être envisagés : soit la soustraction des données sensibles de l'analyse à partir du *Cloud* ; soit l'encryptage de celles-ci. Des travaux commencent à émerger portant sur l'analyse des données encryptées. Ces questions représentent sans doute des pistes de recherche intéressantes. Cependant, elles ne sont pas les seules préoccupations, les nombreux problèmes cités ci-dessus montrent également la diversité des pistes de recherche que suscite cette nouvelle problématique des entrepôts dans le *Cloud*. Celle-ci souffre aujourd'hui d'un manque de conceptualisation du fait de son émergence récente.

Ainsi, les problèmes de sécurité intrinsèques au stockage de données dans le nuage demeurent : espionnage de la part du fournisseur de service ou d'un sous-traitant, garantie de disponibilité des données, croisements incontrôlés de données... (Chow et al., 2009). Il existe cependant des pistes de recherche prometteuses, notamment au niveau de l'anonymisation des données qui, même cryptées, restent interrogeables et utilisables dans certains traitements. Stocker des données volontairement altérées, mélangées dans le Cloud peut être aussi une possibilité pour assurer la confidentialité des données. Cette solution soulève également des questions au niveau du cryptage et décryptage pour les interrogations. De plus, le calcul multi-parties permet à des individus distincts de construire de façon collaborative un résultat d'analyse commun sans pour autant dévoiler leurs sources de données. Ces techniques de cryptographie ne sont toutefois pas encore assez matures pour permettre l'analyse en ligne ou la fouille de données, ni pour un déploiement à l'échelle du nuage. Ce dernier soulève des problèmes de temps de traitement qui pousse à ne sécuriser que certaines données : les plus sensibles, les plus récentes ...

7 Conclusion

Dans cet article, nous avons présenté le domaine de l'aide à la décision au travers du prisme des entrepôts de données et de l'analyse en ligne. Ainsi, l'aide à la décision apparaît ainsi dans ce domaine comme la proposition de méthodes et d'outils permettant aux décideurs de naviguer dans les données consolidées dédiées à l'analyse.

Après avoir présenté les concepts fondateurs de ce domaine, nous nous sommes penchés sur quatre aspects pouvant être considérés comme structurants par rapport à la recherche dans ce domaine, à savoir : les données, les environnements de stockage de ces données, les utilisateurs et la sécurité. Par ailleurs, ce travail a permis de synthétiser les problèmes ouverts de ce domaine, qui se posent dans un nouveau contexte économique et technologique. Ce contexte est fortement corrélé avec l'émergence du *Cloud*, des outils *Open Source* qui modifient en profondeur le rapport des utilisateurs aux données et à leur analyse, posant de réels problèmes de sécurité. L'aide à la décision du point de vue des entrepôts de données et de l'analyse est amenée à évoluer en fonction de ce nouveau contexte, assurant aux professionnels du domaine un développement d'activité croissant et, aussi, un avenir scientifique prometteur avec des verrous identifiables nombreux, comme nous avons pu le constater.

Références

Abello, A., J. Darmont, L. Etcheverry, M. Golfarelli, J.-N. Mazon, F. Naumann, T.-B. Pedersen, S. Rizzi, J. Trujillo, P. Vassiliadis, et G. Vossen (2013). Fusion cubes : Towards self-service business intelligence. *International Journal of Data Warehousing and Mining 9*(2).

Abouzeid, A., K. Bajda-Pawlikowski, D. Abadi, A. Silberschatz, et A. Rasin (2009). Hadoopdb : an architectural hybrid of mapreduce and dbms technologies for analytical workloads. *Proc. VLDB Endow. 2*(1), 922–933.

Agrawal, D., S. Das, et A. El Abbadi (2011). Big data and cloud computing : current state and future opportunities. In *Proceedings of the 14th International Conference on Extending Database Technology*, EDBT/ICDT '11, New York, NY, USA, pp. 530–533. ACM.

Akkaoui, Z. E., E. Zimányi, J.-N. Mazón, et J. Trujillo (2011). A model-driven framework for etl process development. In *14th International Workshop on Data Warehousing and OLAP, Glasgow, United Kingdom (DOLAP 2011)*, pp. 45–52. ACM.

Aligon, J., M. Golfarelli, P. Marcel, S. Rizzi, et E. Turricchia (2011). Mining preferences from olap query logs for proactive personalization. In *15th International Conference on Advances in Databases and Information Systems, Vienna, Austria (ADBIS 2011)*, Volume 6909 of *Lecture Notes in Computer Science*, pp. 84–97. Springer.

Aouiche, K. et J. Darmont (2009). Data mining-based materialized view and index selection in data warehouses. *J. Intell. Inf. Syst. 33*(1), 65–93.

Bédard, Y. et J. Han (2009). *Geographic Data Mining and Knowledge Discovery*, Chapter Fundamentals of Spatial Data Warehousing for Geographic Knowledge Discovery. Taylor & Francis.

Benkrid, S. et L. Bellatreche (2011). Une démarche conjointe de fragmentation et de placement dans le cadre des entrepôts de données parallèles. *Technique et Science Informatiques 30*(8), 953–973.

Bentayeb, F., O. Boussaid, C. Favre, F. Ravat, et O. Teste (2009). Personnalisation dans les entrepôts de données : bilan et perspectives. In *5èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 2009), Montpellier*, Volume B-5 of *RNTI*, Toulouse, pp. 7–22. Cépaduès.

Berthold, H., P. Rösch, S. Zöller, F. Wortmann, A. Carenini, S. Campbell, P. Bisson, et F. Strohmaier (2010). An architecture for ad-hoc and collaborative business intelligence. In *Proceedings of the EDBT/ICDT Workshops*.

Cabanac, G., M. Chevalier, F. Ravat, et O. Teste (2007). An annotation management system for multidimensional databases. In *9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2007), Regensburg, Germany*, Volume 4654 of *LNCS*, pp. 89–98. Springer.

Cao, L., J. Ni, et D. Luo (2006). Ontological engineering in data warehousing. In *8th Asia-Pacific Web Conference (APWeb 2006), Harbin, China*, Volume 3841 of *Lecture Notes in Computer Science*, pp. 923–929. Springer.

Cattell, R. (2011). Scalable sql and nosql data stores. *SIGMOD Rec. 39*(4), 12–27.

Chaudhuri, S. et U. Dayal (1997). An Overview of Data Warehousing and OLAP Technology. *SIGMOD Rec. 26*(1), 65–74.

Chaudhuri, S., U. Dayal, et V. Narasayya (2011). An overview of business intelligence technology. *Commun. ACM 54*(8), 88–98.

Chow, R., P. Golle, M. Jakobsson, E. Shi, J. Staddon, R. Masuoka, et J. Molina (2009). Controlling data in the cloud : Outsourcing computation without outsourcing control. In *First ACM Cloud Computing Security Workshop (CCSW 2009), Chicago, IL, USA*, pp. 85–90.

Codd, E. (1993). Providing olap (on-line analytical processing) to user-analysts : an it mandate. Technical report, E.F. Codd and Associates.

Darmont, J., T.-B. Pedersen, et M. Middelfart (2012). Cloud intelligence : What is really new ? Panel.

Essaidi, M. (2010). ODBIS : towards a platform for on-demand business intelligence services. In *Proceedings of the EDBT/ICDT Workshops*, Lausanne, Switzerland.

Franco, J. M. (1997). *Le Data Warehouse, le Data Mining*. Eyrolles.

Golfarelli, M., F. Mandreoli, W. Penzo, S. Rizzi, et E. Turricchia (2012). OLAP query reformulation in peer-to-peer data warehousing. *Information Systems 5*(32), 393–411.

Gonzalez, H., A. Y. Halevy, C. S. Jensen, A. Langen, J. Madhavan, R. Shapley, W. Shen, et J. Goldberg-Kidon (2010). Google fusion tables : web-centered data management and collaboration. In *2010 ACM International Conference on Management of Data (SIGMOD 2010), Indianapolis, USA*, pp. 1061–1066.

Grabova, O., J. Darmont, J.-H. Chauchat, et I. Zolotaryova (2010). Business intelligence for small and middle-sized enterprises. *SIGMOD Record 39*(2), 39–50.

Han, J., Y. Chen, G. Dong, J. Pei, B. Wah, J. Wang, et Y. Cai (2005). Stream Cube :An Architecture for Multidimensional analysis of Data Streams. *Distributed and Parallel Databases 18*, 173–187.

Inmon, W. H. (1996). *Building the Data Warehouse*. John Wiley & Sons.

Jin, X., J. Han, L. Cao, J. Luo, B. Ding, et C. X. Lin (2010). Visual cube and on-line analytical processing of images. In *19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada*, pp. 849–858.

Jörg, T. et S. Dessloch (2009). Near real-time data warehousing using state-of-the-art etl tools. In *Enabling Real-Time Business Intelligence – Third International Workshop (BIRTE 2009), Lyon, France*, Volume 41 of *LNBIP*, pp. 100–117. Springer.

Jovanovic, P., O. Romero, A. Simitsis, et A. Abelló (2012). Integrating etl processes from information requirements. In *14th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2012), Vienna, Austria*, Volume 7448 of *Lecture Notes in Computer Science*, pp. 65–80. Springer.

Kandel, S., A. Paepcke, J. M. Hellerstein, et J. Heer (2012). Enterprise data analysis and visualization : An interview study. *IEEE Trans. Vis. Comput. Graph. 18*(12), 2917–2926.

Kimball, R., L. Reeves, M. Ross, et W. Thornthwaite (2000). *Concevoir et déployer un data warehouse*. Eyrolles.

Leavitt, N. (2010). Will nosql databases live up to their promise ? *Computer 43*(2), 12–14.

Malinowski, E. et E. Zimányi (2004). Olap hierarchies : A conceptual perspective. In *16th International Conference on Advanced Information Systems Engineering (CAiSE 2004), Riga, Latvia*, Volume 3084 of *Lecture Notes in Computer Science*, pp. 477–491. Springer.

Mazón, J.-N., J. Lechtenbörger, et J. Trujillo (2009). A survey on summarizability issues in multidimensional modeling. *Data Knowl. Eng. 68*(12), 1452–1469.

Messaoud, R. B., O. Boussaid, et S. L. Rabaséda (2006). A multiple correspondence analysis to organize data cubes. In *Databases and Information Systems IV - Selected Papers from the Seventh International Baltic Conference, DB&IS 2006, July 3-6, 2006, Vilnius, Lithuania*, Volume 155 of *Frontiers in Artificial Intelligence and Applications*, pp. 133–146. IOS Press.

Middelfart, M. (2012). Analytic lessons : in the cloud, about the cloud. Indistrial keynote.

Muñoz, L., J.-N. Mazón, et J. Trujillo (2009). Automatic generation of etl processes from conceptual models. In *12th International Workshop on Data Warehousing and OLAP (DOLAP 2009), Hong Kong, China*, pp. 33–40. ACM.

Nguyen, T.-V.-A., L. d'Orazio, S. Bimonte, et J. Darmont (2012). Cost models for view materialization in the cloud. In *Workshop on Data Analytics in the Cloud (EDBT-ICDT/DanaC 12), Berlin, Germany*.

Papastefanatos, G., P. Vassiliadis, A. Simitsis, et Y. Vassiliou (2012). Metrics for the prediction of evolution impact in etl ecosystems : A case study. *J. Data Semantics 1*(2), 75–97.

Pedersen, T. B. (2010). Research challenges for cloud intelligence : invited talk. In *2010 EDBT/ICDT Workshops, Lausanne, Switzerland*.

Ravat, F., O. Teste, R. Tournier, et G. Zurfluh (2008). A top keyword extraction method for olap document. In *International Conference on Data Warehousing and Knowledge Discovery (DAWAK 2008)*, Volume 5182, pp. 55–64. Springer Verlag, LNCS.

Selma, K., B. IlyèS, B. Ladjel, S. Eric, J. StéPhane, et B. Michael (2012). Ontology-based structured web data warehouses for sustainable interoperability : requirement modeling, design methodology and tool. *Comput. Ind. 63*(8), 799–812.

Simitsis, A., D. Skoutas, et M. Castellanos (2010). Representation of conceptual etl designs in natural language using semantic web technology. *Data Knowl. Eng. 69*(1), 96–115.

Stonebraker, M., D. Abadi, D. J. DeWitt, S. Madden, E. Paulson, A. Pavlo, et A. Rasin (2010). Mapreduce and parallel dbmss : friends or foes ? *Commun. ACM 53*(1), 64–71.

Sureau, F., F. Bouali, et G. Venturini (2009). Optimisation heuristique et génétique de visualisations 2d et 3d dans olap : premiers résultats. In *5èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 2009), Montpellier*, Volume B-5 of *RNTI*, Toulouse, pp. 65–78. Cépaduès.

Thiele, M. et W. Lehner (2011). Real-time BI and situational analysis. In *Business Intelligence Applications and the Web : Models, Systems and Technologies*, pp. 285–309. Hershey, PA : IGI Global.

Wang, H., J. Li, Z. He, et H. Gao (2005). OLAP for XML data. In *Proceedings of the* 1^{st} *International Conference on Computer and Information Technology (CIT2005)*, Shangai, China, pp. 233–237. IEEE Computer Society.

Summary

In this paper, we present the background regarding decisional processes in terms of data warehousing and OLAP. We present the main related concepts and the research challenges according to four points of view: data, storage, users and security.

Une aide à la décision pour l'apprenant basée sur le QCM

Igor Crévits, Saïd Hanafi, Najah Kushlaf

LAMIH, Université de Valenciennes de du Hainaut-Cambrésis
Le Mont Houy - 59313 Valenciennes Cedex 9
Igor.Crevits@univ-valenciennes.fr

Résumé. Nous nous intéressons au déploiement du cadre méthodologique proposé par le Processus d'Aide à la décision dans le domaine de l'apprentissage, en particulier dans le cas des langues. Le modèle sur lequel s'appuie l'aide à la décision est représenté par un QCM. Sa fonction concerne essentiellement l'identification de la structure des connaissances de l'apprenant. Cette structure correspond aux processus décisionnels qui se sont construit durant son apprentissage et son vécu. La nature hautement structurée du QCM et sa capacité à réunir un grand nombre de données permettent de la voir selon une logique d'entrepôts et de fouille de données. Sur la base des réflexions méthodologiques, des liens peuvent être établis entre aide à la décision et entrepôts et fouille de données.

1 Introduction

Les travaux du domaine constitué de l'aide à la décision se divisent en deux branches : théorie de la décision (d'orientation mathématique) et système d'information (d'orientation mathématique). L'essentiel du sujet porte sur la représentation extérieure, le modèle, qui permet d'apporter une solution au problème de décision posé. La décision est donc vue comme un résultat. Dans la réalité de problèmes concrets résolus par des décideurs dans le cadre de processus décisionnels, la construction du modèle mérite alors réflexion. En conséquence, il se dégage la nécessité de se pencher sur une orientation méthodologique de l'aide à la décision. Dans cette orientation, le facteur important n'est pas tant la décision comme résultat, mais le processus décisionnel avec lequel cette solution doit être cohérente. Pour cela, une analyse du problème concret, dont le processus décisionnel est une composante, est à mettre en œuvre. Les travaux présentés s'inscrivent dans les réflexions méthodologiques de l'aide à la décision. La réalité abordée concerne l'apprentissage au sens éducatif du terme.

La cadre méthodologique utilisé comme référence est le Processus d'Aide à la Décision (PAD) qui structure l'activité de construction de l'aide en quatre étapes : (1) la situation problématique identifie les facteurs décisionnels clés du problème, (2) la formulation de problème rassemble les solutions envisageables de façon cohérente avec la situation problématique, ainsi que le mode de recherche des solutions les plus appropriées, (3) le modèle d'évaluation permet une représentation numérique fine, basée sur les cadres offerts par la théorie de la décision, permettant de produire les solutions recherchées, (4) la recommandation finale se focalise sur l'intégration des solutions dans la réalité.

Une combinaison de deux PAD est déployée dans le contexte de l'apprentissage afin de structurer la place que peut prendre un QCM dans les décisions d'apprentissage et

ainsi constituer le socle d'une aide à la décision. Une situation d'apprentissage met en jeu deux notions fondamentales de savoir commun à acquérir, capitalisé sous forme écrite, et de connaissance qui résulte de l'intériorisation personnelle du savoir. Le QCM est traditionnellement un outil d'évaluation orientée vers le résultat. Par un principe sommatif, il permet d'identifier le degré de conformité des connaissances personnelles au savoir commun. Or, ce principe sommatif favorable à l'évaluation du savoir peut être complété par un principe formatif. L'évaluation concerne alors les facteurs et priorités qui interviennent dans les décisions d'apprentissage prises par l'apprenant. Le QCM permet également une évaluation de la construction des connaissances de l'apprenant. Le résultat du QCM fournit donc un appui à l'enseignant qui lui permet de mieux argumenter ses recommandations. PAD permet donc de mieux établir les liens entre situation d'apprentissage, vue comme une situation problématique, et QCM, vu comme modèle d'évaluation du savoir comme des connaissances. Ces liens sont établis, à la fois en terme de conception du QCM comme en terme d'exploitation du résultat par l'enseignant et l'apprenant.

Trois particularités du QCM sont utilisées : la collecte structurée possible d'un volume important de données, la construction envisageable de nombreuses combinaisons de ces données et la possibilité de mettre en évidence, par les effets (les résultats), les facteurs clés du processus décisionnel d'apprentissage de l'apprenant. De ce fait, le QCM s'apparente aux concepts d'entrepôts de données en ce qui concerne sa conception, ainsi que de fouille de données en ce qui concerne l'exploitation des données produites par l'apprenant. La situation d'apprentissage constitue le contexte décisionnel que PAD permet de relier de façon cohérente aux moyens offerts par le QCM. Une combinaison est alors envisageable entre les concepts de l'aide à la décision d'une part, et d'entrepôts et de fouille de données d'autre part.

Dans la première partie, nous donnons les éléments constituant une situation d'apprentissage. Puis nous établissons un lien avec la définition de l'aide à la décision pour montrer qu'une situation d'apprentissage s'apparente à une situation d'aide à la décision. L'évaluation constitue une forme de situation d'apprentissage utile aux décisions d'apprentissage. En prenant la forme d'un QCM, l'évaluation constitue le modèle sur lequel l'aide à la décision peut s'appuyer. Dans la deuxième partie, nous présentons PAD qui permet de structurer situation d'apprentissage et modèle d'évaluation des connaissances par le QCM, ainsi que les liens qui les associent. Dans la troisième partie, nous proposons une combinaison de deux PAD pour montrer que le QCM peut fournir une évaluation fine des connaissances de l'apprenant que l'enseignant peut exploiter pour faire évoluer les processus d'apprentissage. Dans la quatrième partie, nous tirons quelques enseignements des particularités d'une situation d'apprentissage. Enfin, dans la cinquième partie, nous dressons un parallèle entre la démarche méthodologique de l'aide à la décision et les principes d'entrepôts et de fouille de données.

2 Situation d'apprentissage et d'aide à la décision

Une situation d'apprentissage met en jeu deux concepts fondamentaux : connaissance et savoir. La connaissance résulte d'une intériorisation du savoir (Labour, 2010) lequel est capitalisé et normé. La connaissance se construit et se transforme au cours du temps (Astolfi, 1993). Une ambigüité peut donc apparaître d'un manque de différenciation entre ces

deux notions. En effet, l'apprenant peut confondre intériorisation et aptitude à retenir le savoir à l'identique de sa capitalisation.

Une situation d'apprentissage représente également un ensemble de conditions et de circonstances susceptibles d'amener une personne à construire ses connaissances (Faerber, 2004). Une situation d'apprentissage s'organise autour de trois composantes : (1) une problématique énoncée par l'enseignant qui nécessite la mobilisation d'une parcelle de savoir, (2) un traitement, encadré par l'enseignant, de cette problématique par l'apprenant qui sollicite ses connaissances, (3) un environnement technologique et social.

