Analyses multidimensionnelles de contenus documentaires dans un ENT universitaire au service de l'acteur enseignant-chercheur

Frédérique Peguiron, Odile Thiery

Analyses multidimensionnelles de contenus documentaires dans un ENT universitaire au service de l'acteur enseignant-chercheur

Frédérique Peguiron (1) Frederique.Peguiron@loria.fr

Odile Thiéry (2) Odile.Thiery@loria.fr

LORIA, Campus Scientifique,B.P. 239, 54506 Vandoeuvre-les-Nancy Cédex, France (1)

LORIA, Campus Scientifique,B.P. 239, 54506 Vandoeuvre-les-Nancy Cédex, France (2)

Mots-clés : Bibliométrie multidimensionnelle, analyse multidimensionnelle, entrepôt de données, OLAP.

Keywords: Multidimensional Bibliometry, analyze multidimensional, Datawarehouse, OLAP.

Résumé : Les universités s'organisent en consortium pour rayonner en Université Numérique de Région ou Université Numérique Thématique et proposent des Espaces Numériques de Travail (ENT) à leurs utilisateurs. A l'instar des pôles de compétitivité en Entreprise naissent les pôles de recherche et d'enseignement supérieur en Université qui font office de levier dans le développement des universités à l'échelon du territoire voire à un échelon mondial. Apparaît la nécessité de penser à des systèmes d'analyse ou d'évaluation au service de la recherche. Ces ENT offrent d'une part de nouvelles perspectives d'analyses de documents à des fins bibliométriques ou scientométriques et d'autre part permettent la mise en œuvre d'analyses par la mise en place d'indicateurs au service d'une nouvelle gouvernance des établissements. Après avoir montré la faisabilité de mener une analyse multidimensionnelle sur du contenu documentaire au travers d'une expérimentation, nous développons une application en open source intégrable à une plate forme numérique d'espace collaboratif. Nous focalisons ici autour de l'acteur enseignant-chercheur. Nos développements témoignent que cette application personnalisable en fonction des acteurs peut être utile à l'évaluation de la recherche. La phase expérimentale analyse de façon multidimensionnelle le contenu informatif de groupes de discussion et de listes de diffusion à l'aide d'un entrepôt de données. Ce travail aboutit à des données que nous récupérons pour les intégrer à une plate forme qui permet des analyses OLAP via le web. Les vues proposées sont en fonction des types d'acteurs. L'orientation de la diffusion des produits de la recherche et la nécessaire visibilité des enseignants-chercheurs via des OAI échappent à l'Impact Factor mené par les éditeurs scientifiques.

Abstract : The universities are organized in consortium to radiate in Numerical University of Area or Numerical University Set of themes. They propose Numerical Spaces of Work to their users. The poles of research and higher education in University are levers in the development of the universities at the level of the territory and a world level. It is necessary to think of systems of analysis or evaluation for research. These numerical spaces offer new prospects for abstracts to bibliometric or scientometric ends. They allow the implementation of analyses by the installation of indicators the service of the establishments. We show feasibility to carry out a multidimensional analysis on documentary contents through an experimentation, then we develop an application in open integrable source at a numerical platform of collaboratif space. We concentrate on the actor teacher-researcher. Our developments testify that this personnalisable application according to the actors can be useful for the evaluation of research. The experimental phase analyzes in a multidimensional way the informative contents of newsgroups and mailing lists using a datawarehouse. This work leads to data which we recover to integrate them into a punt forms which allows analyses OLAP through Web. The sights suggested are according to the types of actors. The orientation of the diffusion of the products of research and the necessary visibility of the teacher-researchers through OAI escape from Impact Factor carried out by the scientific editors.

Introduction

Actuellement nous assistons à des développements rapides des Environnements Numériques de Travail dans plusieurs universités qui ont fait émerger des difficultés relatives à la conduite du changement et une implication des différents acteurs de l'université. Les Systèmes d'Information sont au cœur des organisations. Les universités se sont organisées autour de consortium pour rayonner en Université Numérique de Région ou Université Numérique Thématique pour proposer des Espaces Numériques de Travail (ENT) à leurs utilisateurs. Ces ENT ne sont pas seulement une juxtaposition d'outils, mais proposent des services pour permettre une entrée pédagogique de leurs acteurs. A l'instar des pôles de compétitivité en Entreprise naissent les pôles de recherche et d'enseignement supérieur en Université qui font office de levier dans le développement des universités à l'échelon du territoire voire à un échelon mondial. Ces ENT offrent d'une part de nouvelles perspectives d'analyses de documents à des fins bibliométriques ou scientométriques et d'autre part permet la mise en œuvre d'analyses par la mise en place d'indicateurs au service d'une nouvelle gouvernance des établissements. Les services mis à la disposition des utilisateurs représentent la partie émergée du système d'information. Nous verrons comment en nous intéressant aux données de la partie immergée du système d'information celui-ci peut atteindre la dimension d'un système d'information stratégique au bénéfice d'une facilitation de l'évaluation de l'organisation par ses acteurs.

Dans cet article nous nous intéressons particulièrement à un service qui peut enrichir un ENT. Ce papier propose deux parties : une phase expérimentale montre la faisabilité de mener une analyse multidimensionnelle de contenu documentaire à partir d'un entrepôt de données à l'aide d'un outil décisionnel, puis la seconde phase reprend les résultats de l'expérimentation pour proposer un service d'analyse multidimensionnelle intégrable à un ENT sous forme d'une application à l'aide d'un produit en open source qui offre une valeur ajoutée par la personnalisation des vues proposées aux différents acteurs du système d'information selon leurs besoins.