Une évaluation rassemble en un processus ces trois composantes. L'évaluation permet une explicitation des connaissances que l'environnement technologique et social permet de capitaliser. La trace ainsi laissée peut être exploitée pour prendre des décisions. Ces décisions peuvent avoir une visée institutionnelle et pédagogique (Hadji, 1993). (Bouyssou, 2000) se focalise sur le cadre institutionnel. Nous nous intéressons au cadre pédagogique, dans le contexte de l'apprentissage des langues, où une dualité est reconnue entre processus d'apprentissage et évaluation (Sommer, 2001).

Une situation d'apprentissage est propice à l'aide à la décision dont (Roy, 1985) donne la définition suivante : *« L'aide à la décision est l'activité de celui qui, prenant appui sur des modèles clairement explicités, mais non nécessairement complètement formalisés, aide à obtenir des éléments de réponses aux questions que se pose un intervenant dans un processus de décision, éléments concourant à éclairer la décision et normalement à prescrire, ou simplement à favoriser, un comportement de nature à accroître la cohérence entre l'évolution du processus d'une part, les objectifs et le système de valeurs au service desquels cet intervenant se trouve placé d'autre part.* ».

Cette définition apparait correspondre à une situation d'apprentissage :

- l'apprenant et l'enseignant constituent deux intervenants recherchant un accroissement de cohérence entre le processus de décision de construction et de transformation des connaissances de l'apprenant et le savoir constitué en un système de valeurs dont le garant est l'enseignant,
- le modèle est représenté par le processus d'évaluation dont la trace fournit un appui pour les deux intervenants.

L'appui se concentre sur la confrontation de la connaissance et du savoir par un dialogue amélioré entre l'apprenant et l'enseignant. Le modèle porte sur la standardisation du problème et de son traitement comme une mise en situation d'une parcelle de savoir et des connaissances correspondantes. L'amélioration est permise de deux façons par :

- un report de l'activité de l'enseignant de l'évaluation vers l'investissement du résultat, ce qui constitue les composantes (1) et (2) d'une situation d'apprentissage,
- la disponibilité de données fines qui peuvent être combinées de plusieurs façons.

Le modèle prend donc la forme d'un QCM, un outil répandu en apprentissage (Brady, 2005), lequel est informatisé au moyen d'une base de données. Deux points essentiels sont abordés durant le dialogue entre l'apprenant et l'enseignant :

- l'existence d'une confusion entre connaissance et savoir chez l'apprenant qui le conduit à valoriser la restitution à l'identique du savoir plutôt que de mobiliser ses connaissances,
- l'attention trop importante accordée au savoir qui conduit l'apprenant à ignorer les processus de construction et de transformation des connaissances et à négliger l'investissement de ses erreurs comme traces de non-conformité au savoir de ses connaissances.

Le principe est mis en œuvre dans le contexte de l'apprentissage de la langue française. Le QCM est composé de 30 questions répartis en deux groupes de 15 questions, l'un consacré au passé composé et le second à l'imparfait. Une question est associée à 5 choix linguistiques et trois choix pédagogiques. Chacun des choix linguistiques correspond à une notion parmi cinq : présent, passé composé, imparfait, plus que parfait et passé simple. Pour chaque groupe, les distracteurs (choix différents de la bonne réponse) sont donc harmonisés. Les choix pédagogiques consistent en une déclaration de certitude (choix 'Je suis sûr') ou de non certitude (choix 'Je ne suis pas sûr') dans le choix linguistique sélectionné ou l'indication de l'impossibilité de répondre (choix 'Je ne parviens pas à répondre').

L'évaluation s'opère en deux étapes. Après que l'apprenant ait rempli le QCM, les résultats sont examinés par l'enseignant pour préparer un entretien en cote-à-cote avec l'apprenant. Au cours de cet entretien, les résultats du QCM sont présentés à l'apprenant de plus en plus précisément en sept étapes : (1) score linguistique global, (2) les notions abordées par le QCM (passé composé et imparfait), (3) ventilation du score vers les deux notions du QCM, (4) nombre de choix 'Je ne parviens pas à répondre'), (5) décomposition du score linguistique avec le score pédagogique, (6) décomposition du score linguistique avec le score pédagogique pour les deux notions du QCM, (7) ventilation vers les distracteurs.

L'apprenant est invité à commenter ses résultats et la façon dont il les perçoit. Plus largement, cet entretien est l'occasion de rechercher les éléments déterminants dans la construction de ses connaissances, en particulier son vécu et sa pratique.

Une situation d'apprentissage s'apparente à une situation d'aide à la décision. Le QCM, en fournissant une évaluation fine des connaissances de l'apprenant, constitue le modèle sur lequel l'enseignant peut s'appuyer pour émettre des recommandations fines adaptées à l'apprenant. Pour établir précisément le lien entre une situation d'apprentissage et le QCM, nous nous appuyons sur le cadre méthodologique de PAD.

3 Processus d'Aide à la Décision

La définition d'un cadre méthodologique de l'aide à la décision nécessite de se pencher sur l'interaction entre le client et l'analyste vue comme la mise en œuvre d'un processus décisionnel conduit par l'analyste, au premier plan, et le client, dont le sujet est la construction d'une solution solidement raisonnée au problème du client. Il existe une littérature abondante sur les méthodes de structuration de problème qui mettent en évidence que l'aide à la décision n'est pas limitée à la résolution d'un modèle de décision, mais nécessite une réflexion sur le problème. Nous nous focalisons sur les plus connues : Cognitive Mapping (Eden, 1988), Strategic Choice (Friend, 1987), Soft Systems Methodology (Checkland, 1990), Value Focused Thinking (Keeny, 1992) and Integrating Approach (Belton, 2002). Les limites de ces méthodes, relevées dans (Bouyssou, 2006), portent sur un manque de cohérence entre les points de vue du client et l'analyste qui conduit à une représentation trop proche de l'un et donc défavorable à l'autre, en particulier en terme de compréhension.

On peut relever d'autres limitations. On ne note aucune différenciation forte du problème et de la solution ce qui conduit à un glissement vers la solution qui détermine plus ou moins le problème exprimé par le client, selon la méthode utilisée. L'entrée en aide à la décision est donc plus déterminée par la solution que par le problème. L'exemple le plus marquant est Cognitive Mapping qui favorise l'émergence de consensus orientés vers l'action, c'est-à-dire

une solution au problème. L'orientation vers l'action est également la base de Strategic Choice et Soft Systems Methodology. Integrated Approach tend à améliorer la réponse offerte par le modèle par une utilisation de plusieurs méthodes multicritères. Le raisonnement est donc fortement orienté vers les alternatives ou actions potentielles. Value Focused Thinking volorise la hiérarchisation des objectifs comme point de départ de la recherche de solutions vers les objectifs à haute valeur. Bien que le point de départ soit ici le problème, la différenciation de la solution est faible. L'ensemble de ces remarques attire l'attention sur la nécessité d'un arbitrage entre problème et solution. Le rôle de l'analyste réside d'abord dans cet arbitrage avant la construction du modèle. Le manque de différenciation entre problème et solution vient de ce que l'interlocuteur de l'analyste est unique. Or, deux rôles sont dévolus au client selon qu'il exprime un problème ou qu'il envisage (ou qu'il estime) une solution. Le client est également décideur. La séparation et l'arbitrage des deux rôles, si elle ne correspond pas à la réalité de deux individus différents, relèvent des prérogatives de l'analyste. Deux représentations différenciées et intelligibles peuvent donc améliorer la construction de l'aide.

PAD s'inscrit dans une démarché constructive en structurant les décisions relatives à la conception d'une aide à la décision selon les étapes de rationalité procédurale (intelligence, design, choice, review) de (Simon, 1977). PAD considère deux intervenants : le client (lequel peut être décideur), demandeur de l'aide et l'analyste, qui disposent de connaissances méthodologiques. Tous deux cherchent à construire, sous une contrainte de temps, une représentation partagée structurée en quatre artefacts : (1) la situation problématique, (2) la formulation du problème, (3) le modèle d'évaluation et (4) la recommandation finale. Chacun amène sa connaissance afin de produire une recommandation utile à apporter une réponse au problème du client.

(1) La *situation problématique* a pour objet de représenter l'origine du problème et l'implication du client, d'identifier les conséquences d'une décision et la façon la plus judicieuse d'apporter une solution. Ces précisions sont également utiles à l'analyste pour qu'il clarifie lui-même sur quel point l'aide peut porter.

Formellement, la situation problématique $\mathcal{P}$ est un triplet ($\mathcal{A}$, $\mathcal{O}$, $\mathcal{S}$) où :

- $\mathcal{A}$ est l'ensemble des participants au processus de décision,
- $\mathcal{O}$ est l'ensemble des enjeux que les participants amènent dans le processus de décision,
- $\mathcal{S}$ est l'ensemble des engagements pris par chaque participant sur ses enjeux et ceux des autres.

L'objectif est bien de clarifier, pas de figer. Les précisions apportées permettent de mieux faire évoluer la demande d'aide et la réponse associée.

(2) La *formulation du problème* a pour objet la réponse à apporter au problème clarifié dans le cadre de la situation problématique. La construction de la formulation du problème s'appuie sur des choix autour d'une idée que se fait le client d'une certaine rationalité dans la réponse à apporter au problème. Ces choix font l'objet d'une explicitation par l'analyste en vue de leur capitalisation en une représentation formelle, étape préalable essentielle à l'application d'une méthode d'aide à la décision.

Du point de vue formel, la formulation du problème Γ est un triplet (A, V, Π) où :

- A est l'ensemble des actions potentielles dans le cadre de la situation problématique $\mathcal{P}$,
- V est l'ensemble des points de vue sous lesquels il est envisagé d'observer, d'analyser, d'évaluer et de comparer les actions potentielles,

- Π est la problématique décisionnelle, la typologie d'application envisageable sur A, une anticipation de ce que le client attend.

La formulation du problème est une étape importante dans la construction de la représentation partagée entre le client et l'analyste. Elle constitue la charnière entre le problème de décision vu d'une façon implicite par le client, et une représentation manipulée par l'aide destinée à produire un résultat. La formulation du problème en explicitant le problème de décision permet à l'analyste de préparer son travail de modélisation d'une façon intelligible pour le client, favorisant ainsi l'expression de son point de vue et son intervention.

(3) Le *modèle d'évaluation* a pour objet d'estimer finement, sur la base d'une représentation numérique dotée de propriétés formelles, l'impact de la solution envisagée dans la formulation du problème.

Le modèle d'évaluation $\mathcal{M}$ est un n-uple (A, D, E, H, $\mathcal{U}$, $\mathcal{R}$) où :

- A est l'ensemble des alternatives sur lesquelles le modèle d'évaluation s'applique,
- D est l'ensemble des dimensions, éventuellement muni de propriétés structurelles, par lesquelles les actions potentielles de A sont manipulées par le modèle,
- E est l'ensemble des échelles associées à chaque élément de D,
- H est l'ensemble des critères sous lesquels les éléments de A sont évalués afin de prendre en compte les préférences du client, restreintes à chaque critère,
- $\mathcal{U}$ est l'ensemble des distributions d'incertitudes associées à D et/ou H,
- $\mathcal{R}$ est l'ensemble des opérateurs de synthèse d'information des éléments de A ou de $A \times A$, notamment les opérateurs d'agrégation.

Il s'agit d'une représentation conforme en grande partie aux modèles d'aide à la décision classiquement utilisés. Le modèle d'évaluation peut être soumis à plusieurs validations internes (consistance logique des éléments du modèle, validation expérimentale) et externes vis-à-vis de la réalité du problème et de la décision (validation conceptuelle de la cohérence en amont des éléments du modèle avec le problème et la décision et de la signification de ces éléments pour le client, validation opérationnelle en aval de la cohérence avec le processus décisionnel et les effets imprévus du modèle).

(4) La *recommandation finale* Φ s'attache au retour à la réalité par la mise en cohérence du résultat du modèle d'évaluation avec le langage du client autour de trois questions :

- signification : aptitude de la recommandation à apporter une réponse à l'ensemble des préoccupations du client,
- complétude opérationnelle : aptitude de la recommandation à être mise en oeuvre,
- légitimité : cohérence de la recommandation avec le contexte de la décision qui n'est pas pris en compte dans le modèle d'évaluation.

Une présentation détaillée de PAD dans (Tsoukiàs, 2007) permet de représenter la hiérarchie de construction des artefacts et sous-artefacts de la façon décrite à la figure 1. Les liens de précédence ne doivent pas être perçus comme des étapes qui se succéderaient de façon stricte depuis $\mathcal{P}$ jusque Φ. Ils sont plutôt à envisager comme des guides qui orientent le client et l'analyste dans les questions qui se posent lors la conception de l'aide à la décision. Il en découle certes une temporalité. Mais des allers-retours sont tout à fait envisageables si l'avancée dans PAD montre que la réflexion sur un artéfact nécessite de revenir sur des options précédemment définies ou de les compléter. Le déploiement dans la gestion du trafic aérien en section suivante illustre cette construction. Les participants, les actions potentielles et les alternatives constituent les points de départ respectifs de construction de la situation

problématique, de la formulation du problème et du modèle d'évaluation. On notera que Π et $\mathcal{U}$ sont exclusivement amenés par l'analyste.

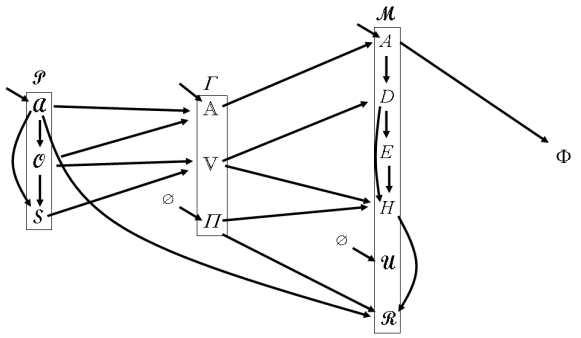


Figure 1 – Construction des artéfacts du Processus d'aide à la décision

La très grande force de PAD réside dans la séparation et l'indépendance du problème et de la solution représentés chacun par deux artefacts distincts, respectivement la situation problématique et la formulation du problème. Ces deux représentations peuvent alors être clairement reliées et leur cohérence assurée par un partage qui ne soit pas plus favorable au client qu'à l'analyste. Ces deux représentations permettent également de séparer clairement l'intervention du client et du décideur. Les artefacts manipulés dans ces deux représentations sont également intelligibles pour le client et l'analyste car ils ne recourent à aucun cadre théorique. Nous illustrerons cette cohérence par la suite dans le contexte de la navigation aérienne. Plus généralement, en étant adossé au principe de rationalité procédurale de Simon, PAD dispose d'un potentiel de déploiement très important.

PAD a été déployé dans différentes situations comme l'organisation administrative (Ostanello, 1993), la conception de logiciels (Stamelos, 2003) ou la maintenance d'infrastructures routières (Tsoukiàs, 2012). L'aide à la décision y est vue comme une résolution de problèmes. L'aide aux décisions dans le domaine de l'apprentissage relève d'une autre nature. Les particularités seront abordées en discussion. Ce domaine constitue un réel challenge pour l'aide à la décision et une épreuve déterminante pour un cadre méthodologique.

PAD fournit un cadre méthodologique synthétique qui permet de construire une aide à la décision. Par les relations entre ses artefacts, il permet également de donner un cadre de référence à des situations qui s'apparentent informellement à de l'aide à la décision, comme c'est le cas d'une situation d'apprentissage. Nous abordons dans la suite le déploiement de PAD dans ce type de situation ce qui permet d'améliorer l'efficacité du QCM.

4 Déploiements du Processus d'Aide à la Décision

Nous proposons une analyse fine du processus d'évaluation en nous appuyant sur PAD. Pour relier de façon appropriée les deux étapes, les relations entre l'apprenant et l'enseignant sont à affiner, la difficulté résidant dans la séparation de lui-même que doit opérer

l'apprenant dans l'examen de ses connaissances. Le concept de compétence clinique sur lequel s'appuie la médecine apparait pertinent. (Wimmers, 2006) définit la compétence clinque par : (1) une dimension cognitive rassemblant connaissances et aptitudes à les appliquer dans une situation particulière, (2) la capacité de diagnostic et résolution de problème adaptée à l'histoire du patient, (3) l'aptitude à communiquer avec les patients et les confrères, (4) les qualités à établir des relations professionnelles et respectueuses avec le patient pour améliorer son bien-être.

La compétence clinique peut donc se résumer à associer de façon approprié un cadre général de la médecine à la particularité de chaque patient. Dans la différenciation entre connaissance et savoir, on retrouve les traits de la compétence clinique. Une différence importante résidant dans le fait que la compétence de l'apprenant est à appliquer à lui-même, l'enseignant est déterminant dans l'acquisition de cette compétence. Pour cela, l'examen de compétence clinique, utilisé dans l'enseignement de la médecine (Wass, 2001), sert de cadre de référence (figure 2). L'examen de compétence clinique est chargé d'évaluer l'aptitude de l'élève médecin à interagir avec le patient afin de déterminer précisément la pathologie dont il souffre. Durant cet examen, un médecin confirmé encadrant est présent afin d'observer et de commenter (durant l'examen ou a posteriori selon l'organisation) la démarche de l'élève.

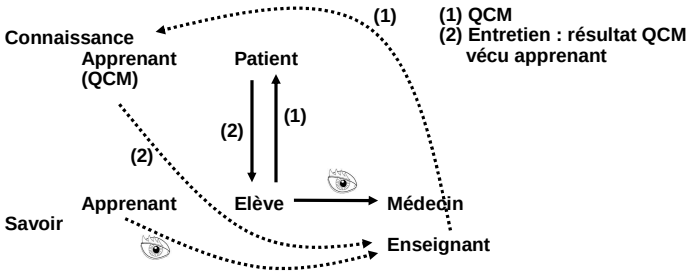


Figure 2 - Représentation du QCM et de l'entretien par un examen de compétence clinique

Il s'agit bien ici de s'appuyer sur un cadre structuré dans un domaine où la structuration de l'individu est examinée. En aucun cas, il ne s'agit de considérer l'apprenant de comme un patient, souffrant d'une pathologie, qui nécessite des soins. Les cadres sont différents mais la démarche est la même.

Une particularité apparaît (figure 2). L'examen de compétence clinique met en jeu trois individus alors que l'évaluation en apprentissage se limite à l'apprenant et l'enseignant. Or, compte tenu d'un état des connaissances qui peuvent être mal structurées, une séparation peut être considérée entre l'apprenant en tant que personne physique et un état de ses connaissances révélé par le QCM. On notera qu'une seule évaluation peut suffire à rendre compte de défauts de structure. L'enseignant joue un rôle de médiateur qui permet à l'apprenant d'exprimer la partie de lui-même que représentent ses connaissances (figure 2, liens (2)). La présentation des résultats lors de l'entretien et les questions posées par l'enseignant à l'apprenant correspondent donc aux sollicitations du patient auxquelles procède l'élève médecin (figure 2, liens (1)). En révélant des facteurs de structure des connaissances, les résultats du QCM correspondent aux informations qu'apporte le patient qui permettent à l'élève d'établir un diagnostic. Les réponses que donne l'apprenant à l'enseignant correspondent à une observation directe de la part du médecin chargé d'estimer

la qualité des compétences cliniques de l'élève (figure 2, liens d'observation exprimés par l'œil). L'observation n'est donc pas directe, elle est provoquée par l'enseignant.

En l'absence de demande d'aide, la première question abordée concerne la définition du client et du décideur. Par la demande d'aide qu'il formule, le client est l'entité a la vision la plus complète de la structure du problème de décision posé. Lorsqu'il est différent du client, le décideur dispose d'éléments permettant de construire des solutions. En faisant la différence entre savoir et connaissance et en accordant une importance plus grande aux connaissances, l'enseignant constitue le client de l'aide. A priori, il n'est pas établi que l'apprenant fasse une différence entre connaissance et savoir. L'aide à donc un double objectif de vérifier que cette différence est faite et de faire en sorte que l'apprenant accorde une importance plus grande aux connaissances par un ajustement de ses processus d'acquisition et de transformation. De ce fait, l'apprenant occupe la position de décideur. On notera que d'autres situations d'apprentissage admettent la configuration inverse. Dans le cas d'examens comme le TOEIC, par la valorisation du score, il est recherché un degré de conformité dans la restitution du savoir. La construction et la transformation des connaissances est donc secondaire. Le client est donc l'apprenant. L'enseignant se trouve en position de décideur de la façon la plus pertinente pour l'apprenant de retenir le savoir adapté.