Les enjeux

Le déploiement des pôles de recherche et d'enseignement supérieur en Université et l'intégration des plates formes de dépôts d'archives ouvertes dans les systèmes d'information offrent de nouvelles perspectives d'analyses à prendre en compte pour mettre en œuvre un système d'observatoire au service d'une nouvelle gouvernance des universités. Le cadre de cet article s'inscrit dans les objectifs du S.3I.T. 2008 - Schéma Stratégique des Systèmes d'Information et des Télécommunications qui à horizon 2008 définit la stratégie pour le numérique dans l'éducation nationale, l'enseignement supérieur et la recherche[1].

L'existant

Le Système d'Information de l'Université est complexe et hétérogène. Résumons de façon non exhaustive les services classiquement proposés dans un ENT qui constituent une entrée pédagogique pour les acteurs et représentent la partie émergée du Système d'Information :

- Courriel, forum, agenda, plan de travail,
- Podcasting[2] : fichiers MP3, Mpeg,
- Portefeuille de compétences ou Portfolio : porte document partageable, flux RSS,
- Banques d'images et d'animations,
- Content Management System (CMS) ou système de gestion de contenu,
- Volet pédagogique : cours, exerciseurs,
- Volet documentaire : catalogues, bases de données, encyclopédies.

Le système d'information est constitué d'une juxtaposition d'applications. «L'éclatement des technologies se traduit par une multiplication des degrés de liberté pour créer des applications» [23]. Ce phénomène accroît les difficultés pour les systèmes d'information qui sont pensés en termes de processus transversaux.

1. Analyse bibliométrique multidimensionnelle : phase expérimentale

Nous proposons par exemple à l'acteur «enseignant-chercheur» de compléter un état de l'art par l'analyse de listes de diffusion à partir d'un entrepôt de données. D'après Franco [12], l'architecture de l'entrepôt de données comporte trois niveaux fonctionnels essentiels : le niveau acquisition des données, le niveau stockage des

[1] http://www.education.gouv.fr/cid4180/le-2008-schema-strategique-des-systemes-information-des-telecommunications-horizon-2008.html#qu'est-ce-que-le-s3it-2008-la-reference-pour-les-tic-e-et-les-s-i-

[2] Podcasting (un terme composé autour des mots iPod , webcasting , et broadcasting) est une technique qui permet de transférer et d'écouter automatiquement sur son baladeur MP3 les programmes audio d'un site, sans avoir à le visiter.

données et le niveau analyse de données. L'entrepôt de données doit intégrer les données les unes avec les autres afin d'assurer une cohérence sémantique globale. Il se compose d'un data warehouse, de bases de données multidimensionnelles ou hypercubes et d'un ensemble d'outils permettant l'alimentation du data warehouse, son interrogation et la production de rapports, l'extraction intelligente des données par techniques de data mining enfin l'analyse décisionnelle. Nous restituons ici une méthode adoptée pour faire ressortir les tendances émergentes propre au thème de l'intelligence économique qui intéresse l'acteur enseignant-chercheur entre 2001 et 2005.

Nous proposons une méthode pour prendre connaissance d'un vocabulaire existant autour d'une thématique pour en mesurer son évolution et analyser son contexte. Le thème retenu «l'intelligence économique» est un processus qui couvre plusieurs champs disciplinaires de façon transversale. Il s'agit de cerner les différents concepts propres à cette thématique. Cette façon de procéder permet de repérer les tendances émergentes, les acteurs, les réseaux, les parutions d'ouvrages et les conférences ou colloques en rapport avec ce thème. Pour cela nous nous sommes abonnées à des listes de diffusion et à des groupes de discussion autour :

- des outils de recherche d'information,
- des moteurs de recherche,
- de l'intelligence économique,
- de la gestion des connaissances,
- des outils de veille,
- des outils spécifiques à la documentation,
- des outils spécifiques aux bibliothèques.

Voici le nom des groupes et des listes, ainsi que les thèmes abordés qui ont servi de support de réflexion :