		QCM	**Entretien**		
$\mathcal{P}_q$	$\mathcal{A}_q$	Apprenant	Apprenant, QCM rempli, vécu de l'apprenant	$\mathcal{A}_e$	$\mathcal{P}_e$
	$\mathcal{O}_q$	Savoir, connaissance	Connaissance, savoir	$\mathcal{O}_e$	
	$\mathcal{S}_q$	Note	Regard porté par l'apprenant sur ses connaissances et ses erreurs	$\mathcal{S}_e$	
Γ_q	$\mathbb{A}_q$	QCM	Entretien	$\mathbb{A}_e$	Γ_e
	$\mathbb{V}_q$	Score, certitude	Score, certitude	$\mathbb{V}_e$	
	Π_q	Description (de l'ordre de $\mathbb{V}_q$)	Rangement (sur $\mathbb{V}_e$)	Π_e	
$\mathcal{M}_q$	A_q	30 questions (15 sur le passé composé, 15 sur l'imparfait)	∅	A_e	$\mathcal{M}_e$
	D_q	Linguistique, pédagogique	∅	D_e	
	E_q	{présent, passé composé, imparfait, plus que parfait, passé simple} {sûr, pas sûr, ne parviens pas à répondre}	∅	E_e	
	H_q	D_q	∅	H_e	
	$\mathcal{U}_q$	30 questions	∅	$\mathcal{U}_e$	
	$\mathcal{R}_q$	∅	Etapes de présentation graduelle des résultats du QCM	$\mathcal{R}_e$	
Φ_q		∅	Engagements de l'apprenant pris sur lui-même et des intervenants ponctuels		Φ_e

Tableau 1 – Synthèse des déploiements de PAD dans l'apprentissage des langues

Compte tenu des éléments présentés, le processus d'évaluation peut se représenter par le déploiement d'une séquence de deux PAD dont une synthèse est représentée par le tableau 1. Sous cette représentation, plusieurs particularités apparaissent clairement. L'ordre des enjeux s'inverse du QCM à l'entretien. L'objet de l'aide à la décision concerne la structure des enjeux plutôt que la partition de l'ensemble des actions potentielles comme le montre les ensembles $\mathbb{A}_q$ et $\mathbb{A}_e$ qui ne sont composés que d'un seul élément. La prédominance accordée à l'enjeu du savoir dans $\mathcal{O}_q$ constitue une hypothèse qui sera vérifiée durant l'entretien, l'objectif du dispositif étant de replacer l'enjeu des connaissances au premier plan, si cela s'avère nécessaire. L'enjeu des connaissances est donc prioritaire dans $\mathcal{O}_e$. En complétant l'évaluation par l'interaction avec l'enseignant, il est possible de vérifier l'importance qu'accorde l'apprenant au résultat que représente la note. De ce fait, l'apprenant apparaît être le seul participant dans $\mathcal{A}_q$. La note représente donc l'engagement $\mathcal{S}_q$ qu'il prend vis-vis de lui-même au moment du QCM, sur l'enjeu du savoir qu'il peut vouloir tenter de restituer avec le plus fort degré de conformité. L'intérêt que porte l'apprenant à la constitution de ses connaissances peut être identifié par l'acceptation de ses erreurs comme base d'un travail d'amélioration. Dans ce cadre, le témoignage de son vécu de situations d'apprentissage antérieures est déterminant. Ce vécu, en intervenant dans la constitution des connaissances de l'apprenant, est considéré comme un participant à part entière dans $\mathcal{A}_e$.

La structuration du modèle d'évaluation de l'entretien se limite au protocole d'entretien qui aborde graduellement les résultats du QCM. De ce fait, le seul élément du modèle d'évaluation $\mathcal{M}_e$ est l'opérateur de synthèse d'information $\mathcal{R}_e$. Il complète le modèle d'évaluation $\mathcal{M}_q$ qui ne nécessite pas un tel opérateur. Cet opérateur $\mathcal{R}_e$ est hiérarchisé et procède de façon désagragative depuis le score global jusque la ventilation des réponses sur les distracteurs. Comme il s'agit d'identifier les facteurs individuels, tel le vécu, qui interviennent dans la constitution des connaissances une trop forte structuration de cette évaluation ne permettrait pas d'obtenir des résultats sur lesquels des recommandations pourraient prendre appui. La fonction d'évaluation est donc remplie par l'enseignant. Le protocole d'entretien permet de mettre en évidence des situations contradictoires qui permettent de mieux identifier la constitution des connaissances. La prise en compte de l'incertain s'opère uniquement durant le QCM par l'accumulation des 15 questions, représentée par $\mathcal{U}_q$, portant chacune sur les deux notions centrales du QCM.

Aucune recommandation Φ_e n'est émise à l'issue du QCM. Au delà de la collecte de données, le QCM est chargé de déclencher la recherche du score à tout prix par une survalorisation du savoir. Cette attitude porte le nom de washback effect qui est défini comme la résultante, positive ou négative, de la rencontre entre apprentissage et évaluation (Djurić, 2008). Les recommandations sont données durant la seconde phase dans Φ_e. Comme ce qui est déterminant dans le dispositif est l'hypothèse sur la priorité au savoir accordé par l'apprenant, elles concernent essentiellement les engagements que peut prendre l'apprenant vis-à-vis de son apprentissage. Selon le cas, il peut s'agir :

- d'un investissement plus grand des erreurs, ce qui est fréquemment le cas d'apprenants d'un bon niveau,
- la création d'engagements en se faisant corriger dans sa pratique quotidienne,
- la prise de recul critique quant à des habitudes acquises durant l'apprentissage,
-le relâchement d'engagements pris dans le cadre de l'apprentissage et qui ne sont plus adapté dans une situation de mise en œuvre des connaissances,
- …

Le dispositif a été déployé dans le cadre d'enseignements en Français Langue Etrangère (FLE) sur une quinzaine d'étudiants en 2005 et 2006. Ce type de recommandation est particulièrement adapté aux étudiants FLE qui se focalisent trop sur un savoir qu'ils ont déjà acquis d'une façon suffisante, plutôt que sur la pratique. Ainsi, leur situation d'étudiant étranger en France constitue une opportunité qu'ils ne saisissent pas suffisamment. Il apparait en effet qu'une trop grande exigence est portée sur la perfection de leur maitrise de la langue. Cet excès, souvent favorisé par les enseignants de leur vécu passé, amène des comportements particuliers :

- une posture selon laquelle la population française accorde un haut niveau d'exigence à la langue parlée,
- une retenue dans les dialogues qu'ils peuvent entretenir quotidiennement avec la population française,
- un droit insuffisamment accordé à l'erreur qui leur permettrait d'apprendre en se faisant corriger.

Ainsi, la particularité essentielle du dispositif mise en évidence par PAD, consiste à affirmer l'enseignant comme un facteur de constitution des connaissances et non de mise en œuvre. Ainsi, l'enseignant ne fait pas partie des participants de $\mathcal{A}_e$. Ceci accentue la particularité d'une situation d'apprentissage que les engagements déterminants concernent l'apprenant comme un double participant. La création d'engagements avec des tiers constitue un moyen de rendre concret ce dédoublement lorsque cela s'avère nécessaire.

Le processus d'évaluation s'apparente à un examen de compétence clinique utilisé dans l'enseignement de la médecine. La représentation de ce processus par une combinaison de deux PAD met en évidence l'intérêt du QCM. Le QCM permet de placer l'apprenant dans la situation contradictoire d'être à la fois participant aux décisions et objet de la décision au moment de l'évaluation. Cette situation, appelée washback effect, permet de vérifier si le comportement décisionnel de l'apprenant est réellement basé sur la mobilisation de ses connaissances ou si d'autres enjeux apparaissent. Par les réponses au QCM, l'enseignant dispose alors de données objectives qui lui permettent de rechercher les causes d'une mauvaise constitution des connaissances de l'apprenant et d'émettre les recommandations appropriées. Cette recherche de données subjectives complète l'évaluation. PAD montre donc que le QCM constitue un réel modèle d'évaluation utile dans une démarche d'aide à la décision.

5 Discussion

Une première constatation s'impose sur la séparation entre client et décideur Il s'agit d'une particularité en matière d'aide à la décision qui considère un seul et même individu. On notera que le client est le mieux à même de garantir les hiérarchies qui s'opèrent dans les décisions. L'aide a donc pour rôle primordial d'offrir une garantie de respect de ces hiérarchies. C'est particulièrement le cas dans ce problème d'aide à l'apprentissage qui, dès les situations problématiques, s'appuie sur des priorités entre enjeux : le savoir dans $\mathcal{O}_q$ et la connaissance dans $\mathcal{O}_e$.

Il résulte de ces priorités que les problématiques décisionnelles Π_q et Π_e concerne les points de vue et non les actions potentielles comme c'est habituellement le cas en aide à la décision. Cependant, les problématiques de base de description, choix, tri et rangement

conservent tout leur intérêt. Il en résulte que l'aide à la décision peut être envisagée comme une combinaison de partition de l'ensemble des actions potentielles $\mathbb{A}$ et des points de vue $\mathbb{V}$. Vue de cette façon, la représentation des préférences présente dans l'operateur de synthèse d'information $\mathcal{R}$ du modèle d'évaluation $\mathcal{M}$ peut tirer avantage d'une recherche plus précoce de relations entre critères par l'intermédiaire des relations entre points de vue.

On remarque que les deux évaluations se complètent. Compte tenu qu'aucune recommandation n'est émise à l'issue de la première phase, le QCM constitue le pivot du dispositif. Il sert donc à établir un état de la situation de l'apprenant pour pouvoir réellement rechercher des pistes d'aide et émettre des recommandations adaptées. Il constitue donc un élément déterminant d'analyse et de construction de $\mathcal{P}_e$ et Γ_e. Ici, l'aide à la décision apporte un appui à des décisions et un processus décisionnel en place. Cet appui peut porter sur une modification de $\mathcal{P}$.

Plus largement, cet enchainement de deux PAD tend à montrer qu'une évaluation de la structure du problème peut être déterminante dans une aide à des décisions déjà mises en œuvre. En effet, en ne s'appuyant sur aucun opérateur de synthèse d'information $\mathcal{R}_q$, $\mathcal{M}_q$ est dénué de toutes préférences pour ne s'attacher qu'à fournir une représentation fine d'un état des connaissances. En précédant $\mathcal{P}_e$ et Γ_e, le modèle $\mathcal{M}_q$ peut s'apparenter à un modèle d'évaluation du problème qui compléterait le modèle $\mathcal{M}$ d'évaluation des actions potentielles $\mathbb{A}$. Ce modèle d'évaluation du problème permet d'envisager une construction argumentée de $\mathcal{P}$ et Γ, par un complément de validation, tout comme le modèle $\mathcal{M}$ permet, par plusieurs types de validations, de produire des recommandations significatives, applicables et légitimes. De ce fait, ce modèle d'évaluation est de nature qualitative. Cependant, il peut s'appuyer sur un dispositif numérique comme c'est le cas ici avec le QCM. On notera dans le cas de l'apprentissage, comme c'est révélé par la succession des deux PAD, que l'évaluation quantitative précède l'évaluation qualitative. La première évaluation vise à identifier la structure des connaissances et la seconde à intervenir sur cette structure via les engagements. Il existe donc une structure de décision. L'ordre des évaluations qualitative et quantitative résulte donc essentiellement de la constitution du problème de décision. Un problème présentant une structure établie nécessitera une évaluation fine donc quantitative, alors qu'un problème non structuré nécessitera une évaluation qualitative. Cette différence de structure conduit à envisager l'aide à la décision selon une approche descriptive ou constructive.

6 Aide à la décision, entrepôts et exploration de données

Les liens entre artefacts de PAD (figure 1) permettent de mieux identifier des phénomènes imprévus dans la conduite des décisions qui ne résultent pas de la construction de l'aide. Ainsi, le phénomène de washback peut être vu comme un lien partant des alternatives *A* de $\mathcal{M}$ vers les points de vue $\mathbb{V}$ de Γ. De ce fait, la complémentarité des deux modèles $\mathcal{M}_q$ et $\mathcal{M}_e$ permet de voir les deux PAD non pas comme une séquence mais comme un aller-retour au sein d'un seul et même PAD (figure 3).

A l'aller, l'orientation des liens traduit la nature constructive de l'aide à la décision. Le retour permet un affinement du contenu de la représentation proche de la réalité du problème

c'est-à-dire la situation problématique $\mathcal{P}$. Il en va de même pour le modèle d'évaluation du problème qui reste à structurer.

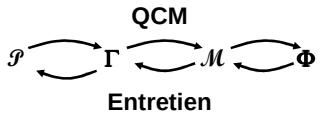


Figure 3 – Complémentarité en aller retour des PAD

La situation problématique fournit donc un cadre décisionnel commun structurant :

- la collecte de données opérationnelles structurées telle que le décrit le principe des entrepôts de données (rappelons d'ailleurs que le QCM décrit préalablement est supporté par une base de données) (Inmon, 1992),
- les objectifs de l'exploration des données tels qu'énoncé par les principes de fouille de données (Fayyad, 1996).

L'aide à la décision permet d'envisager un complément méthodologique synthétique spécifiquement décisionnel aux modes de structuration et d'exploration des données. Ce complément permet d'envisager une identification des schémas récurrents les plus pertinents parmi ceux qui peuvent apparaitre dans les données. Ainsi, l'aide à la décision se base fondamentalement sur des structures relationnelles, tel que :

- les engagements $\mathcal{S}$ au sein de la situation problématique $\mathcal{P}$ qui structurent les relations établies ou à établir entre participants $\mathcal{A}$ et enjeux $\mathcal{O}$,
- les problématiques décisionnelles Π au sein de la formulation de problème Γ qui mettent en relation des sous-ensembles d'actions potentielles,
- les opérateurs de synthèse d'information $\mathcal{R}$ au sein du modèle d'évaluation qui représentent les préférences comme des relations entre actions potentielles.

De leur coté, **les principes des entrepôts et de l'exploration de données permettent d'envisager des compléments en outils sur lesquels les cadres méthodologiques de l'aide à la décision peuvent s'appuyer** pour une compréhension des problèmes de décision dont la structure n'est pas toujours précisément définie.

7 Conclusion

A la lumière de cette réalité de l'apprentissage, il apparaît que l'aide à la décision concerne un problème de décision déjà structuré. L'aide apporte donc un appui à cette structure qui peut conduire à procéder à des ajustements. La première question concerne donc l'identification de cette structure. L'aide à la décision s'inscrit donc dans une logique descriptive et non pas seulement constructive, ce qui en soi constitue une particularité. Pourtant, PAD offre un cadre méthodologique adapté bien qu'il s'inscrive dans une démarche constructive. Un complément de représentation du problème de décision à PAD s'avère une piste d'investigation intéressante.

Les concepts d'entrepôts et de fouille de données concernent l'identification de connaissances sur la base de gros volumes de données opérationnelles. Ces concepts ont engendré la définition de nombreux outils.

Dans un cadre décisionnel, ces deux familles de concepts se retrouvent autour de la notion commune d'effets des décisions qui sont vus de façon spéculative en aide à la décision et selon un point de vue opérationnel avéré en entrepôts et fouille de données. Par ailleurs, ces deux cadres partagent une idée commune de relation entre objets de base qui sont établis ou recherchés selon la famille de concepts. L'aide à la décision peut donc apporter un éclairage en termes de structure décisionnelles qui permet de mieux délimiter les problèmes de décision abordés par les entrepôts et la fouille de données. En particulier, la structuration du problème posé par le client et géré par le décideur est une question déterminante. Les approches d'entrepôts et de fouille de données fournissent quant à elles de nombreux outils permettant d'identifier la structure du problème de décision.

Des perspectives mutuellement enrichissantes s'ouvrent alors de l'échange entre les deux familles de concepts.

Bibliographie

Astolfi, J.P. (1993). Comment les élèves apprenent-ils ?, *Sciences Humaines 32*.

Belton, V., T. Stewart (2002). *Muliple Criteria Decision Analysis: An Integrated Approach. Kluwer*, Academic.

Bouyssou, D., T. Marchant, M. Pirlot, P. Perny, A. Tsoukiàs, P. Vincke (2000). *Evaluation and Decision Models: A Critical Perspective*, Springer Verlag.

Bouyssou, D., T. Marchant, M. Pirlot, A. Tsoukiàs, P. Vincke (2006). *Evaluation and Decision Models: stepping stones for the analyst*, Springer Verlag.

Brady, A.M. (2005). Assessment of learning with multiple-choice questions, *Nurse Education in Practice 5*, 218-242.

Checkland, P., P. Scholes (1990). *Soft Systems Methodology in Action.* J. Wiley.

Djurić, M. (2008). Dealing with Situations of Positive and Negative Washback, *Scripta Manent* 4, 14-27.

Eden, C. (1988). Cognitive Mapping. *European Journal of Operational Research 36*, 1-13.

Faerber, R. (2004). *Caractérisation des situations d'apprentissage*, STICEF.

Fayyad, U., G. Piatetsky-Shapiro, P. Smyth, R. Uthurasamy (1996). *Advances in Knowledge Discovery and Data Mining*, AAAI Press/ The MIT Press.

Friend, C.K., A. Hickling (1987). *Planning Under Pressure: The Strategic Choice Approach.* Pergamon Press.

Hadji C. (1993). *L'évaluation, règles du jeu ; Des intentions aux outils*, Editions ESF.

Inmon B. (1992). *Building the Data Warehouse*. Wiley and Sons.

Keeney, R.L. (1992). *Value-Focused Thinking. A Path to Creative Decision Making.* Harvard University Press.

Labour, M. (2010). *MEDIA-REPERES. Une méthode pour l'explicitation des construits de sens au visionnage.* Habilitation à Diriger des Recherches, Université de Valenceinnes.

Ostanello A., A. Tsoukiàs (1993). An explicative model of public interorganizational interactions. *European Journal of Operational Research 70*, 67-82.

Roy, B. (1985). *Méthodologie Multicritère d'aide à la décision*. Economica.

Simon, H.A. (1977). *The new Science of management decision*, Prentice-Hall.

Sommer S. (2001). *La nécessaire interaction entre évaluation et processus d'apprentissage en langues*, ASP 34, http://asp.revues.org/1706.

Stamelos, I., A. Tsoukiàs (2003). Software evaluation problem situations, *European Journal of Operational Research 145*, 273-286.

Tsoukiàs, A. (2007). On the concept of decision aiding process. *Annals of Operations Research 154*, 3-27.

Tsoukiàs, A., H. Ralijaona (2012). Rural Road Maintenance in Madagascar: the GENIS project, in R. Bisdorff, L. Dias, V. Mousseau, M. Pirlot (eds.), *Evaluation and Decision Models: real case studies*, to appear with Springer-Verlag.

Wass, V., C. Van der Vleuten, J. Shatzer, R. Jones (2001). Assessment of clinical competence, *The Lancet 357*, 945-949.

Wimmers, P. F., (2006). *Developing Clinical Competence*, Thèse de doctorat, Université Erasmus de Rotterdam.

Summary

This paper presents a decision aid devoted the student in the context of language learning. This is based on the Multiple Choice Question (MCQ). The high structure of MCQ could be considered as a data warehouse and the management of the student response of MCQ as a data mining by the teacher. So we propose some links between the concepts of data mining, data warehouse and methodological framework of decision aiding.

Combination Framework of BI solution & Multi-agent platform (CFBM) for multi-agent based simulations

Truong Minh Thai*, Frédéric Amblard*
Benoit Gaudou*

* UMR 5505 CNRS, Institut de Recherche en Informatique de Toulouse
Université Toulouse 1 Capitole
tmthai@cit.ctu.edu.vn, frederic.amblard@ut-capitole.fr, benoit.gaudou@ut-capitole.fr

Abstract. Integrated environmental modeling in general and specifically Multi-agent-based modeling and simulation approach are increasingly used in decision-support systems with, as a major consequence, to manipulate and generate a huge amount of data for their functioning (parametrization, use of real data in the simulation, ...). Therefore there is a need to manage efficiently these data being either used or generated by the simulation. Practically, existing generalist simulation platforms lack database access and analysis tools and simulation outputs are usually stored as text files or spreadsheets to be manipulated later by dedicated tools. In this paper, we propose a solution to handle simulation models data, i.e. their outputs as well as corresponding real data. We designed a conceptual framework based on a combination of two components, a Business Intelligence (BI) solution and a multi-agent platform. Such a framework aims at managing simulation models data throughout the lifespan of the simulation, from its execution and its coupling with real data to the generation of simulation results in order to use the simulation model as an effective decision-support system with what-if scenarios.
Keywords: Decision-Support System, Multi-Agent Simulation, BI, Data warehouse, OLAP.

1 Introduction

Integrated socio-environmental modeling in general and the multi-agent based simulation approach applied to socio-environmental systems in particular are increasingly used as decision-support systems in order to design, evaluate and plan public policies linked to the management of natural resources (Laniak et al., 2013). The main idea behind such approaches is to combine and couple information available from different sources and scientific fields (like water management, climate sciences, sociology, economics). Such information mainly take the form of empirical data gathered from the field and of simulated models regarding some aspects of the studied phenomenon (for instance the behavior of related actors from the field). This information is usually combined in an ad hoc manner, combining database access and management of text files, and simulation results are often stored as text files, spreadsheets or xml files. The basic statement we can make is that currently, if the design and simulation of

models has benefited from the advances in computer science through the popularized use of simulation platforms (like Netlogo (Wilensky, 1999), GAMA (Taillandier et al., 2012)...), it is not the case for the management of data yet, which are still managed in an ad hoc manner, despite the advances in the management of huge datasets (data warehousing for instance). Such a statement is rather pessimistic if we consider recent tendencies toward the use of data-driven approaches in simulation aiming at using more and more data available from the field into simulated models.

Therefore, we stand that there is definitely a need for a robust data management solution of huge datasets in multi-agent based simulations and we will in this article, propose the first steps towards such a solution. This solution will combine two aspects. The first one deals with the status of data, as the proposed solution should be able to manage empirical data gathered from the studied phenomenon or system as well as simulated data produced by simulations considered as *in silico* experiments on the same system. The second aspect concerns the use of a Business Intelligence (BI) solution envisaged as a system of data warehouse and analysis tools. A data warehouse corresponds to a collection of data that supports decision-making processes (Inmon, 2005). Analysis tools may be data mining, statistical analysis, prediction analysis and so on. The features of a BI solution will help us to manage huge amount of historical data and make several analysis on such data.

In the following, we first present a state of the art of the links between these two systems (Section 2). We then present the global architecture of our combined framework (Section 3), before discussing its strengths and weaknesses (Section 4).

2 Related works and methodologies

Lot of works have been done to deal with huge amount of data and provide analysis tools in the field of Data Warehouse (DW) and BI tools. Many researches have been conducted on the use of DW and simulation to develop a decision support system or a prediction system. The combination of simulation tools and DW is widely increasingly used and applied in different areas. (Madeira et al., 2003) proposed a new approach dedicated to analyze and compare a large amount of output data from different experiments or similar experiments across different systems. They gathered data from raw data sources (text file or spreadsheet) into multidimensional database and use OLAP tools to analyze or compare them. (Sosnowski et al., 2007) proposed a data warehouse for collecting and analyzing simulation results. Although this is only an application of OLAP technologies to a special problem, namely the system dependability evaluation using fault injections into running programs, this work demonstrates that dimensional tables can store several hundreds of thousand records of simulation results. The multidimensional database of the simulation results can be analyzed, mined and enabled reporting by using standard OLAP tools. In (Vasilakis et al., 2008) and (Ehmke et al., 2011), the authors developed systems involving simulation models, multidimensional database and OLAP tools. These systems are called "decision support system" or "forecast system". Although these researches only solve specific problems, they demonstrated a potential in gathering and analyzing simulation results by using data warehouse and OLAP. (Mahboubi et al., 2010) presents a research on using multidimensional model to develop data warehouse in systems, coupling complex simulation models such as biological, meteorological and so on. These models are usually coupled models and they generate a huge amount of output data.

Most articles cited above focus on the application of data warehouse and OLAP technologies for collecting and analyzing output of simulations with huge amount of data. Although authors have successfully combined simulation models, data warehouse and OLAP technologies to solve concrete problems, they only focused on building specific simulation models and multidimensional schemas for solving specific problems. The state of the art demonstrates therefore the practical possibility and the usefulness of the combination of simulation, data warehouse and OLAP. It also shows the potential of a general framework that is, to our knowledge, not yet proposed in the literature.

Hence, the purpose of our research is to design a logical framework for the combination of both BI and MABS solutions. The framework can help us to exploit the useful features of BI solution and multi-agent platform to build a model, manage multiple models (especially their inputs/outputs) and analyze results of simulations. In particular, the framework should support a distributed and collaborative environment.

Our proposed Combination Framework of BI solution and Multi-agent platform, named CFBM for short, is detailed in the next session.

3 Combination Framework of BI solution & Multi-agent platform (CFBM)

3.1 Computer simulation system

Fishwick (1997) defined that computer simulation is the discipline of designing a model of an actual or theoretical physical system, executing the model on a digital computer, and analyzing the execution output. On the basis of Fishwick's definition, we can define a computer simulation system as below:

> *A Computer simulation system is a computation system with four components and the intercommunications between them:*
> – *Model design tool: a software environment that supports a modeling language, notations and user interface for modeling an actual or theoretical physical system.*
> – *Model execution tool: a software environment that can run models.*
> – *Execution analysis tool: a software environment that supports statistical analysis features for analysis of output data of models.*
> – *Database tool: a software environment that supports appropriate database and database management features for overall components in the system.*

The components of a computer simulation system and their intercommunications are illustrated in Figure 1. From this figure, we have designed a conceptual framework for multi-agent based simulations. The framework is illustrated in figure 2. In this framework, we use a BI solution as a database tool, a multi-agent platform as model design tool and model execution tool. For execution analysis, we can either use OLAP analysis tool or use analysis features as an external plug-in of the multi-agent platform (for instance R scripts).

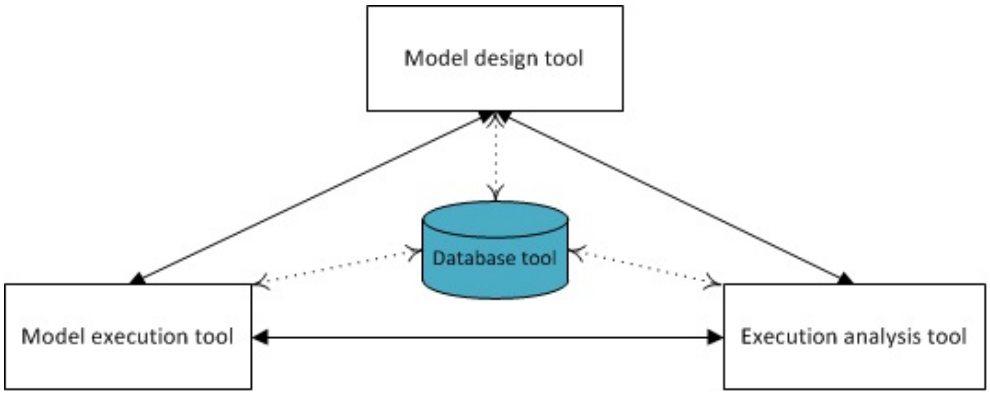


FIG. 1 – *Computer simulation system.*

3.2 CFBM architecture

In Figure 2, the CFBM is divided into three systems with seven layers. The function of each part is detailed below.

3.2.1 Simulation system

The simulation system is composed of a multi-agent platform with a relational database. This system helps to implement simulation models and to handle the various models and their input/output data. The multi-agent platform supports the model design and its execution.

The simulation is composed of the following parts. The **Simulation interface** is an user environment that helps the modeler to design and implement his models, execute them and visualize simulation results. **Multi-agent simulation models** are a set of multi-agent based models and can be considered either as alternative models of the same phenomenon or as different versions of the same model. They are used to simulate phenomena that the modeler aims at studying. A multi-agent simulation model reads input values from a database and store simulation results via SQL agents. The **SQL Agent** is a particular kind of agents which supports SQL features to query data from a relational database. As these agents are integrated into the simulation, they play an intermediary role between the simulation and the database layers.

The data source layer is composed of two relational databases. **Real data** is a database used to store empirical data gathered from the target system that are needed for simulation and analysis. This information can be used as input data for the simulation phase or as validation data for the analysis phase. **Simulation data** is a database used to manage the simulation models, simulation scenarios and output results of the simulation models.

In principle, the real data and the simulation data are separated, even if they are related to the "same objects" (one being from the real world, the other from the simulated world) and that both will be used to answer questions from the decision-maker in many cases. One of the important feedback to make to the decision-maker is related to the quality of the given answer, and a part of this quality can be represented in the rate of simulated versus empirical data used

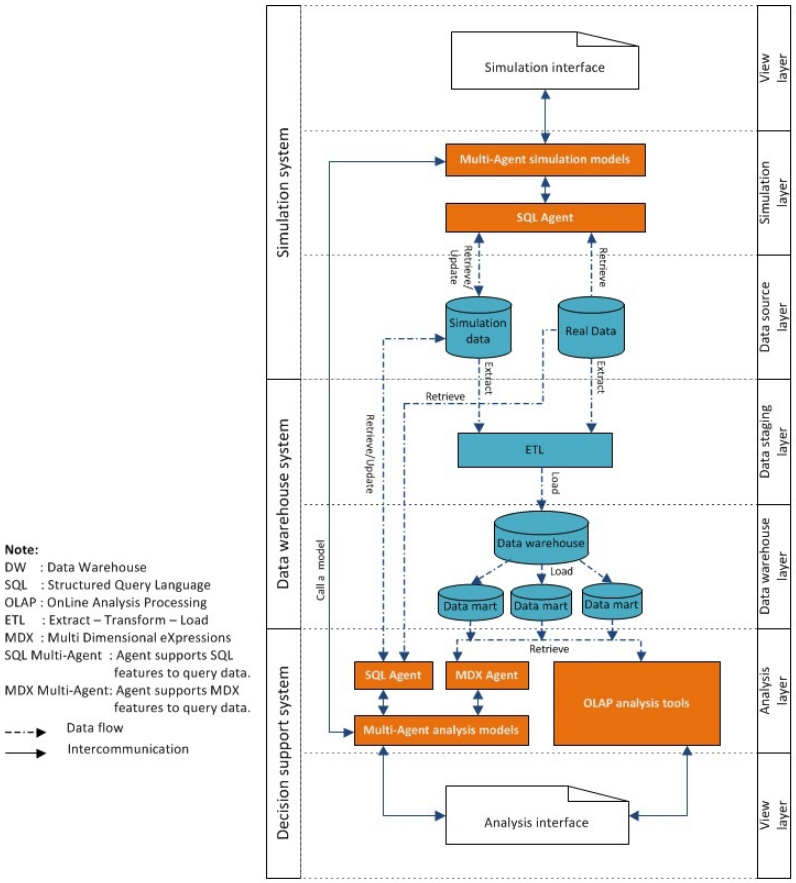


FIG. 2 – *Combination Framework of BI solution and Multi-agent platform (CFBM) architecture.*

to answer the question. In any case, these two data sources will be used to feed the second part of the framework, namely the **Data Warehouse system**.

3.2.2 Data warehouse system

The data warehouse system is used to store historical data about the actual system and simulation data. It is divided into three parts.

ETL is a set of processes with three responsibilities. First, ETL (Extract-Transform-Load) extracts data from the real and the simulation databases. Second, it transfers extracted data from these two sources into an appropriate data format. Finally, it loads transferred data into the data warehouse. **Data warehouse** is a relational database used to store historical data which are loaded from simulation system by ETL. A **Data mart** is a subset of data stored in the data warehouse. It is used to gather and store data for one or many analysis. We can create several

data marts depending on our analysis requirements. Data marts are particularly useful to help users to improve the performance of analytic processes.

We propose to use a BI solution with OLAP technologies to implement the data warehouse system. OLAP is the most popular way to exploit information in a data warehouse (Golfarelli and Rizzi, 2009). OLAP technologies help experts to analyze multidimensional databases having a complex analysis requirement, with some common operators of OLAP such as roll-up, drill-down, slide and dice, analysts can aggregate data, navigate data to more details, slice data into specific datasets and dice data in different viewpoints easily. In addition, our framework has been designed modular enough to use other technologies instead of OLAP. In particular, OLAP could be replaced by an open-source alternative.

3.2.3 Decision support system

In our architecture, the decision support system part is a software environment supporting analysis, decision-making features and visualization of results. There exist many tools that could be used to implement such a decision support system. In our framework, we give the user the opportunity to use either existing OLAP analysis tools, or a multi-agent platform with analysis features or a combination of both of them.

The decision support system is thus composed of four parts. The **Analysis interface** is an user interface that is used to handle analysis models and visualize results. It should be generic enough to be used by the two decision support tools: the multi-agent and the OLAP analysis tools.

Multi-agent analysis models are a set of multi-agent-based analysis models. They are created based on analysis requirements and handled via analysis interface. One of the key points of our framework is the fact that multi-agent analysis models and multi-agent simulation models are implemented with the same modeling language in the same platform hence they can communicate easily with each others. The **MDX Agent** is a special kind of agent which supports MultiDimensionnal eXpressions (MDX) features to query data from a multidimensional database. Analysis model can access data from relational database or multidimensional database by using SQL agent (the same agent as in the simulation system) or MDX agent appropriately.

OLAP analysis tools is an analysis software that supports OLAP operators.

The key points of the CFBM architecture are that the model design, the model execution and the execution analysis functions are integrated into one multi-agent platform, all related data are managed by data warehouse, and analysis models and simulation models can interact with each others. They are very useful features for building a computer simulation system.

4 Discussion: Advantages and disadvantages of CFBM

CFBM is a conceptual framework that we have designed to manage interactions between multi-agent based simulations and large amounts of data. The framework has some advantages listed below.

CFBM is an open and modular architecture. As for the implementation of the CFBM, we can use any BI solution and multi-agent platform depending on which technology is the most adapted. We can choose an open source software or a commercial one. While choosing

such a software, however we should consider important points such as the selected software adaptation with our requirements on model design, model execution and execution analysis, the integration seems feasible.

CFBM can be used in a distributed environment. Assuming many modelers and analyzers working together on the same project but being located in different places, we can setup a simulation system and an analysis system in each location and all the data from each location can be integrated and shared via a centralized data warehouse.

CFBM allows to handle a complex simulation system. In particular, we can build several simulation models to simulate the same phenomenon, conduct lot of simulations on each of them and compare simulation results of each model (e.g. to determine which one is better for which parameters value domain). In this case, it is very difficult for modelers to manage, analyze or compare output data of simulations if modelers do not have an appropriate tool. With help of SQL agents and relational data in simulation system part of CFBM, modelers can create a database to manage and store simulation models, scenarios and outputs of simulation models easily. In addition, ETL will load all the outputs of a simulation and appropriate empirical data into data warehouse then it also allows modeler to deal with several analyses to compare simulation results of a simulation in different scenarios as well as to validate simulation models with empirical data.

CFBM is a combination framework of BI solution and multi-agent platform, it supports simulation modeling, database management, database analysis and interaction environment between them hence CFBM is suitable to implement many kinds of system such as what-if simulation systems, prediction/forecast systems or decision support systems.

Although CFBM has many advantages, it still has some drawbacks such as:

- It is very hard to implement CFBM. Because CFBM is a mix system of different applications, multi-agent platform, BI solution and analysis tools, the integration of all these software components and expertise is a complex work, however similar complexity is actually present when working on integrated modelling.
- CFBM is not suitable for building a simple simulation system such as one or two models working with a small amount of data because it may take more time and workforce than other approaches.

5 Conclusion and future works

In this paper, we proposed a conceptual framework adapted to multi-agent-based simulations with high volume of data. Not only CFBM supports experts to model a phenomenon and to execute the models via a multi-agent based simulation platform, but our framework also helps experts to manage sets of models, input and output of models, to aggregate and analyze output data of models via data warehouse and OLAP analysis tools. The key features of CFBM are that it supplies four components (model design, model execution, execution analysis and database management) and it also assures communication in-between these components in a computer simulation system. The most important point of CFBM is the integration power of data warehouse, OLAP analysis tools and a multi-agent based platform that is useful to develop complex simulation systems such as what-if simulation system, prediction/forecast system or decision support system with a large amount of input/output data.

As future works, on the one hand we will choose an appropriate BI solution and a multi-agent based platform to implement our framework. The prototype of the architecture is being developed using the open-source GAMA platform (Taillandier et al., 2012). On the other hand, we will apply our framework on the management of simulation models and their input/output, analysis output simulations and evaluation models of two projects (the DREAM Project [1] and the MAELIA Project [2]). These two research projects aim at building integrated models, dealing thus with several models (and several versions of the same models) and huge quantities of data.

References

Ehmke, J. F., D. GroSShans, D. C. Mattfeld, and L. D. Smith (2011). Interactive analysis of discrete-event logistics systems with support of a data warehouse. *Computer Industry 62*(6), 578–586.

Fishwick, P. A. (1997). Computer simulation: Growth through extension. *Transactions of the Society for Computer Simulation International 14*(1), 13–23.

Golfarelli, M. and S. Rizzi (2009). *Data Warehouse Design: Modern Principles and Methodologies*, Chapter 1. Mc Graw Hill.

Inmon, W. H. (2005). *Building the Data Warehouse, 4th Edition* (4th Edition ed.). Wiley Publishing, Inc.

Laniak, G. F., A. E. Rizzoli, and A. Voinov (2013). Thematic issue on the future of integrated modeling science and technology. *Environmental Modelling & Software 39*(0), 1–2.

Madeira, H., J. Costa, and M. Vieira (2003). The olap and data warehousing approaches for analysis and sharing of results from dependability evaluation experiments. In *IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'03)*, Los Alamitos, CA, USA, pp. 86. IEEE Computer Society.

Mahboubi, H., T. Faure, S. Bimonte, G. Deffuant, J. P. Chanet, and F. Pinet (2010). A multidimensional model for data warehouses of simulation results. *IJAEIS 1*(2), 1–19.

Sosnowski, J., P. Zygulski, and P. Gawkowski (2007). Developing data warehouse for simulation experiments. *Lecture Notes in Computer Science (LNCS) 4585*, 543–552.

Taillandier, P., A. Drogoul, D. A. Vo, and E. Amouroux (2012). GAMA: a simulation platform that integrates geographical information data, agent-based modeling and multi-scale control. In Springer (Ed.), *The 13th International Conference on Principles and Practices in Multi-Agent Systems (PRIMA)*, Volume 7057 of *Lecture Notes in Coputer Science*, India, pp. 242–258.

Vasilakis, C., E. El-Darzi, and P. Chountas (2008). A decision support system for measuring and modelling the multi-phase nature of patient flow in hospitals. *Studies in Computational Intelligence 109*, 201–217.