Liste de diffusion	Thèmes
adbs-info@cru.fr	L'association des professionnels de l'information et de la documentation en 1994, a pour objectif de faciliter les échanges d'informations, d'idées et d'expériences.
adest@grenet.fr	Bibliométrie, scientométrie, infométrie, recherche théorique et appliquée. Cette liste de diffusion permet aux professionnels de la bibliométrie d'engager des discussions sur différentes questions, de s'informer sur les nouveaux outils, méthodes, traitements des données, essais.
agents@yahoogroupes.fr	Consacrée aux agents intelligents. Imbriqué au site AgentLand.fr, cet espace permet d'échanger des solutions pour mieux maîtriser les agents, poser des questions, suggérer des améliorations, donner des avis sur un agent.
biblio-fr@cru.fr	Cette liste de diffusion regroupe bibliothécaires et documentalistes francophones, et toute personne intéressée par la diffusion électronique de l'information documentaire.
cybercrise@yahoogroupes.fr	Ce groupe de discussion est destiné à échanger et à faire évoluer la réflexion sur la gestion de crise.
gredoc@grenet.fr	Liste dédiée à la mesure des sciences et techniques
ienetwork@yahoogroupes.fr	Ce groupe a pour objectif de regrouper les nouveaux et anciens étudiants, les professionnels de l'intelligence économique, promouvoir notre profession, échanger des offres d'emploi et des informations sur l'actualité de l'intelligence économique au niveau international.
i-KM@yahoogroupes.fr puis i-KMFORUM@yahoogroupes.fr	Forums et listes de discussion du secteur de l'information-documentation
intelligence-economique@yahoogroupes.fr	Ce groupe de discussion est consacré à l'intelligence économique au sens large, c'est à dire la gestion de l'information externe : mise en place d'un système de veille, outils et méthodes, les aspects de protection de l'information, de renseignement, de benchmarking, d'influence, de knowledge management.
miste-esiee@yahoogroupes.fr	Ce forum dédié à la créativité et à l'utilisation des cartes heuristiques, cartes d'organisation d'idées, topogrammes, arbres à sens, schémas arborescents, cartes mentales et autres «mind maps»
motrech@yahoogroupes.fr	Cette liste, consacrée aux moteurs de recherche sur Internet, est un lieu d'échanges sur les problématiques, techniques, développements et évaluations/comparaisons des outils de recherche d'information sur Internet.
netkm@egroupes.fr puis netkm@yahoogroupes.fr	Club du Knowledge Management et de l'Intelligence Economique.
newsletter@afnet.fr	Liste de diffusion de l'AFNeT (Association Francophone des utilisateurs du Net de l'e-business et de la société en réseau)
prospective@egroups.fr	Prospective sur Internet. De quelle manière Internet peut être un excellent outil pour détecter les nouvelles tendances, constituer un réseau d'experts, identifier les réseaux de collaborations.
veille@egroups.com puis veille@yahoogroupes.fr	Liste consacrée aux thématiques de veille sur Internet. Historiquement, il s'agit de la première mailing-list française sur l'intelligence économique et stratégique sur Internet (1998). Créée initialement pour les lecteurs du livre «Intelligence Stratégique sur Internet : comment développer efficacement des activités de veille et de recherche sur les réseaux» (Dunod).

Figure 1. Listes de diffusion et groupes de discussion étudiés de 2001 à 2005

1.1 Récapitulatif des groupes et listes comme support d'analyses

Ces listes de diffusion et groupes de discussion nous offrent un corpus de 724 messages après une équation de recherche autour du mot «colloque» dont nous ne retenons que les messages ayant trait aux événements concernant notre sujet de recherche (l'intelligence économique) via les événements qui s'y rapportent. Un premier tableau analytique est réalisé après dépouillement des listes de diffusion et groupes de discussion consacrés à l'intelligence économique – il permet la réalisation de tableaux synthétiques autour d'indicateurs en vue d'une fouille de données. Le tableau analytique est construit autour des rubriques suivantes : «**Evénement**», «**Date**», «**Lieu**», «**Site dédié**», «**Organisateur**» et «**Objectifs**». Une simple lecture de ce tableau dense en quantité d'informations ne permet pas de faire une analyse fine et de mettre l'accent sur une évolution des événements. C'est pourquoi une seconde lecture dite «intelligente» permet de proposer un second tableau simplifié où sont rajoutées des rubriques pour chaque événement : Thèmes, Secteur, Type d'organisateur et Spécialité de l'organisateur. Par ces nouvelles rubriques nous mettons en place des indicateurs qui constituent des clés de lecture en vue d'une fouille de données pour mettre l'accent sur des émergences, des évolutions ou encore des tendances. La rubrique «**Themes**» concerne au principal thème abordé lors de l'événement. Le champ «**Secteur**» identifie le secteur global touché ciblé par l'événement. «**Organisateur_type**» type l'organisateur par rapport à sa raison social relative à notre sujet de recherche.

1.2 Préalable à l'analyse multidimensionnelle

Nous obtenons à l'aide d'Excel plusieurs tableaux pour chaque année de 2001 à 2005 dont nous restituons pour exemple en figure 2 le tableau relatif à l'année 2004 :

Evenements	Themes	Secteur	Objectifs	Date	Organisateur_type	Organisateur_specialite
EGC	méthodes	Km	extraction de connaissances à partir de données	2004	institut	information
journées veille	veille	entreprise	recherche d'information	2004	université	information-communication
journées linguistiques	outils-méthodes	km	gestion des contenus	2004	association	information-économie
VSST	méthodes	km	exploitation efficace des grandes masses de documents	2004	université	informatique-bibliométrique
congré	projet	entreprise	compétitivité et innovation	2004	association	information
journées ADBS	exploitation	documentation	weblogs dans la publication et diffusion de l'information : enjeux	2004	association	documentation-information

Figure 2. Tableau synthétique autour d'indicateurs pour l'année 2004.

1.3 Analyse multidimensionnelle en vue d'une fouille de données

A partir de ce type de tableau, les fonctionnalités d'Excel, hormis les fonctions de calcul, de tri et de représentation graphique, ne nous permettent pas de mettre en valeur des tendances : Excel ne montre qu'un vue planaire des données. Pour avoir une vue multidimensionnelle des données il faut se tourner vers des outils décisionnels. Nous utilisons un outil d'analyse et de reporting Cognos et procédons à une fouille de données pour faire émerger visuellement des dates clés en rapport avec les événements autour de l'intelligence économique et mesurer l'apparition de concepts. Cognos regroupe les outils Powerplay[3] et Transformer[4] et permet l'exploration d'une base multidimensionnelle. Les tableaux de 2001 à 2005 réalisés lors de la phase expérimentale sont importés à la suite Cognos. Transformer permet de créer un arbre de dimensions où apparaissent les sources de données, les mesures, les cubes ainsi que la grille des dimensions autour de «**Secteur**», «**Date**» et «**Organisateur_type**» en relation avec les mesures «T**hemes**», «**Organisateur_specialite**», «**Evenements**» et «**Objectifs**» comme en témoigne la figure 3. PowerPlay permet une analyse multidimensionnelle en faisant varier les niveaux d'analyse : le secteur «Intelligence Economique» est le plus ciblé des événements en 2002. D'une façon très simple, on modifie les vues par exemple autour des types d'organisateurs. On s'aperçoit ainsi que les universités, au cours des années, représentent régulièrement le nombre le plus important d'organisateur. Le secteur de la documentation est largement concerné depuis 2003.