Wilensky, U. (1999). Netlogo. Technical report, Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

1. http://www.ctu.edu.vn/dream/
2. http://maelia1.wordpress.com/

Variations autour du "palmarès des villes étudiantes" du magazine l'Etudiant

Antoine Rolland* & Jérôme Kasparian**

* Laboratoire ERIC - Université Lumière Lyon II
av Pierre Mendes-France, 69676 BRON Cedex
antoine.rolland@univ-lyon2.fr
** Université de Genève, GAP-Biophotonics, 20 rue de l'Ecole de Médecine,
CH-1211 Geneva 4, Switzerland - jerome.kasparian@unige.ch

Résumé. Le magazine "L'Etudiant" a publié en septembre 2012 un classement des "villes où il fait bon étudier", présentant un palmarès multi-critère des 41 villes françaises de plus de 8000 étudiants. Nous proposons ici une étude visant à analyser ce palmarès à travers des méthodes d'agrégation diversifiées, afin de faire ressortir les invariants d'un tel classement de villes, et au contraire des effets directement liés à l'utilisation de telle ou telle méthode.

1 Introduction

Le magazine "L'Etudiant" a publié en septembre 2012 un classement des "villes où il fait bon étudier" (cf Bertereau et al. (2012)), présentant un palmarès multi-critère des 41 villes françaises [1] de plus de 8000 étudiants. Il faut saluer ici, même s'il n'est pas totalement abouti, l'effort de transparence effectué par le magazine : l'ensemble des classements des villes sur chacun des critères est disponible, ainsi que le classement global, et, sur simple demande à l'auteur du dossier, les pondérations de chaque critère. Les valeurs des données brutes et la méthode d'agrégation retenue pour obtenir chaque classement monocritère ne sont pas indiquées, mais les sources de chaque indicateur sont mentionnées. Nous ne critiquerons donc pas ici le choix des indicateurs retenus, considérant que cela concourt de l'expertise du magazine, qui sait ce qui peut caractériser une "ville où il fait bon étudier". Nous pouvons aussi noter positivement la répartition proposée en trois catégories (grandes villes, villes moyennes, petites villes) des villes retenues, ce qui permet de comparer ce qui est comparable.

Notre objectif ici est de proposer des analyses et des méthodes d'agrégation diversifiées afin de faire ressortir les invariants d'un tel classement de villes, et au contraire des effets directement liés à l'utilisation de telle ou telle méthode. Cette étude critique a par exemple déjà été proposée pour le classement de Shangai par Billaut et al. (2010) ou pour l'indice de bien-être de l'OCDE par Kasparian et Rolland (2012). Pour cela, nous allons dans la partie suivante présenter plus précisément le classement effectué par le magazine l'Etudiant. En partie 3, nous utiliserons une approche purement statistique pour décrire les différentes villes. Puis dans les

1. En fait les parties françaises des "unités urbaines" définies par l'INSEE
http ://www.insee.fr/fr/methodes/default.asp ?page=definitions/unite-urbaine.htm

parties suivantes nous analyserons les résultats obtenus à l'aide de méthodes d'agrégation multicritère qualitatives en partie 4, ou ordinales en partie 5. Enfin nous nous interrogerons sur le lien entre les classements obtenus et l'attractivité réelle des villes considérées (partie 6).

2 Le palmarès du magazine l'Etudiant : description

Le palmarès des "villes où il fait bon étudier" du magazine "l'Etudiant" consiste à comparer à l'aide d'un classement multicritère les 41 villes de France ayant plus de 8000 étudiants. D'après les auteurs, neuf thèmes (études, rayonnement international, sorties, culture, sports, transports, logement, environnement, emploi) et 37 indicateurs ont été retenus. La liste de ces indicateurs est présentée dans l'enquête, mais les modalités prises par ces indicateurs ne sont pas précisées. Les villes ont ensuite été classées sur chacun des neuf thèmes. La ville classée première reçoit alors un score de 41, la deuxième un score de 40, jusqu'à la dernière qui reçoit un score de 1. Il est à noter la présence éventuelle d'ex-aequo. Ces neuf thèmes seront appelés critères dans la suite de l'article.

En ce qui concerne la méthode d'agrégation retenue, nous pouvons lire dans un encart présentant la méthodologie de l'enquête que *" Nous avons attribué [aux critères] des coefficients (de 1 à 4) et nous avons créé 4 catégories - la formation (classements études et rayonnement international), la vie étudiante (classements sorties, culture et sports), le cadre de vie (classements logement, transports et environnement) et l'emploi (classement emploi) - de poids équivalent. Cela nous a permis de classer les 41 unités urbaines françaises de 8.000 étudiants et plus."* Nous sommes donc en présence d'une agrégation par moyenne pondérée sur les notes obtenues sur chaque critère.

Le classement final présenté dans le tableau 1 a fait l'objet dans le journal d'une présentation séparée par catégories de taille d'agglomération, détaillant les métropoles (plus de 400.000 habitants), les grandes villes (entre 250.000 et 400.000 habitants) et les villes moyennes (moins de 250.000 habitants).

3 Approche statistique

Nous proposons dans cette partie une approche purement statistique de description des résultats obtenus. Nous avons effectué une analyse en composantes principales normée sur les 41 villes. Les résultats sont détaillés sur les graphes 1 et 2. Nous avons retenu les deux premiers axes, qui expliquent 58 % de la variance totale. Nous pouvons voir sur le graphe des critères que le premier axe s'expliquent par l'opposition entre d'une part les critères "international", "culture", "emploi", "études" d'une part, et les critères "sport" et "logement" d'autre part. Le deuxième axe est marqué par le critère "environnement" et, dans une moindre mesure, le critère "sortie". Le classement final est assez fortement corrélé avec le premier axe. L'analyse du graphe des individus montre bien que la taille de la ville joue un grand rôle dans le classement : cela conforte le fait qu'il est bon de présenter les vainqueurs de chaque catégorie. Schématiquement, nous pouvons dire qu'il y a d'un côté les grands pôles universitaires (Paris, Lyon, Aix-Marseille, Lille...), correspondant à des villes dynamiques, qui proposent une grande diversité d'études, de bons débouchés, et les services et loisirs d'une grande métropole ; d'un autre côté les petites villes (Chambéry, Limoges, Poitiers, Dijon...) qui n'ont pas ces avan-

tages, mais qui où le logement est peu cher, les installations sportives plus accessibles et qui se distinguent entre elle par leur environnement. Enfin, au milieu se trouvent les villes moyennes (Nancy, Rouen, Grenoble, Montpellier...) qui pour certaines rivalisent avec les grandes métropoles et pour d'autres sont plus proches des petites villes [2]. Un test de Kruskal-Wallis nous montre d'ailleurs que le classement final dépend de la taille de la ville (p-value de 0,1 %).

4 Approche multicritère quantitative

Nous avons noté en partie 2 que la méthode d'agrégation retenue pour obtenir un classement final à partir du classement des alternatives est une moyenne pondérée. Or il apparait que cet agrégateur est assez particulier, et limite le type de solutions mises en avant (voir Grabisch et al. (2009); Marichal (2009) pour une étude complète). Ceci est d'autant plus vrai que les valeurs prises par les villes sur chacun des critères sont des valeurs fondamentalement ordinales, et non quantitative : par exemple, les valeurs des indicateurs peuvent être très proches pour les villes classées première et deuxième sur un critère, et très éloignées pour les villes classées troisième et quatrième, mais au final l'écart sera le même entre le rang 1 et le rang 2, et entre le rang 3 et le rang 4.

Dans cette partie, nous allons donc nous attacher à interroger le classement final obtenu suivant trois axes :

- tout d'abord, nous allons mesurer la robustesse du classement obtenu aux variations des pondérations des critères. En effet, ceux-ci sont assez subjectifs, et il est très difficile de se rendre compte a priori de l'influence de telle ou telle valeur d'un poids sur le classement final.
- ensuite, nous évoquerons une méthode fondée sur l'optimisation multicritère, en appliquant la méthode TOPSIS (Hwang et Yoon (1981)).
- enfin, prenant acte que les valeurs obtenues ne sont que des valeurs ordinales, nous allons appliquer une méthode purement ordinale d'agrégation des critères, issue de la méthode ELECTRE (Roy (1996)).

4.1 Analyse de robustesse

L'analyse de robustesse du classement obtenu consiste à étudier les variations du classement final en fonction des variations des pondérations affectées aux critères. Pour cela, nous pouvons avoir deux visions complémentaires :

1. étant donné une ville, quel est le jeu de poids qui optimise son classement ? Autrement dit, existe-t-il un jeu de poids permettant de classer une ville donnée en première position, et sinon, quel est le meilleur classement auquel cette ville peur prétendre ?
2. Dans l'espace (à 9 dimensions) des critères, nous pouvons observer quelle ville est classée première pour chaque jeu de poids : la compilation de ces classements nous indique quelles sont les villes les plus souvent classées en première position si les poids sont tirés aléatoirement.

2. Il est à noter que ces villes (au centre du graphe) sont en réalité assez mal représentées par la projection sur le plan des deux premiers axes factoriels

Dans le premier cas, nous avons utilisé un algorithme de programmation linéaire afin d'optimiser le classement de chacune des villes. Dans le second cas, nous avons utilisé une méthode de Monte-Carlo en effectuant 400.000 tirages aléatoires de jeux de poids. Un tel nombre de tirages nous assure d'obtenir un intervalle de confiance à $\pm 1\ {}^{0}/_{00}$. Les résultats sont présentés dans le tableau 2. Nous pouvons constater que si 28 des 41 villes peuvent être classées premières avec un jeu de poids *ad hoc*, seules trois villes sont classées premières dans plus de 10% des cas (Grenoble, Toulouse et Montpellier) et trois autres dans plus de 1% des cas (Aix-Marseille, Poitiers, Bordeaux). Cela conforte le choix de l'Etudiant d'avoir distingué Toulouse, Montpellier et Poitiers chacune dans leur catégorie respective. Il est cependant à noter que le résultat obtenu pour le classement général (Toulouse première devant Grenoble) est trois fois moins probable que l'inverse (Grenoble classée première).

4.2 Approche de l'optimisation multicritère

Comme indiqué précédemment, la moyenne pondérée n'est qu'un des opérateurs possibles d'agrégation pour obtenir un score global à partir de plusieurs critères. D'autres méthodes, telle que le minimum, la moyenne ordonnée pondérée (OWA) ou l'intégrale de Choquet amènent à valoriser d'autres profils d'alternatives (Grabisch et al. (2009); Marichal (2009) *op. cit.*). Cependant ici, en l'absence de dialogue avec les promoteurs du palmarès, il est délicat de déterminer quel type de profil devrait être privilégié parmi les villes étudiantes, et donc encore plus délicat de fixer des paramètres ces méthodes.

Nous proposons ici d'employer une approche issue de l'optimisation multicritère à travers la méthode TOPSIS (Hwang et Yoon (1981)). TOPSIS est basée sur le fait que l'alternative classée première doit être en même temps à la plus petite distance possible du point idéal (ici celui qui serait classé premier sur tous les critères) et à la plus grande distance du point anti-idéal (celui classé dernier sur tous les critères). L'utilisation de la distance euclidienne vise à favoriser les individus ayant des scores plutôt équilibrés sur les différents critères par rapport à des individus pouvant avoir de très bons scores, mais également de très mauvais.

Nous pouvons constater en regardant les résultats présentée dans le tableau 3 que les résultats sont assez cohérents avec le classement proposé par l'Etudiant. Les cinq premiers du classement sont globalement les mêmes. Les différences notables concernent Bordeaux, Brest, Poitiers, St-Etienne et Metz qui descendent de 4 à 9 places, et Orléans, Avignon, Chambéry, Toulon, Valenciennes et Douai, qui gagnent de 5 à 11 places. Il est difficile de donner une interprétation claire des différences obtenues suivant une méthode ou une autre. On peut cependant noter que la plupart des villes sont classées relativement de la même manière par les deux méthodes. Cela implique que les villes pour lesquelles il existe une grande différence entre méthodes (Poitiers par exemple) ne doivent pas être sur- ou sous-valorisée dans les conclusions de l'enquête.

5 Approche multicritère ordinale

Comme il a été noté précédemment, les données disponibles ont été traitées comme des données quantitatives, mais sont fondamentalement des données ordinales : pour chaque critère nous disposons du classement des villes sur ce critère. Le fait que le magazine l'Etudiant

ai traité ces données comme des données quantitative justifie les analyses de robustesses effectuées précédemment. Cependant, nous pouvons de manière adaptée utiliser une méthode ordinale, par exemple la méthode ELECTRE (que nous choisirons sans veto, et avec seuil de préférence égal à ϵ), à base de comparaison par paires. L'inconvénient des méthodes à base de comparaison par paire est de conduire à des relations qui peuvent ne pas être transitives. Cependant, ces méthodes conduisent à des résultats qui sont tout de même exploitables. Le principe de la méthode ELECTRE consiste à comparer deux à deux les villes, puis à calculer un indice dit "de concordance" entre une ville a et une ville b correspondant à la somme des degrés d'importance des critères où a est mieux classée que b. Dans le cas présent, nous avons pris des importances égales aux pondérations retenues par l'Etudiant dans sa méthode de calcul. Il s'agit ensuite de déterminer une relation de préférence par une coupe des indices de concordances supérieurs à un certain seuil. Dans le cas présent, nous avons choisi le seuil minimisant le nombre d'ex-aequo dans le classement obtenu. Il a est donc à noter que la relation obtenue peut ne pas être transitive, c'est à dire qu'une ville a peut être préférée à une ville b, la ville b à une ville c, et la ville c à la ville a ! Cela met en lumière que certaines villes possèdent des atouts sur des critères très divers, et sont de ce fait difficile à comparer entre elles.

Afin de présenter des résultats exploitables, nous avons choisi de séparer les villes par taille, et ainsi d'obtenir dans les tableaux 4, 5 et 6 des relations de préférences à l'intérieur de chaque groupe. Ces tableaux montrent les difficultés qu'il peut y avoir à vouloir comparer des alternatives très dissemblables.

Parmi les conclusions que nous pouvons tirer de ces tableaux, retenons les suivantes :

- pour les métropoles, si Toulouse et Lyon semblent dominer le classement, il existe de nombreuses incomparabilités entre les différentes villes, ce qui tendrait à souligner qu'elles ont toutes des avantages et des inconvénients relativement limités les unes par rapport aux autres. Plutôt qu'une stricte relation de préférence entre les différentes villes, on voit apparaître un regroupement en trois classes : Toulouse, Lyon, Marseille et Grenoble dans une première classe, Bordeaux, Nantes, Nice et Paris dans une deuxième classe, puis Strasbourg, Lille, Toulon et Douai dans une troisième classe.
- pour les grandes villes, le classement semble plus solide, dans sa première partie tout du moins. On peut noter qu'Orléans n'est que très peu comparables aux autres villes, n'étant classé moins bon que de deux villes, mais n'étant préféré également qu'à deux villes. Les villes de la deuxième moitié du tableau (St-Etienne, Tours, Metz, Avignon) semblent former une seule classe de villes incomparables entre elles.
- le tableau des villes moyennes fait ressortir ici aussi, dans une certaine mesure, la singularité de Poitiers qui est la seule ville préférée à Dijon, tout en étant dominée par trois autres villes. Il est à noter que plusieurs cycles de préférences apparaissent, tels que Le Havre - Amiens - Reims.

6 Approche inverse

Le palmarès de l'Etudiant vise à ordonner les villes suivant leur attractivité auprès des étudiants suivant des critères définis. Cependant, force est de constater que les étudiants n'ont pas attendu la parution du journal pour se déterminer et ainsi "voter avec leurs pieds". Suivant l'exemple dans l'immobilier exposé par Alexandre et al. (2010), il peut être intéressant de rapprocher le classement effectué par l'Etudiant avec un indice d'attractivité des villes françaises

pour les étudiants, calculé à partir des données disponibles auprès de l'INSEE de la manière suivante :

$$Att = \frac{\text{Nombre d'étudiants}}{\text{Nombre de jeunes entre 15 et 29 ans}}$$

Cet indicateur permet, de manière grossière, de mesurer la capacité d'une ville à accueillir des étudiants au delà de son public naturel. Plus l'indice est élevé, plus la ville est réputée attractive. Les résultats sont présentés dans le tableau 7 et la figure 3. Le coefficient de corrélation est significativement non nul, sans être très élevé (0,43). Cela signifie qu'il y a une légère corrélation entre l'indice d'attractivité tel que nous l'avons défini et le classement de l'étudiant. Il est à noter que Poitiers est la ville en tête de l'attractivité, confirmant ainsi l'intérêt que lui porte l'Etudiant. De manière générale, les métropoles sont relativement moins bien classées par l'indice d'attractivité que par le magazine l'Etudiant : il semble que malgré toutes leurs qualités, elles n'arrivent pas à attirer une proportion significative d'étudiants extérieurs par rapport à leur population. Au contraire, les grandes villes (Rennes, Montpellier, Angers, Clermont, Nancy) et les villes moyennes (Poitiers, Besançon, Caen, Dijon) bien classées dans leurs catégories sont aussi bien classées en terme d'attractivités. Il est à noter l'attractivité particulière d'Amiens qui contraste avec sa 35^{eme} place dans le classement de l'Etudiant, qui peut s'expliquer comme étant le seul pôle universitaire de sa région.

Conclusion

Nous avons vu que des variations de paramètres et/ou de méthode, sans révolutionner totalement le classement proposé, peuvent cependant amener des changements significatifs. Il est alors dommage de présenter un seul classement, précis, comme étant un palmarès incontestable. Il semblerait préférable de procéder à des regroupements des villes en niveaux afin de pouvoir distinguer des ensembles relativement homogènes de villes comparables, sans privilégier particulièrement telle ou telle, et ce d'autant plus que les critères sont très divers et assez nombreux.

Références

Alexandre, H., F. Cusin, et C. Juillard (juillet 2010). L'attractivité résidentielle des agglomérations franç0aises. enjeux, mesure et facteurs explicatifs. *Université Paris-Dauphine, rapport interne*.

Bertereau, V., T. Brisson, M. Brochard, et P. Falga (2012). Dossier : le palmarès 2012-2013 des villes où il fait bon étudier. *L'Étudiant septembre*.

Billaut, J.-C., D. Bouyssou, et P. Vincke (2010). Should you believe in the shanghai ranking ? an mcdm view. *Scientometrics 84*, 237–263.

Grabisch, M., J.-L. Marichal, R. Mesiar, et E. Pap (2009). *Aggregation functions*, Volume 127 of *Encyclopedia of Mathematics and its Applications*. Cambridge, UK : Cambridge University Press.

Hwang, C. L. et K. Yoon (1981). *Multiple attribut decision making : Methods and applications : a state-of-the-art survey*. Springer-Verlag.

Kasparian, J. et A. Rolland (2012). Oecd's'better life index : can any country be well ranked ? *Journal of Applied Statistics 39*(10), 2223–2230.

Marichal, J.-L. (2009). *Aggregation functions for decision making, Decision-Making Process - Concepts and Methods*. ISTE/John Wiley.

Roy, B. (1996). *Multicriteria Methodology for Decision Aiding*. Kluwer Academic Publisher.

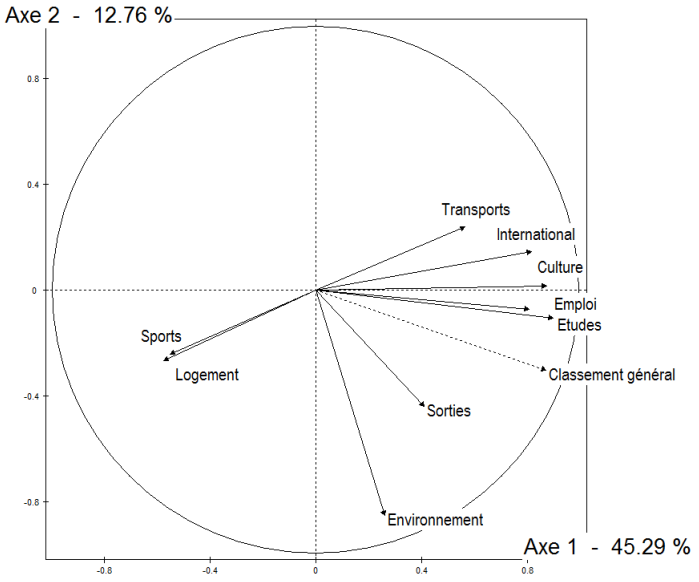


FIG. 1 – *Graphe des critères de l'analyse en composantes principales*

Summary

The magazine "L' Etudiant" published in September 2012 a ranking of the "cities where it well makes study ", presenting a multi-criterion ranking of 41 French cities of more than 8000 students. We propose here a study to analyze this ranking through diversified aggregation methods to highlight the invariants of such a city ranking, and the effects of the use of such or such method.

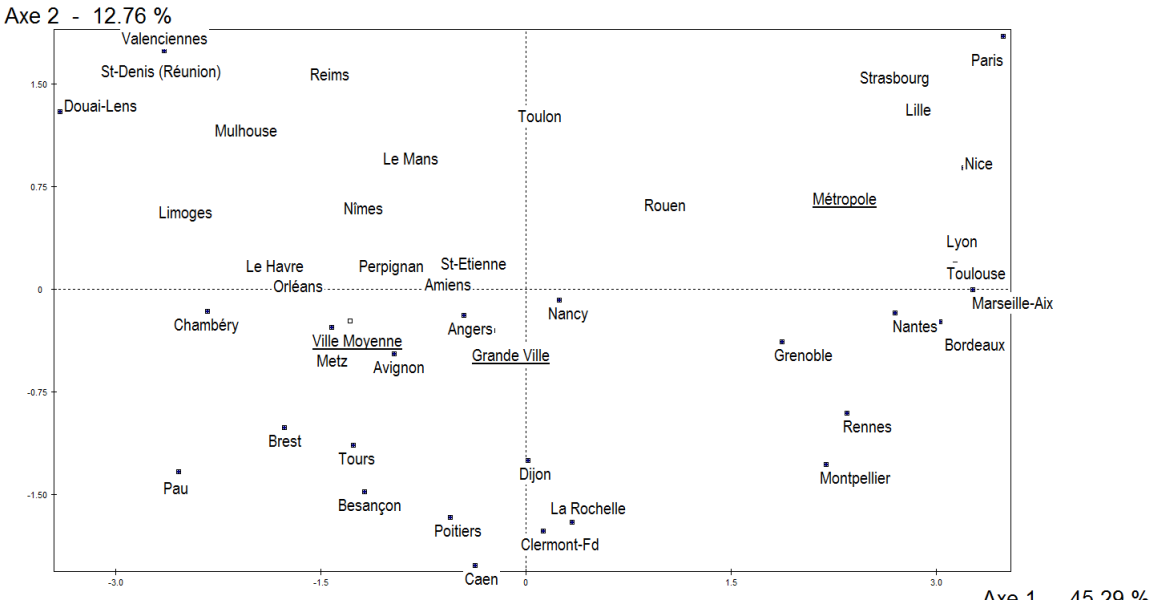


FIG. 2 – *Graphe des individus de l'analyse en composantes principales*

Villes	taille	Classement général
Toulouse	Métropole	1
Grenoble	Métropole	2
Montpellier	Grande ville	3
Marseille-Aix	Métropole	4
Lyon	Métropole	5
Bordeaux	Métropole	6
Nantes	Métropole	7
Rennes	Grande ville	8
Nice	Métropole	9
Paris	Métropole	10
Strasbourg	Métropole	11
Nancy	Grande ville	12
Rouen	Grande ville	13
Lille	Métropole	14
Clermont-Fd	Grande ville	15
Poitiers	Ville moyenne	16
Dijon	Ville moyenne	17
Caen	Ville moyenne	18
Angers	Ville moyenne	19
Orléans	Grande ville	20
Tours	Grande ville	21
St-Etienne	Grande ville	22
Besançon	Ville moyenne	23
Metz	Grande ville	24
Brest	Ville moyenne	25
Pau	Ville moyenne	26
La Rochelle	Ville moyenne	27
Avignon	Grande ville	28
Nîmes	Ville moyenne	29
Chambéry	Ville moyenne	30
Toulon	Métropole	31
Perpignan	Ville moyenne	32
Mulhouse	Ville moyenne	33
Le Mans	Ville moyenne	34
Amiens	Ville moyenne	35
Limoges	Ville moyenne	36
Reims	Ville moyenne	37
Le Havre	Ville moyenne	38
Douai-Lens	Métropole	39
St-Denis (Réunion)	Ville moyenne	40
Valenciennes	Grande ville	41

TAB. 1 – *Les différentes villes et leur classement par le magazine l'Etudiant*

	Villes	Meilleur rang	% premier
1	Amiens	5	0
2	Angers	1	ϵ
3	Avignon	2	0
4	Besançon	1	ϵ
5	Bordeaux	1	2,02 %
6	Brest	2	0
7	Caen	1	ϵ
8	Chambéry	2	0
9	Clermont-Fd	1	0,08 %
10	Dijon	1	ϵ
11	Douai-Lens	5	0
12	Grenoble	1	61,70 %
13	La Rochelle	1	ϵ
14	Le Havre	7	0
15	Le Mans	1	ϵ
16	Lille	1	ϵ
17	Limoges	1	ϵ
18	Lyon	1	0,42 %
19	Marseille-Aix	1	3,22 %
20	Metz	1	ϵ
21	Montpellier	1	13,07 %
22	Mulhouse	1	ϵ
23	Nancy	1	ϵ
24	Nantes	1	0,05%
25	Nice	1	ϵ
26	Nîmes	1	ϵ
27	Orléans	1	ϵ
28	Paris	1	ϵ
29	Pau	1	ϵ
30	Perpignan	7	0
31	Poitiers	1	2,15 %
32	Reims	7	0
33	Rennes	1	0,20 %
34	Rouen	1	ϵ
35	St-Denis (Réunion)	4	0
36	St-Etienne	1	ϵ
37	Strasbourg	3	0
38	Toulon	4	0
39	Toulouse	1	17,09 %
40	Tours	2	0
41	Valenciennes	24	0

TAB. 2 – *Les différentes villes et leur meilleur classement possible avec une moyenne pondérée et la portion de l'espace des poids où ces villes sont classées premières. Lecture : le meilleur classement de Nantes est premier, et cela arrive dans 0,05 % des jeux de poids. le meilleur classement possible de Strasbourg est troisième. Un ϵ signifie que la ville a été classé première dans moins de 10 cas (voire aucun) sur 400.000 essais.*

Ville	Classement Etudiant	Classement TOPSIS	Différence
Toulouse	1	1	0
Lyon	5	2	+3
Grenoble	2	3	-1
Montpellier	3	4	-1
Marseille-Aix	4	5	-1
Nantes	7	6	+1
Rennes	8	7	+1
Paris	10	8	+2
Nice	9	9	0
Bordeaux	6	10	-4
Strasbourg	11	11	0
Lille	14	12	+2
Rouen	13	13	0
Nancy	12	14	-2
Orléans	20	15	+5
Clermont-Fd	15	16	-1
Dijon	17	17	0
Tours	21	18	+3
Caen	18	19	-1
Angers	19	20	-1
Toulon	31	21	+10
Avignon	28	22	+6
Chambéry	30	23	+7
Poitiers	16	24	-8
Pau	26	25	+1
Besançon	23	26	-3
Nîmes	29	27	+2
Douai-Lens	39	28	+11
Brest	25	29	-4
La Rochelle	27	30	-3
St-Etienne	22	31	-9
Valenciennes	41	32	+9
Metz	24	33	-9
Perpignan	32	34	-2
Le Mans	34	35	-1
Mulhouse	33	36	-3
Amiens	35	37	-2
Reims	37	38	-1
Limoges	36	39	-3
Le Havre	38	40	-2
St-Denis (Réunion)	40	41	-1

TAB. 3 – *Les différentes villes et leur classement suivant la méthode TOPSIS et suivant le magazine L'Etudiant. Lecture : Rennes est classée* 8^{eme} *par l'Etudiant et* 67^{eme} *par TOPSIS, ce qui donne une place de mieux dans le deuxième classement que dans le premier.*

ville	Toulouse	Lyon	Marseille-Aix	Grenoble	Bordeaux	Nantes	Nice	Paris	Strasbourg	Lille	Toulon	Douai-Lens
Toulouse		1	0	1	1	1	1	0	1	1	1	1
Lyon	-1		0	0	1	1	1	1	1	1	1	1
Marseille-Aix	0	0		1	0	0	1	0	1	1	1	1
Grenoble	-1	0	-1		1	0	0	0	1	1	1	1
Bordeaux	-1	-1	0	-1		0	0	0	1	1	1	1
Nantes	-1	-1	0	0	0		0	0	0	1	1	1
Nice	-1	-1	-1	0	0	0		0	0	1	1	1
Paris	0	-1	0	0	0	0	0		0	1	1	1
Strasbourg	-1	-1	-1	-1	-1	0	0	0		0	1	1
Lille	-1	-1	-1	-1	-1	-1	-1	-1	0		1	1
Toulon	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1		1
Douai-Lens	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	

TAB. 4 – *Relation de préférence obtenue par la méthode ELECTRE sur les métropoles. Lecture : Toulouse est considérée comme préférée à Grenoble, mais Lyon n'est pas considérée comme préférée à Grenoble, ni l'inverse.*

Ville	Montpellier	Rennes	Rouen	Nancy	Clermont-Fd	Orléans	St-Etienne	Tours	Metz	Avignon	Valenciennes
Montpellier		1	1	1	1	1	1	1	1	1	1
Rennes	-1		1	1	1	1	1	1	1	1	1
Rouen	-1	-1		1	1	0	1	1	1	1	1
Nancy	-1	-1	-1	1	1	0	1	1	1	1	1
Clermont-Fd	-1	-1	-1	-1		0	1	0	1	1	1
Orléans	-1	-1	0	0	0		0	0	0	1	1
St-Etienne	-1	-1	-1	-1	-1	0		0	1	0	1
Tours	-1	-1	-1	-1	0	0	0		0	0	1
Metz	-1	-1	-1	-1	-1	0	-1	0		0	1
Avignon	-1	-1	-1	-1	-1	-1	0	0	0		1
Valenciennes	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	

TAB. 5 – *Relation de préférence obtenue par la méthode ELECTRE sur les grandes villes. Lecture : Rouen est préférée à Nancy mais n'est pas comparable à Orléans.*

Ville	Dijon	Angers	Caen	Poitiers	La Rochelle	Brest	Besançon	Nîmes	Pau	Chambéry	Perpignan	Mulhouse	Le Havre	Amiens	Le Mans	Reims	Limoges	St-Denis (Réunion)
Dijon		1	1	-1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Angers	-1		0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Caen	-1	0		0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Poitiers	1	-1	0		-1	-1	1	1	1	1	1	1	1	1	1	1	1	1
La Rochelle	-1	-1	-1	1		1	0	0	0	0	1	1	1	1	1	1	1	1
Brest	-1	-1	-1	1	-1		-1	0	-1	1	1	1	1	1	1	1	1	1
Besançon	-1	-1	-1	-1	0	1		1	0	1	1	1	1	0	1	1	0	1
Nîmes	-1	-1	-1	-1	0	0	-1		-1	1	1	1	1	1	1	1	1	1
Pau	-1	-1	-1	-1	0	1	0	1		-1	1	1	1	1	1	0	0	1
Chambéry	-1	-1	-1	-1	0	-1	-1	-1	1		1	0	1	1	1	1	0	1
Perpignan	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1		1	1	0	1	1	1	1
Mulhouse	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1		1	1	0	1	1	1
Le Havre	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1		1	-1	1	0	1
Amiens	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1		1	-1	1	1
Le Mans	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	1	-1		0	1	1
Reims	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	1	0		1	1
Limoges	-1	-1	-1	-1	-1	-1	0	-1	0	0	-1	-1	-1	-1	-1	-1		1
St-Denis (Réunion)	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	

TAB. 6 – *Relation de préférence obtenue par la méthode ELECTRE sur les villes moyennes. Lecture : Pau est préférée à Brest.*

Villes	attractivité	classement attractivité	classement étudiant	différence
Poitiers	1,32	1	16	+15
Rennes	1,31	2	8	+6
Montpellier	1,29	3	3	0
Angers	1,20	4	19	+15
Amiens	1,19	5	35	+30
Clermont-Fd	1,16	6	15	+9
Nancy	1,15	7	12	+5
Besançon	1,10	8	23	+15
Caen	1,07	9	18	+9
Dijon	1,05	10	17	+7
Strasbourg	1,01	11	11	0
Reims	0,95	12	37	+25
Grenoble	0,91	13	2	-11
La Rochelle	0,91	14	27	+13
Toulouse	0,88	15	1	-14
Bordeaux	0,85	16	6	-10
Brest	0,84	17	25	+8
Lille	0,82	18	14	-4
Limoges	0,81	19	36	+17
Lyon	0,79	20	5	-15
Nantes	0,75	21	7	-14
Tours	0,75	22	21	-1
Rouen	0,72	23	13	-10
Perpignan	0,62	24	32	+8
Pau	0,61	25	26	+1
Nîmes	0,61	26	29	+3
Metz	0,60	27	24	-3
St-Etienne	0,59	28	22	-6
Le Mans	0,59	29	34	+5
Chambéry	0,58	30	30	0
Marseille-Aix	0,57	31	4	-27
Paris	0,56	32	10	-22
St-Denis (Réunion)	0,56	33	40	+7
Orléans	0,50	34	20	-14
Nice	0,49	35	9	-26
Le Havre	0,42	36	38	+2
Valenciennes	0,37	37	41	+ 4
Mulhouse	0,34	38	33	-5
Toulon	0,28	39	31	-8
Avignon	0,26	40	28	-12
Douai-Lens	0,16	41	39	-2

TAB. 7 – *Comparaison du classement par l'attractivité supposée des villes avec le classement de l'Etudiant.*

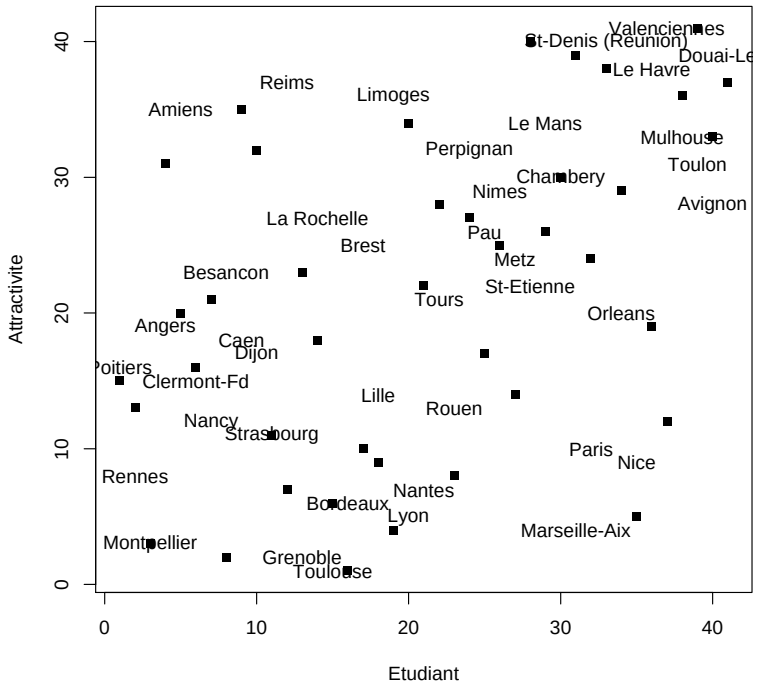


FIG. 3 – *Graphe de la corrélation entre le classement par attractivité et le classement de l'Etudiant*

Eléments d'appréciation et d'analyse d'une base de connaissance : l'expérience industrielle d'Ardans

François VEXLER*, Alain BERGER*, Jean-Pierre COTTON*, Aline BELLONI*

* *Ardans SAS*,
6 rue Jean Pierre Timbaud, "Le Campus" Bâtiment B1,
78180 Montigny-le-Bretonneux, France
{ fvexler, aberger, jpcotton, abelloni } @ ardans.fr
www.ardans.fr

Résumé. Dans l'industrie, on appelle « Recette » l'acte d'acceptation de la recevabilité contractuelle d'une fourniture. La signature d'un contrat qui a pour objectif la livraison d'une base de connaissance est intrinsèquement un paradoxe. Cependant, dans la vie, il faut bien décider et conclure : la base de connaissance est-elle recevable ?

Cet article a pour objectif de montrer les résultats obtenus en mettant en œuvre des principes d'analyse cartographique de bases de connaissance développées pour le compte de différents clients. Le matériel d'étude est donc un matériel réel, opérationnel se composant généralement de bases d'un volume variant entre 250 et 500 éléments de connaissance.

Construites à l'aide de l'outil Ardans Knowledge Maker®, et donc fortement structurées en termes de modèles de contenus et d'arborescences de classification (ou vues), ces analyses répondent tout d'abord à la préoccupation majeure de la qualité des bases fournies et, plus généralement, à la qualité des produits de capitalisation. On s'intéresse ici à la complétude et à la cohérence de ces produits ce qui pose toujours un problème dès que l'on atteint un certain volume.

Ces travaux sont issus de l'industrie et de la douzaine d'années d'opérations d'ingénierie des connaissances réalisée par Ardans® en France et en Europe. Ils préfigurent le prochain module compagnon d'AKM appelé "KB-SCOPE".

1 Introduction

L'ingénierie des connaissances [ArdansSas (2011)] propose aux organisations de s'appuyer sur une méthode voire des outils logiciels afin de modéliser, structurer leurs métiers et les connaissances de leurs meilleurs experts. Depuis sa création, Ardans® conforte sa méthode Ardans MAKE® [ArdansSas (2006b)] au fil de ses opérations industrielles et implante dans sa plate-forme Ardans Knowledge Maker®, (ci-après AKM [ArdansSas (2006a)] et [Mariot et al. (2007)]) ses retours d'expérience et les artéfacts qui satisfont aux exigences des utilisateurs. Mais, le sujet qui préoccupe les responsables des organisations [Berger et Cotton (2011)] est bien celui de la qualité de la base de connaissance qui est mise à disposition de leur

acteurs opérationnels. Comment apprécier un tel dispositif, à sa livraison et au cours de sa vie ? Nous vous proposons l'éclairage pragmatique de notre expérience quotidienne d'industriel du domaine [1].