[3] Powerplay comporte EXPLORER et REPORTER qui permettent la création de rapports et la mise en évidence de résultats pertinents pour l'aide à la décision

[4] Transformer crée des hypercubes à partir d'une base multidimensionnelle

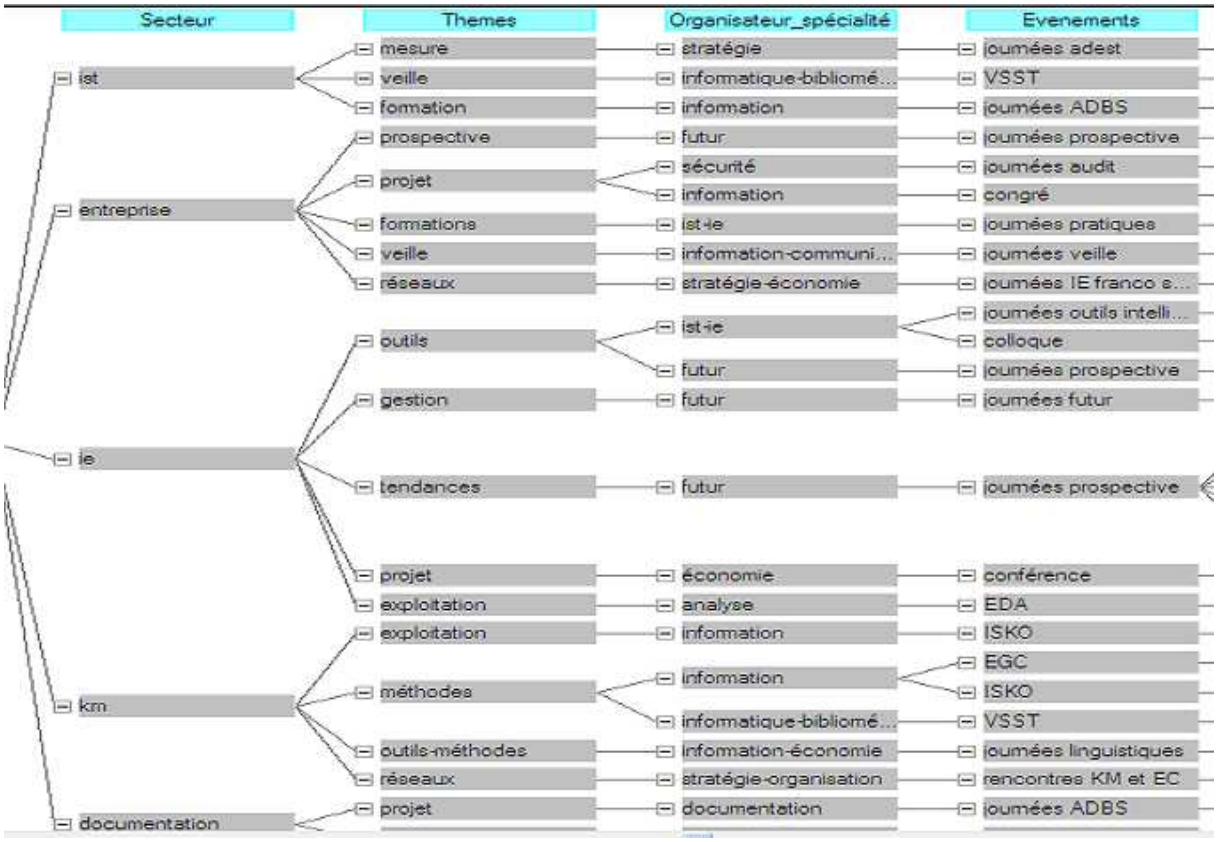


Figure 3. Arbre des dimensions

A partir des résultats de cette première phase expérimentale d'analyse bibliométrique multidimensionnelle, nous avons pu mettre en perspective les thèmes abordés de l'état de l'art regroupés autour de quatre concepts qui sont : les tendances (idées, expériences, gestion de crise), l'information (actualité, national, international, protection), les outils (nouveaux outils, réseaux d'experts, méthodes) et la représentation (traitement des données, cartes heuristiques, cartes d'organisation d'idées) autour de l'intelligence économique – intelligence économique à forte connotation «gestion des connaissances». Deux verbes d'action sous tendent les relations entre concepts et thèmes : collaborer d'abord, puis partager. Cette phase expérimentale au service d'un état de l'art apporte une aide pour les pistes de recherche à l'acteur enseignant-chercheur et permet d'ébaucher un plan de rédaction pour l'acteur thésard. Toutefois ce genre d'outil comporte des inconvénients dans la mesure où il est onéreux, nécessite une installation sur chaque poste de travail et ne permet pas d'analyse via le web.