2 Structuration des bases et outils utilisés

Le méta-modèle de gestion des connaissances implanté dans AKM met en œuvre deux principes simples. Les contenus sont stockés dans des fiches ou éléments de connaissances typés par des modèles définis en fonction du besoin. Les exemples d'analyses présentés dans cet article se réfèrent principalement à une base de connaissance utilisant quatre modèles (Fondamental, Procédé, Résultat et Fiche Technique). Chacune des fiches produites peut être liée à une ou plusieurs fiches via un lien bidirectionnel enregistré en base et permettant la navigation de proche en proche. Ce premier réseau de liens est appelé le réseau « FI-FI ». Par ailleurs, des

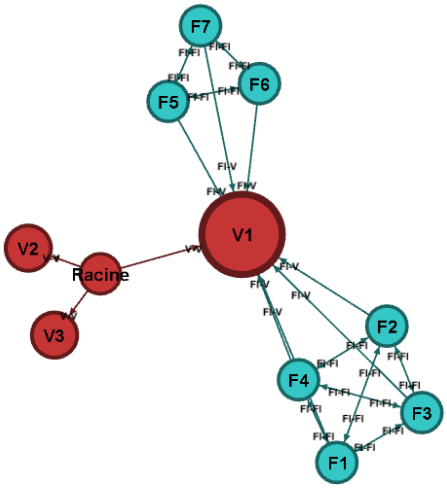


Fig. 1 *Exemple simple de structuration : les réseaux FI-FI et FI-V.*

structures arborescentes sont déclarées et jouent le rôle de support de classification pour les éléments de connaissance que l'on accroche en posant un lien. Ces structures arborescentes, que l'on appelle « Vues » dans le jargon AKM, représentent les concepts métier et se déclinent en autant d'arbres que nécessaire. Un élément de connaissance pouvant être lié à plusieurs items d'arbre via l'accrochage « multi-vues », jouera de facto un rôle de transversalité. Ce second réseau de liens est appelé le réseau « FI-V ». On retrouve ici les principes utilisés dans la gestion des documentations techniques modulaires (Data-Module et Applicabilité) mais avec un

1. L'expérience industrielle avait été présentée à EGC'2009 [Berger et al. (2009)], les problématiques de cahier de laboratoire développées à IC'2011 [Belloni et al. (2011)], celles de l'industrie automobile [Coppens et al. (2006)], de la santé [Chanliau et al. (2006)], ou de l'aéronautique [Verhagen et Curran (2011)] et [Verhagen et al. (2010)], voire de l'industrie des procédés [Louis (2012)].

progiciel de type CMS [2] qui garantit une évolutivité des structures, aussi bien des vues que des modèles. Bien entendu, les structures décrites précédemment sont autant de points d'entrée [3] et d'accès aux contenus. C'est donc sur ces deux réseaux que nous nous appuyons pour mener les analyses. Le logiciel de cartographie choisi est Gephi 0.8 (http ://www.gephi.org) [Bastian et al. (2009)] qui a l'avantage d'être disponible sous licence GPL3 et d'être facilement couplé au SGBD de AKM par des requêtes SQL. Par ailleurs, il dispose des moyens nécessaires à l'exploration des graphes (algorithmes de spatialisation et de partitionnement notamment). Cet aspect d'exploration des graphes est plus important qu'il n'y paraît. En effet, nous avions déjà développé des analyses de base, toujours dans le même objectif de qualité, mais avec des techniques d'analyse de données. Même si certains résultats semblaient intéressant, les aspects fastidieux de la mise en œuvre ne nous permettaient pas de pratiquer ces analyses en séance, au cours des revues de projets. L'aspect visuel immédiat d'un logiciel de cartographie a donc été un critère d'orientation évident. Le schéma 1 représente ces principes de structuration des bases sur quelques fiches et trois points de vue. On y distingue le réseau FI-FI bidirectionnel et le réseau FI-V unidirectionnel et convergeant sur V1.

3 Cartographie de base de connaissance : considérations

Avant de donner ces exemples d'examen, il convient de tracer les limites *a priori* de ce type de méthode. Il est bien évident notamment que nous nous situons dans un domaine de représentation cartographique et on aura conscience à ce titre que « la carte n'est pas le territoire ». La méthode d'élaboration et de constitution de la base de connaissance qui est l'émanation du territoire, est peu impactée et demande toujours l'immersion nécessaire, sinon à la « compréhension [4] », du moins à la structuration. Dans tous les cas, on jugera d'abord la qualité de la base au contenu des éléments de connaissance et à leur état de validation donné par le ou les experts. C'est donc dans un second temps qu'apparaît l'intérêt des analyses cartographiques. D'une part, et nous l'avons déjà dit, au-delà d'un certain volume, il est difficile de se faire une représentation des contenus et d'autre part, ce type d'exercice mené en réunion, va permettre de partager une vision du « patrimoine connaissance » en termes de qualité, de pertinence, de décisions voire de stratégie. Dans tous les cas, la prise de conscience des volumes et de la complexité de ce patrimoine sera bénéfique à l'entreprise. Pour ce qui concerne l'utilisation de Gephi, elle privilégie l'exploitation de l'algorithme de spatialisation « Atlas [5] » et de celui de partitionnement des sous-graphes fortement connectés (*clusterisation* [Blondel et al. (2008)] Les fonctionnalités de partitionnement sur les autres critères de fiche (modèle, auteur, état), les filtres sur les objets (fiches ou liens) ou encore les classements en fonction du nombre de liens entrant ou sortant sont largement utilisés. Les exemples sont des cas réels anonymisés où les libellés des objets ont été remplacés par leur codification interne afin de respecter la confidentialité industrielle.

2. CMS pour Content Management System

3. Il s'agit ici de points d'entrée suivant une logique d'ensemble avec de multiples sélections. Par ailleurs, il est possible de bénéficier des moteurs d'indexation pour des recherches par mots-clés

4. On peut parler d'une « certaine compréhension des sujets » qui semble nécessaire à la constitution d'une base de connaissance. Il ne faut pas attendre, en revanche, que le réalisateur de la base devienne expert.

5. Algorithme orienté force de répulsion entre objets avec force d'attraction due aux liens.

3.1 Exploitation du réseau FI-FI

L'exploitation du réseau FI-FI est le premier examen à mener. C'est l'examen le plus simple en matière de manipulation du logiciel et d'interprétation. Les renseignements fournis sont directement exploitables en matière de qualité. On sera attentif aux points suivants :

Le classement des fiches en fonction du degré de lien et donc des fiches voisines :
cela identifie de manière immédiate les « Hub de connaissance » et peut être un critère de rang pour les résultats de requête sur la base. Il ne s'agit pas uniquement de fiches traitant de généralités ou de fondamentaux et bien souvent leur identification fait prendre conscience de leur large applicabilité ou de leur importance dans le patrimoine.

La répartition en modèles reflète l'équilibre global de la base de connaissance :
l'examen local peut déceler des manques ou au contraire montrer une abondance dans un domaine. Cette représentation remplace avantageusement les simples ratios donnés par des histogrammes.

La connexité du graphe :
elle dépend certes du sujet traité, mais on peut dire de manière générale que dans le cas où un seul expert est engagé, la connexité doit être totale. Ce n'est pas toujours le cas lorsqu'il s'agit d'une base à plusieurs experts où il peut arriver d'observer des ilots de connaissance isolés ce qui ne manque pas de susciter des interrogations. Le plus intéressant est d'observer l'interconnexion entre les éléments de connaissance des différents experts. Cela donne indirectement les interfaces de travail.

La recherche des sous-graphes fortement connectés (*clusterisation*) :
elle se justifie dans notre cas par la tendance naturelle à poser des liens sur les sujets proches. De manière pratique, au cours d'un recueil, nous laissons le soin à l'expert de choisir les liens de voisinage. La colorisation de ces sous-graphes permet l'identification par examen des différents sujets abordés ainsi que leur articulation. A l'inverse, lorsque l'identification ne donne pas un résultat cohérent, elle contribue à entamer une réflexion plus approfondie.

Les trois exemples qui suivent illustrent le propos.

« Base 1-Expert » :
il rend compte d'une base constituée de 234 éléments répartis dans quatre modèles ; résultats de plusieurs séances de recueil auprès d'**un** expert.

« Base 2-Experts » :
il est la continuité du précédent résultant d'un enrichissement des connaissances (39 éléments) d'un **second** expert ayant eu, dans le passé, à utiliser les connaissances produites par le premier.

« Base N-Experts » :
il rend compte d'une base constituée de 350 éléments répartis dans sept modèles et résultant d'une phase de recueil auprès d'une **vingtaine** d'experts.

3.1.1 La Base 1-Expert

La base 1-Expert montre le réseau FI-FI constitué avec les fiches de différentes tailles. Les deux fiches « Hub » en haut à gauche ne sont pas des « généralités » mais au contraire les

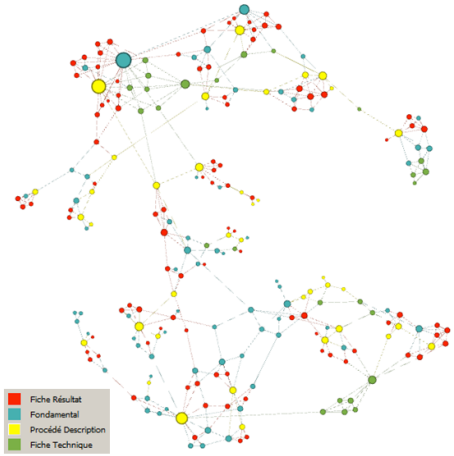


Fig. 2 *Colorisation modèles.*

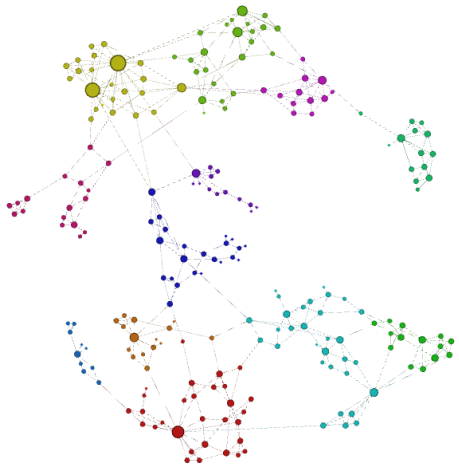


Fig. 3 *Colorisation sous-graphes.*

TAB. 1 – *Base 1-Expert*

deux facettes très rapprochées d'une même expérimentation (modèle de procédé d'un côté et de Fondamental-Enseignement de l'autre). Il est intéressant de constater que la répartition spatiale rendue par l'algorithme « Atlas » regroupe bien les éléments liés et introduit la notion de distance (ou d'adhérence) entre ces objets. La connexité du graphe est complète. Elle a été voulue et a demandé dans certain cas des explications très serrées pour comprendre certaines discontinuités. Le graphe de la figure 2 a été colorisé sur la partition des modèles et le suivant sur la recherche des sous-graphes. Le graphe de la figure 3 montre les 12 sujets abordés dans la base de connaissance. Les points de transition d'un sujet à un autre peuvent être de différents modèles mais supportent la plupart du temps une explication soit technique (utilisation d'un matériel commun), soit historique (intégration d'une sous-expérimentation dans une expérimentation plus large) soit stratégique (évolution des sujets). On notera tout l'intérêt de cette technique au regard de la problématique de complétude d'une base de connaissance.

3.1.2 La Base 2-Experts

La base 2-Experts est la stricte continuité de la précédente, élaborée à quelques mois d'intervalle. Cette continuité a consisté à introduire des éléments de connaissance provenant d'une phase de recueil auprès d'un expert qui avait autrefois, dans le cadre d'un projet, utilisé les connaissances produites par le premier. Il s'agit donc de deux profils très différents d'experts qui, bien qu'ingénieur tous les deux, ont eu des missions différentes. Le graphe de la figure 4 montre la cohabitation des connaissances des deux experts et leurs points d'ancrage.

Bien entendu, cet aspect peut être contesté par le simple fait que le premier expert n'ait pas été présent lors de la pose de ces liens et il n'y a pas de doute que le second expert ait eu une connaissance plus large des connaissances produites par le premier. Mais le focus, au

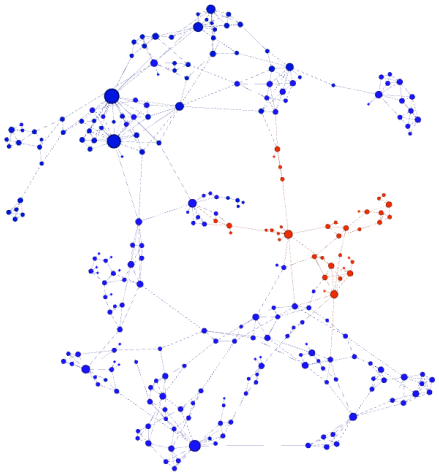


Fig. 4 *Cohabitation des deux expertises.*

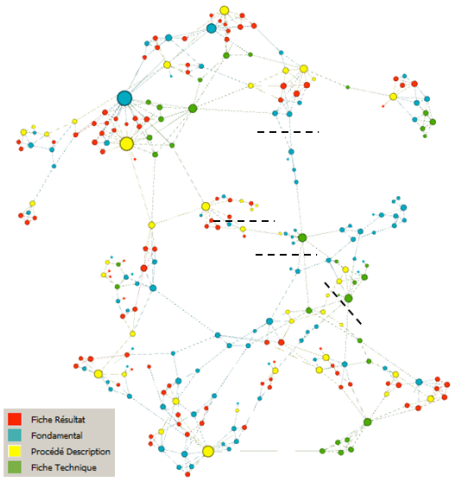


Fig. 5 *Colorisation modèles.*

TAB. 2 – *Base 2-Experts*

cours de cette seconde phase de recueil était centré sur les connaissances du second expert et l'utilisation qu'il avait faite des connaissances du premier. Sur ce sujet, et en reprenant le même graphe que précédemment mais colorisé par les modèles (cf. figure 5), on note que :

Le second expert a généré un ratio important de fondamentaux :
il s'est appuyé pour son projet et le sujet en question sur les connaissances du premier expert. Les autres modèles de connaissance générés sont relatifs à la mission dont il avait la charge.

Les points de connexions (traits en pointillés)
ils sont d'ailleurs pour beaucoup des liens vers les fondamentaux du premier expert et quand ce n'est pas le cas, il s'agit de liens entre fiches de même modèle (procédé vers procédé, fiche technique vers fiche technique).

Dans l'ensemble, on peut juger là de l'appropriation qu'il y a eu et ce type d'analyse met en exergue les concepts de réutilisation [6] des connaissances. Enfin, si on refait une partition par sous-graphe, on trouve bien un 13ème sujet qui s'individualise (graphe non présenté).

3.1.3 La Base N-Experts

Pour terminer le chapitre de l'exploitation du réseau FI-FI, nous présentons dans la figure 6, une base de connaissance élaborée avec une vingtaine d'experts travaillant dans un même service. Cette base antérieure de trois ans, a été conçue pour initialiser un dispositif de gestion des

6. Le cas présenté correspond à une analyse *a posteriori*. On imagine fort bien une démarche pro-active où, partant d'une base de connaissance existante, on se poserait la question des points de connexion et des éléments à acquérir.

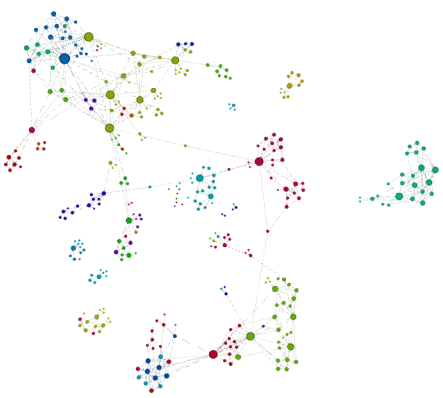


Fig. 6 *Colorisation selon les experts.*

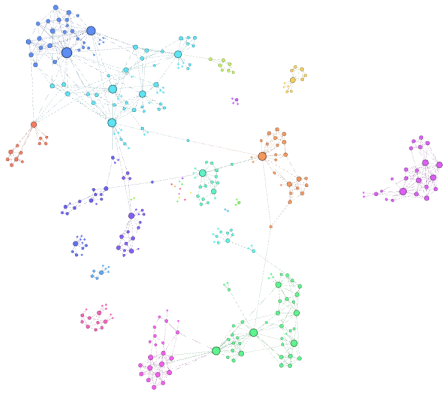


Fig. 7 *Colorisation sous-graphes.*

TAB. 3 – *Base N-Experts*

connaissances. Nous utilisions alors des représentations de type « indicateurs numériques [7] » que nous utilisons toujours d'ailleurs. Avec cette représentation nouvelle « cartographique » les discontinuités du graphe apparaissent. Après une analyse fine, nous constatons qu'elles sont compatibles avec les différentes spécialités d'expert que nous avons rencontrées. Il est toutefois des cas où cette discontinuité n'existe pas et là encore, l'analyse montre un aspect collaboratif entre ces connaissances. Lorsqu'on recherche les sous-graphes, on retrouve bien les différentes spécialités (cf. figure 7).

3.2 Exploitation du réseau FI-V

L'exploitation du réseau FI-V est le second examen à mener. Il est plus compliqué en termes de manipulation du logiciel (utilisation des filtres) et l'examen en séance est délicat à pratiquer à moins d'avoir préparé très exactement les configurations à montrer. Dans cet examen, on s'intéresse avant tout à la qualification des arborescences d'applicabilité, ou Vues d'accrochage dans le jargon AKM et, cela va de pair, à la qualité de l'accrochage. Ces arborescences, qui représentent les « concepts métier » sont l'émanation de l'expert ou des experts et sont construites au fur et à mesure du projet jusqu'à stabilisation. L'opération d'accrochage se fait alors, fiche par fiche [8], en examinant les items et en cochant ceux qui sont concernés par le contenu. Nous avons déjà présenté sur la figure 1, un schéma d'exemple mettant en évidence les liens FI-V créés. Il s'agit en fait d'une indexation manuelle sur « mots-clés hiérarchisés » ou « taxonomie ». Cette manière de faire est motivée par une volonté de maîtrise des connaissances et n'empêche pas l'utilisation par ailleurs de moteurs d'indexation.

7. Ces « indicateurs numériques » illustrent la dynamique de production par modèle et répartition des volumes par modèle pour l'essentiel. Ils ont été depuis industrialisés dans le module AKM Analytics.

8. AKM propose une fonctionnalité d'accrochage de masse dans le cas où il serait nécessaire de pratiquer des réorganisations importantes. Il en est de même pour la réorganisation des arborescences elles-mêmes qui se font sans perte des liens FI-V.

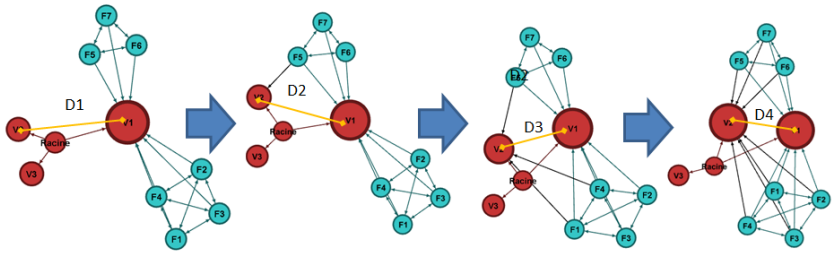


Fig. 8 *Évolution d'une distance entre vues par ajout de liens d'applicabilité.*

Les renseignements fournis sont exploitables en matière de qualité. On sera attentif aux points suivants :

Le classement des items d'arborescence en fonction du degré de lien entrant :
cela identifie de manière immédiate les concepts métiers les plus forts ou qui peuvent être les préoccupations les plus fortes. Dans cet examen, on en profite aussi pour détecter des points de vue non utilisés. C'est un phénomène assez fréquent et qui peut avoir plusieurs causes (niveau de détail trop fin ou décomposition mal adaptée). Le plus souvent, on découvre que le nœud d'arbre directement supérieur (nœud père) est au contraire bien rempli ce qui renseigne sur le niveau de décomposition utile. Certains cas particuliers sont à souligner [9] et qui ont demandé l'utilisation d'un filtre « modèle » pour distinguer les accrochages. Moins fréquent, mais qui peut arriver, la fiche isolée non accrochée. Elle apparaît alors de manière immédiate et le remède est rapide.

La répartition spatiale de l'ensemble des objets en jouant sur la totalité des liens :
cet examen sur les liens (FI-FI + FI-V) montre de manière très simple les distances des concepts métier entre eux et contribue de manière significative à la qualité de la définition de l'arborescence. Nous avions déjà eu cette préoccupation lors de l'élaboration de la « base N-Experts » étant donné sa complexité. A l'époque, nous avions tenté d'utiliser des techniques d'analyse de données qui avaient fourni des résultats cohérents mais fastidieux à mettre en œuvre. Il s'agissait de partir d'une grande matrice de contingence « Vues-Vues », que l'on devait choisir pour aboutir à un *clustering* sous forme d'arbre dendritique. On jugeait après des distances entre concepts à partir de cet arbre. L'utilisation en séance d'un logiciel de cartographie est d'un apport évident. Les schémas de la figure 8 montrent un exemple de résultats à partir de la figure 1 pour éclairer la mécanique utilisée. Pour cela, nous sommes partis de cet exemple et nous avons rajouté des liens d'applicabilité (FI-V) sur la vue V2. On visualise nettement le déplacement de celle-ci, son rapprochement vers V1 et son éloignement de V3. Bien entendu, lors d'un examen de base de connaissance, et par l'intermédiaire des filtres, et si cela est nécessaire, on regardera ces graphiques selon différentes approches (distance des concepts selon un expert ou un autre, selon un modèle ou un autre).