1.4 Vers un enrichissement des services de l'ENT

Au vue des bénéfices ramenés à utiliser un entrepôt de données pour procéder à des analyses documentaires, nous avons eu l'idée d'enrichir l'ENT en proposant un service qui pourrait venir s'intégrer au système d'information et ainsi enrichir la palette des services proposés pour répondre aux besoins des acteurs enseignants-chercheurs. Rappelons que pour favoriser l'intégration de ces services au niveau du système d'information les technologies employées pour le développement des applications reposent sur des logiciels uPortal[5]. Consciente de cette pluralité de possibilités de développement l'Agence de mutualisation des universités (AMUE) [1] travaille à un rapprochement des consortiums dans le souci de pérenniser les développements, d'en favoriser leur réutilisabilité et leur interopérabilité. L'AMUE met l'accent sur l'intérêt de développer des applications autour d'une architecture SOA[6] en faveur des Web services. Les architectures orientées services permettent d'expliquer la nouvelle complexité de la conception d'applications distribuées. Interopérabilité, standardisation, démarches de conception plus collaboratives et orientées processus, applications composites, solutions de management des processus sont des éléments qui contribuent à modifier les portails d'établissement.

[5] uPortal : Framework open source basé sur Java, XML et XSL servant à créer rapidement des portails dédiés aux campus universitaires. Il est développé sous l'égide de JA-SIG. uPortal n'est pas un logiciel prêt à l'emploi, mais plutôt une bibliothèque de classes Java et de documents XML/XSL permettant de développer le portail.

[6] SOA : Service Oriented Architecture ou Architecture Orientée Services

2. Analyse bibliométrique multidimensionnelle : phase applicative

Les portails d'applications se réinvente progressivement en repartant d'une conception documentaire de l'information. XML en forme le socle omniprésent. L'impact de XML dans les systèmes d'informations décisionnels se révèle considérable. Les bases de données XML stockent des documents de manière transactionnelle, tout en gardant la capacité de les extraire grâce à de multiples graphes, à l'instar des bases relationnelles. Nous appréhendons ce nouveau modèle pour notre application où est abordé un langage de développement autorisant la manipulation de bases de données en vue d'analyses. Nous portons au travers d'une application toutes les données issues de la phase expérimentale pour relever le défi d'utiliser un logiciel en open source : Openi. Openi offre des perspectives innovantes quant au traitement du contenu des informations puisqu'il repose sur des schémas XMLA[7] pour l'analyse des données. Nous appréhendons ce nouveau modèle d'analyse pour notre application où est utilisé un langage de développement autorisant la manipulation de bases de données par requêtes MDX[8] en vue d'analyses. MDX est le langage de requêtes utilisé pour les bases de données multidimensionnelles, de la même manière que SQL est utilisé pour les requêtes sur les bases de données relationnelles. Cet outil décisionnel libre repose sur le moteur OLAP Mondrian[9] destiné à la création et à la publication de rapports. Nous avons élaboré des schémas d'analyses en XMLA dont les requêtes MDX permettent de procéder à des analyses multidimensionnelles via une interface web.

2.1 Présentation de Mondrian

Les fonctionnalités attendues d'un système décisionnel sont les rapports statiques, les rapports dynamiques, la navigation multidimensionnelle et les indicateurs synthétiques. Les fonctionnalités du serveur OLAP Mondrian sont disponibles sous la forme d'une application accessible facilement. C'est un nouveau mode de fonctionnement que l'on découvre peu à peu avec les bases de données multidimensionnelles. Le serveur OLAP Mondrian se compose de quatre couches travaillant depuis l'utilisateur final vers le centre des données. Ces couches sont : la **couche de présentation**, la **couche de calcul**, la **couche d'agrégation** et la **couche de stockage**.

La couche de présentation détermine ce que l'utilisateur final voit sur son moniteur et comment il peut interagir pour effectuer de nouvelles requêtes. Il y a beaucoup de manières de présenter des ensembles de données multidimensionnelles, incluant des histogrammes et des outils de visualisation avancés tels que des cartes cliquables et des graphiques dynamiques.

La seconde couche est la couche de calcul. La couche de calcul analyse, valide et exécute des requêtes MDX.

La troisième couche est la couche d'agrégation. Une agrégation est un ensemble de valeurs de mesures «cellules» dans la mémoire, qualifiée par un ensemble de valeurs de dimensions colonnes.

La quatrième couche est la couche de stockage.

2.1.1 Présentation d'un schéma Mondrian

Un schéma Mondrian définit une base de données multidimensionnelle. Il contient un modèle logique constitué de cubes, de hiérarchies, de membres et une projection de ce modèle vers un modèle physique. Le modèle logique est composé de balises utilisées pour écrire les requêtes dans le langage MDX. Le modèle physique est la source des données qui est présentée à travers le modèle logique. C'est en général un schéma en étoile qui est un ensemble de tables dans une base de données relationnelle : une table centrale du modèle multidimensionnel (table des faits) contient les données numériques ayant un intérêt pour les analyses et des colonnes clés étrangères vers les autres tables du modèle. C'est à partir de ces autres tables satellites que seront construites les dimensions.

2.1.2 Application à notre analyse

La récupération de données offre des analyses pré-calculées. Elle concerne les données de l'analyse bibliométrique exposé en phase expérimentale toujours pour répondre aux besoins de l'acteur enseignant chercheur qui fait un état de l'art sur un thème précis «l'intelligence économique». Le but est de lui offrir la possibilité de faire des analyses multidimensionnelles dans une interface web et à distance. Les données récupérées et analysées aboutissent à des vues métiers par type d'acteur. Une authentification via Openi par

[7] XMLA : Extensible Markup Language Analysis.
[8] MDX : MultiDimensional eXpression.
[9] Mondrian : Serveur OLAP écrit en Java.

login et mot de passe permet de proposer différentes vues selon le profil de l'acteur qui se logue. Avant une explication détaillée dans les paragraphes suivants la figure 4 schématise le scénario de notre application :

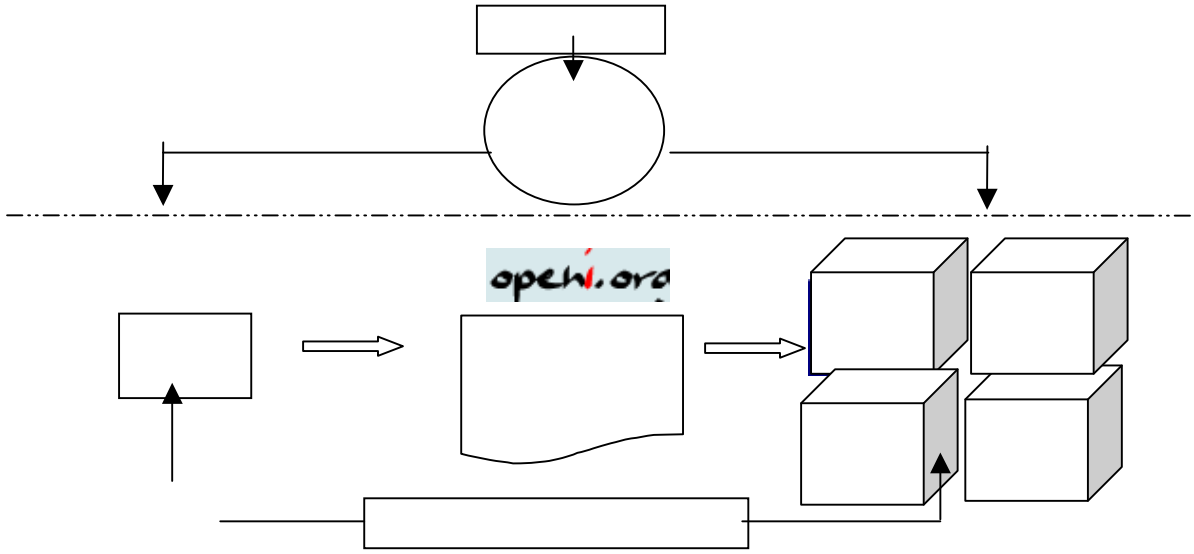


Figure 4. Scénario de notre application

1. Transformation des données

Nous repartons des tableaux excel pour 2001 à 2005 décrits en paragraphe 1.2 *Préalable à l'analyse multidimensionnelle* de ce papier pour en transformer les données en données SQL. Nous avons une base Sql appelée «foodmart» dans laquelle nous créons deux tables : une table de faits et une table satellite.

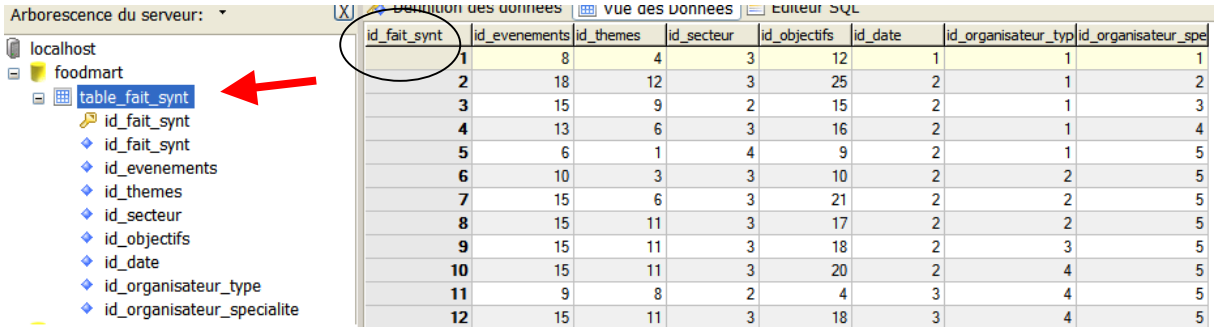
Nous créons une table satellite «synthese» qui comporte des données textes comme le représente la figure 5 ci-dessous :

id_fait_synt	evenements	themes	secteur	objectifs	date	organisateur_t
1	journees adest	mesure	ist	information strategiq	2000	association
2	VSST	veille	ist	traitement des donne	2001	association
3	journees prospective	prospective	entreprise	intelligence strategiq	2001	association
4	journees outils intellig	outils	ie	outils pour managem	2001	association
5	ISKO	exploitation	km	filtrage de l'informatic	2001	association
6	journees futur	gestion	ie	gestion dans l'econo	2001	cab prive
7	journees prospective	outils	ie	prospectives strategi	2001	cab prive
8	journees prospective	tendances	ie	perspectives geopoli	2001	cab prive
9	journees prospective	tendances	ie	prospective socioder	2001	consultant
10	journees prospective	tendances	ie	prospective territorial	2001	ecole inge
11	journees audit	projet	entreprise	enjeux strategiques (	2002	ecole inge
12	journees prospective	tendances	ie	prospective socioder	2002	ecole inge

Figure 5. Données textes de la table «synthese»

Les données de la table «synthese» sont regroupées sous des items qui constituent des métadonnées auxquelles fait appel le shéma en XMLA décrit ultérieurement en figure 8 pour spécifier des dimensions.