Le troisième point est une réflexion sur la corrélation des concepts des items de Vue :
c'est une extension du point précédent mais qui joue non pas sur la distance perçue mais

9. Typiquement des accrochages de fondamentaux à un niveau « item père » et des accrochages « Fiches technique ou résultats » à un niveau « item fils ».

sur l'appartenance de concepts à des sous-graphes fortement connexes. Il s'agit donc d'une visualisation de couleur d'appartenance. De manière générale, nous avons jusqu'à maintenant utilisé cette technique pour détecter des incohérences, c'est à dire des résultats qui irait à l'encontre de la compréhension. Il est pourtant des cas où le renseignement obtenu est plus pertinent encore et permet d'apprendre de la base de connaissance. C'est en quelque sorte, une véritable « métaconnaissance » ou « connaissance sur la connaissance » telle que qualifiée par Jacques Pitrat.

Nous reprenons les mêmes bases pour illustrer le propos.

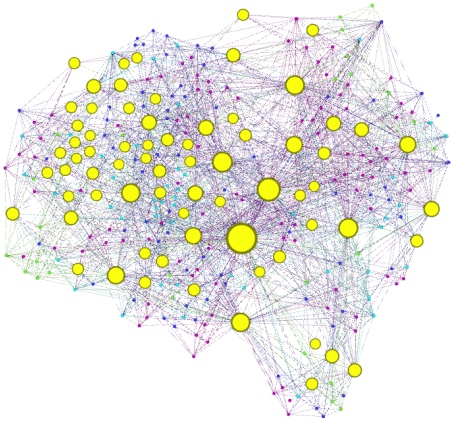


Fig. 9 *Classement des Vues et répartition avec tous les modèles.*

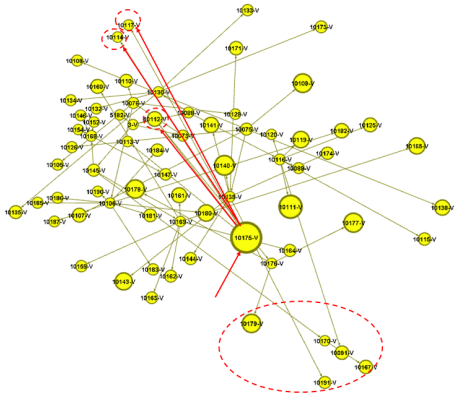


Fig. 10 *Graphe d'analyse des distances.*

TAB. 4 – *Base 1-Expert*

3.2.1 La Base 1-Expert

La base 1-Expert nous montre déjà une répartition homogène des Items de Vue dans les éléments de connaissance. On peut dire que l'expert est cohérent avec lui-même. Après avoir enlevé les objets Fiche (cf. figure 9), les vues apparaissent plus clairement et l'analyse des distances entre ces items confirme la consistance et l'absence d'incohérence. Le graphe de la figure 10 utilisé pour l'analyse des distances est expurgé des objets fiches. Il fait apparaître de plus le réseau V-V dont nous n'avons jamais parlé et qui n'est là que pour reconstituer l'arborescence. Dans l'examen suivant, nous réorganiserons ce réseau pour une meilleure lisibilité. Nous livrons ici quelques éléments d'analyse. On remarquera la vue centrale « 10175-V » de taille prédominante. Cette taille est justifiée compte-tenu de la mission qui était celle d'alors. Pourtant, il s'agit d'un nœud père et les nœuds fils n'ont été que peu remplis. On voit d'ailleurs ces nœuds fils pointés par une flèche rouge (10117-V, 10114-V et 10112-V). Le fils 10112-V est plus proche que les autres ce qui s'avère exact. Le cercle en pointillé entoure 5 vues qui semblent proches l'une de l'autre. Cette proximité s'explique particulièrement bien dans le

contexte de la mission de l'expert. Cet aspect est d'ailleurs flagrant dans l'examen suivant. Le reste de l'examen est intéressant et confirme la cohérence de l'ensemble.

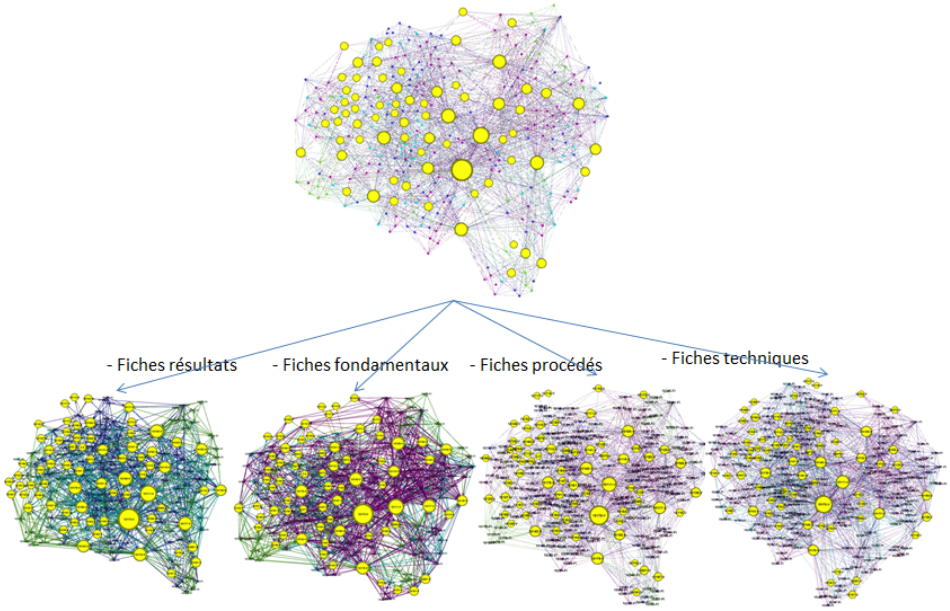


Fig. 11 *Base 1-Expert : disposition des vues en filtrant sur les modèles.*

Le schéma de la figure 11 reprend ce même principe en opérant une sélection sur les modèles afin de voir s'il existe une prédominance. De manière pratique cela consiste à filtrer sur les différents modèles tout en laissant tourner l'algorithme de spatialisation. C'est donc à l'œil que l'on juge des transformations au moment où elles s'opèrent. On voit alors les déplacements des Vues les unes par rapport aux autres. On constate que, sur ces bases déjà d'un certain volume, les déplacements eux-mêmes peuvent être catégorisés : existence de points fixes et phénomènes de rotation entre points fixes. Le graphique 11 montre que si on enlève un modèle de fiche, la disposition générale des vues ne change que peu même si localement il y a des déplacements des nœuds légers. A l'œil, on peut dire que le retrait qui conserve le mieux l'aspect général est celui des « Fiches résultat » ce qui n'était pas évident du fait que ces fiches représentent 31% du volume total. Cela signifie, en d'autres termes, que les « Fiches résultats » sont en concordance avec la configuration spatiale imprimée par les autres modèles, notamment les fondamentaux et les procédés compte-tenu du fait que les « Fiches techniques » ne représentent que 8,5% du total. Ceci est plutôt rassurant pour ce qui concerne les résultats. On constate aussi cette même conservation pour le retrait des Fiches procédés, ce qui tend à montrer une bonne constitution entre les résultats et les Fondamentaux-Enseignements. Nous terminerons l'analyse de la base 1-Expert par l'examen de corrélation des items de Vues (cf. figure 12). Comme nous l'avons fait avec l'examen précédent, nous pourrions choisir les modèles qui nous intéressent le plus, ce que l'on ne manque pas de faire quand on pratique une analyse complète.

Nous nous contenterons dans cet article de pratiquer l'examen avec l'ensemble complet des fiches. Le premier résultat est donc l'intégration des Vues dans les sous-graphes fortement connexes. On y retrouve ce que l'on soupçonnait déjà par la proximité (figure 10) des Vues « 10091-V », « 10191-V », « 10167-V » et « 10179-V ». En filtrant sur les modèles et en réorganisant le graphe, on obtient le résultat présenté dans la figure 13. La colorisation notée en pointillés intègre la vue « 10140-V » qui n'était pas à proximité mais qui est toutefois pertinente dans l'information détectée.

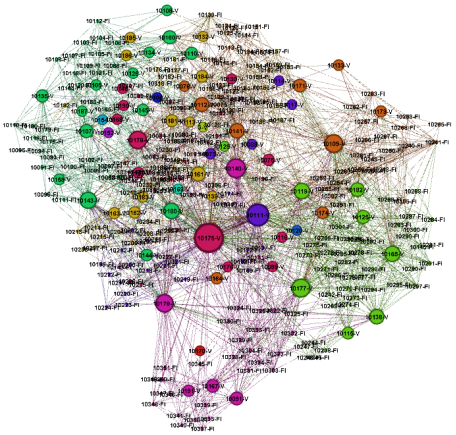


Fig. 12 *Clustering intégrant les vues.*

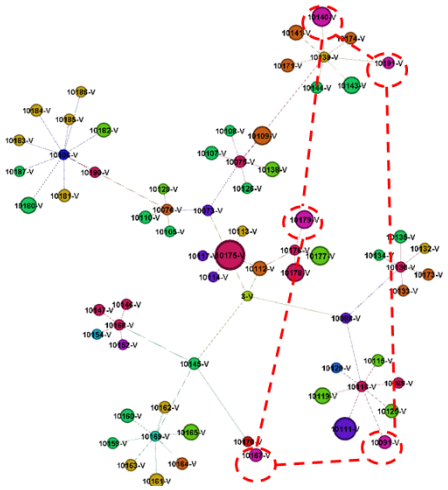


Fig. 13 *Corrélation des vues.*

TAB. 5 – *Base 1-Expert : les Analyses*

3.2.2 La Base 2-Experts

Dans la base 2-Expert, nous nous contenterons de regarder les distances des vues (figure 14) en ne retenant que les fiches produites par le second expert. Ceci complètera la réflexion précédente sur les points d'accroche des connaissances de l'un sur l'autre. On constate l'absence complète d'utilisation d'un nombre important d'items. L'analyse de corrélation des vues (figure 15) ne détecte pas d'incohérence et confirme les sujets d'intérêt du second expert.

3.2.3 La Base N-Experts

Nous terminerons avec la base N-Experts en montrant seulement une analyse de corrélation (figure 16). Etant donné la complexité de l'arborescence, ceci ne peut être qu'intéressant. En effet, cette analyse de corrélation apporte les éléments suivants :

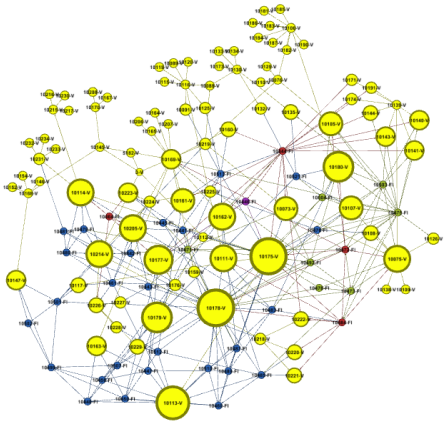


Fig. 14 *Distance de fiche du second expert.*

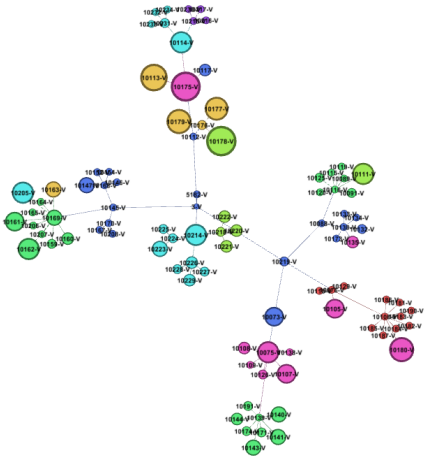


Fig. 15 *Corrélation des vues.*

TAB. 6 – *Base 2-Experts : les Analyses*

- D'une part, le classement des vues correspond bien à la réalité avec deux gros sujets porteurs d'un gros volume d'éléments de connaissance chacun et avec une corrélation évidente entre la vue « 1515-V » et la vue « 1613-V ».
- D'autre part, des corrélations très justes, comme entre la vue « 1512-V » et les autres vues de même couleurs, notamment celles qui ont un certain volume.

Globalement, il n'y a pas d'incohérence dans l'accrochage de cette base qui, rappelons-le, a été constituée avec plus de vingt experts.

4 Conclusions et perspectives

Cet article s'est efforcé de montrer l'intérêt à pratiquer l'analyse des bases de connaissance structurées dans le but avant tout d'en contrôler la qualité, mais aussi de disposer d'information pour contribuer à une gestion pro-active et anticipatrice. Pour ce qui concerne la qualité, ceci est d'autant plus important que nous sommes fréquemment amenés à développer ce type de livrables pour des entreprises dont les connaissances sont très largement d'ordre technique et/ou scientifique. Pour cela, et en rapport avec notre logiciel de production AKM, nous avons exploré, à partir de bases réelles que nous avons constituées et dont nous connaissons le contenu, l'intérêt d'utiliser des méthodes cartographiques telles qu'on les utilise le plus souvent dans l'exploration des réseaux sociaux. La différence tient pour beaucoup dans le mode de structuration qui différencie nettement deux catégories de liens et deux catégories d'objets. Bien entendu, le fait d'avoir nous-même développé ces bases nous offre par leur maîtrise induite, un œil critique sur nos propres conclusions. Les résultats obtenus sont très encourageants. Ils demandent un aspect méthodologique assez rigoureux car il s'agit d'un usage par examen

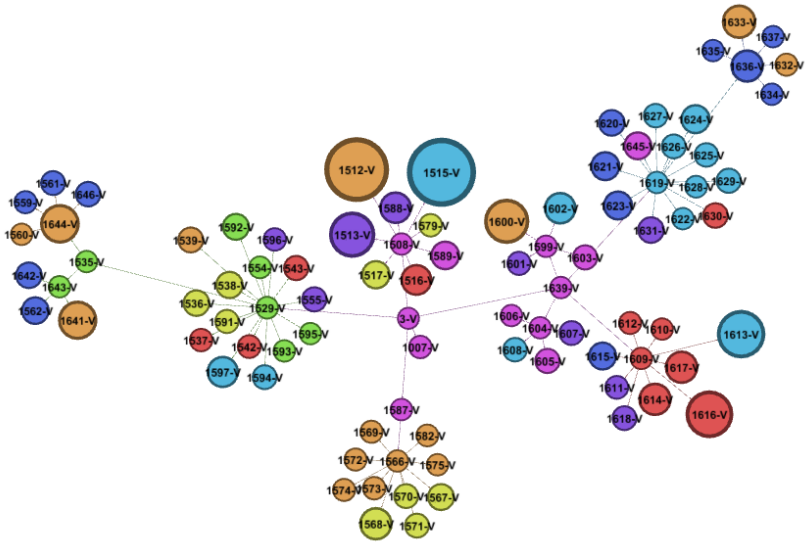


Fig. 16 *Base N-Experts : Analyse de corrélation des Vues.*

dont il faudra dégager un protocole d'examen. Sur l'aspect technique, nous avons directement connecté aux bases de connaissance le logiciel Gephi 0.8 en attendant un prochain export au format graphml. Concernant l'avenir, Ardans a obtenu le soutien d'OSEO pour aider à la maturation de la technique d'auscultation des bases de connaissances spécifié dans cet article vers un environnement que nous avons appelé "KB-SCOPE".

Références

ArdansSas (2006a). Ardans Knowledge Maker : introduction, principes et philosophie implantés dans cet environnement de gestion des connaissances. Article de synthèse technique, www.ardans.fr/doc/AST2006-354%20NKM%20AKM%20v1_3.pdf.

ArdansSas (2006b). Ardans MAKE : méthode d'élaboration de la mémoire en continu collective. Article de synthèse technique, www.ardans.fr/doc/AST2006-152%20NKM%20Make%20v1_2.pdf.

ArdansSas (2011). Mémoire en continu : capitaliser et valoriser le savoir-faire et les expériences en continu. Article de synthèse technique, www.ardans.fr/doc/AST2011-132%20NKM%20MC%20v2_0.pdf.

Bastian, M., S. Heymann, et M. Jacomy (2009). Gephi : An open source software for exploring and manipulating networks (www.aaai.org/ocs/index.php/icwsm/09/paper/view/154).

Belloni, A., A. Berger, J. Cotton, et F. Devoret (2011). De la gestion des connaissances structurées au cahier de laboratoire électronique à valeur probatoire : naissance de CLEOPATRE (www.ardans.fr/doc/IC2011_CLEOPATREpublibook.pdf).

Berger, A. et J. Cotton (2011). Construire une mémoire collective de l'entreprise : la gestion des connaissances (www.ardans.fr/doc/AFIA_Communication.pdf). *AFIA n° 72*, 70–73.

Berger, A., J. Cotton, et P. Mariot (2009). Accompagner au début du 21ème siècle les organisations dans la mise en place d'une gestion des connaissances : retour d'expérience (www.ardans.fr/doc/090128-EGC2009-AB-JPC-PM.pdf). *RNTI EGC 2009 E15*, 475–479.

Blondel, V., J. Guillaume, R. Lambiotte, et E. Lefebvre (2008). Fast unfolding of communities in large networks (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.139.5771). *Journal of Statistical Mechanics: Theory and Experiment P10008*, 1742–5468.

Chanliau, J., A. Mariot, A. Caillette-Baudoin, P. Mariot, A. Berger, et F. Vexler (2006). L'efficience des organisations de santé : La bascule dans la gestion des connaissances (www.irit.fr/SDC2006/cdrom/contributions/Jacqueschaliau_SDC2006.pdf).

Coppens, C., J. Laroque-Malbert, A. Berger, P. Mariot, et J. Cotton (2006). La capitalisation des connaissances dans l'industrie : Implanter un référentiel métier et le déployer : l'exemple d'Icare (www.irit.fr/SDC2006/cdrom/contributions/Alainberger_SDC2006.pdf).

Louis, A. (2012). ALPIN, une application collaborative de partage de connaissances, autour des documents techniques clés de l'ingénierie d'AIR LIQUIDE Engineering (www.imdr.fr/docs/LM18_Actes%20congres/textes/lm18_com_3C-5_079_A_Louis.pdf).

Mariot, P., C. Golbreich, , J. Cotton, et A. Berger (2007). Méthode, Modèle et Outil Ardans de capitalisation des connaissances (www.med.univ-rennes1.fr/lim/doc_166.pdf). pp. 187–206.

Verhagen, W., P. Bermell Garcia, P. Mariot, J. Cotton, D. Ruiz, R. Redon, et R. Curran (2010). Knowledge-based cost modelling of composite wing top cover (http://peer.ccsd.cnrs.fr/peer-00649053). *Int.Journal of Computer Integrated Manufacturing IJCIM-0124*, TCIM–2010.

Verhagen, W. et R. Curran (2011). Ontological modelling of the aerospace composite manufacturing domain (www.icas.org/icas_archive_cd1998-2010/icas2010/papers/483.pdf). In *Improving Complex Systems Today*, Advanced Concurrent Eng., pp. 215–222. Springer.

Summary

In industry, the signing of a contract which aims to deliver a knowledge base is inherently paradoxical. However, in life, we must decide and conclude: the knowledge base is it admissible or not? This paper aims to show the results obtained by implementing the principles of cartographic analysis of knowledge bases developed on behalf of different clients. The study material is real hardware, made of operational bases, typically consisting of a volume between 250 and 500 articles.

Built using the tool Ardans Knowledge Maker®, and therefore highly structured in terms of content models and classification trees (or views), these analyzes address the quality of bases provided and, more generally, the quality of capitalization products. We therefore focus here on the completeness and consistency of these products which is always a problem when one reaches a certain volume.

These works are lessons learnt from industry and of a dozen years of knowledge engineering operations performed by Ardans in France and Europe. They announce the next AKM module called "KB-SCOPE".

Index

Partenaires :