La table de faits appelée «table_fait_synt» comportent les données numériques, ainsi que le montre la figure 6 :



id_fait_synt	id_evenements	id_themes	id_secteur	id_objectifs	id_date	id_organisateur_typ	id_organisateur_spe
1	8	4	3	12	1	1	1
2	18	12	3	25	2	1	2
3	15	9	2	15	2	1	3
4	13	6	3	16	2	1	4
5	6	1	4	9	2	1	5
6	10	3	3	10	2	2	5
7	15	6	3	21	2	2	5
8	15	11	3	17	2	2	5
9	15	11	3	18	2	3	5
10	15	11	3	20	2	4	5
11	9	8	2	4	3	4	5
12	15	11	3	18	3	4	5

Figure 6. Données numériques de la table de faits «table_fait_synt»

2. Les cubes

Un cube est formé d'une collection de dimensions et de mesures dans un secteur particulier ainsi que d'une table de faits qui lui est associée. La seule chose que les dimensions et les mesures d'un cube ont en commun est la table de faits de ce cube. Un cube peut contenir des dimensions qui lui sont propres et des dimensions partagées. Le cube permet de créer une dimension en faisant appel à une dimension partagée par une jointure de sa table de faits avec la table des dimensions. Une fois les dimensions créées, on liste les mesures de celui-ci. Une mesure est une quantité qu'il est intéressant de quantifier dans ce cube au travers de ses dimensions. Chaque mesure a un nom, une colonne dans la table des faits et un agrégateur. Cet agrégateur peut être une somme, un maximum ou encore une moyenne. Un cube virtuel est défini par la combinaison de dimensions et de mesures appartenant à d'autres cubes. Dans notre application nous avons créé un cube nommé «synthese» dont nous allons expliciter la création.

3. Création du cube «synthese»

Nous tirons parti des figures 5 et 6 pour montrer les corrélations entre la table satellite «synthese» et la table de faits «table_fait_synt». Tout d'abord voici le principe de fonctionnement :

- Création de la table des dimensions «synthese»

La table de faits permet de définir des mesures. La table de faits ne contient que des clés secondaires et des données numériques.

- Création de la table «table_fait_synthese»

La table de faits ne comporte que des données numériques. Il faut établir une base de données relationnelle. A cette phase nous créons des clés secondaires et des clés primaires qui permettent les relations.

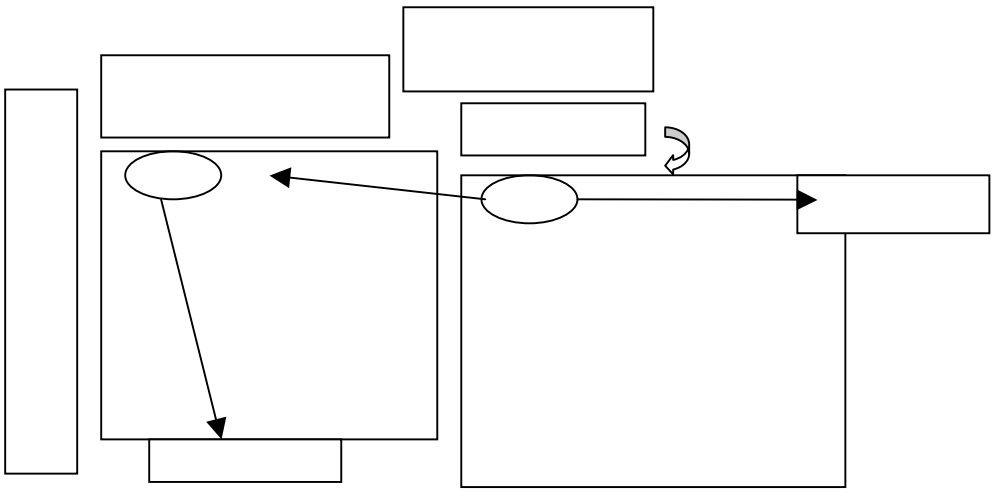


Figure 7. Cube «synthese» au niveau de la base sql

Dans notre application la corrélation entre ces deux tables nous permet par exemple de savoir pour l'acteur «enseignant-chercheur» quel type de colloque le concerne et d'affiner jusqu'à sa spécialité à partir des méta données «evenements» puis «themes». A présent si cet acteur est pluridisciplinaire en informatique, sciences de l'information et communication, documentation alors l'analyse multidimensionnelle permet d'aller plus loin dans la recherche pour révéler une transversalité entre différents colloques.

Après nous être focalisées sur les données de base de données en sql «*foodmart.sql*» concentrons nous à présent sur le schéma en XMLA qui s'appuie sur les données «*foodmart.sql*». Le schéma est un fichier eXtensible Markup Language (XML) qui permet de définir une base de données multidimensionnelle. Dans notre applicatif il est nommé «FoodMart.xml». Le schéma «FoodMart.xml» peut comporter plusieurs cubes distincts que l'on distingue par des balises ouvrantes et des balises fermantes à l'instar des fichiers en XML qui sépare des niveaux de données. Dans notre schéma «FoodMart.xml» nous identifions notre cube propre à notre analyse bibliométrique sous le nom de **«synthese»** où l'on trouve entre les balises ouvrantes <Cube name="**synthese**"> et les balises fermantes </Cube> les dimensions et les mesures de ce cube.

Nous y avons défini 7 mesures agrégées ou sommées autour des items de colonnes "**id_themes**", "**id_evenements**", **id_secteur**", **id_objectifs**", **id_date**", "**id_organisateur_type**" et "**id_organisateur_specialite**".

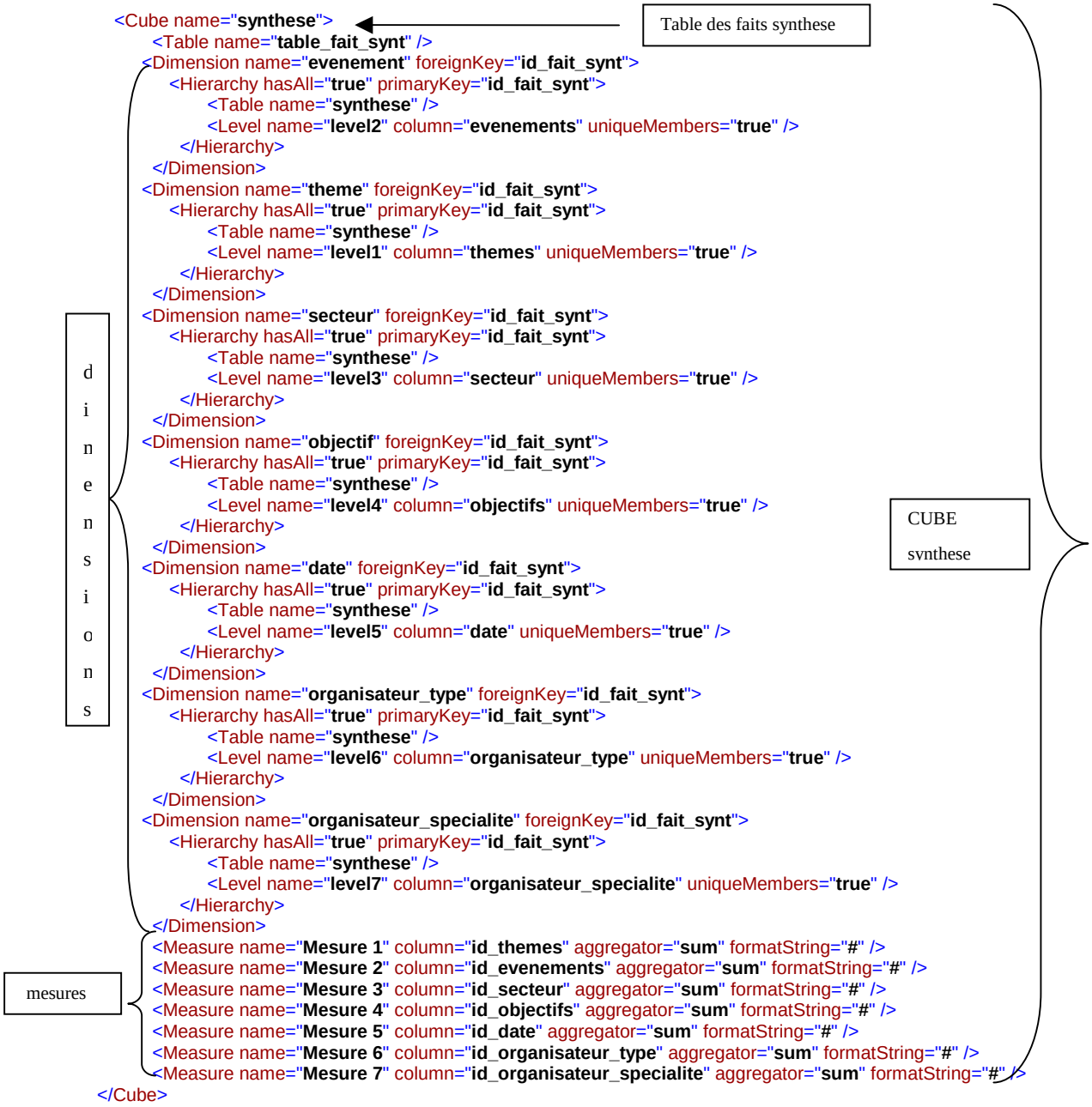


Figure 8. Cube «synthese» au niveau du schéma

Nous avons défini 7 dimensions qui ont pour nom Dimension name=«evenement», Dimension name=«theme», Dimension name=«secteur», Dimension name=«objectif», Dimension name=«date», Dimension name=«organisateur_type» et Dimension name=«organisateur_specialite». Elles sont composées d'une seule hiérarchie. La hiérarchie est constituée d'un niveau par dimension : Level name=«evenements», Level name=«themes», Level name=«secteur», Level name=«objectifs», Level name=«date», Level name=«organisateur_type», Level name=«organisateur_specialite». Chacun de ces niveaux fait intervenir une colonne de la *table_fait_synthese*.

4. Requête MDX associée

Openi offre un éditeur MDX avec une génération automatique de code lors de la création des analyses. Il offre une gestion des projets. Si l'utilisation interactive des cubes est intéressante, leur structure multidimensionnelle associée à la puissance de ce langage en fait des outils de prédilection pour le reporting opérationnel d'entreprise ou d'administration. Au-delà de cet aspect, le MDX permet d'explorer les données des cubes et crée la vraie valeur autour des données et c'est ici que se trouve la plus value.

Notre applicatif décisionnel est accessible à différents acteurs après authentification. Selon les profils des acteurs ils ont des vues adaptées à leur rôle. Selon qu'ils sont étudiants, enseignants-chercheurs, responsable ou administratif ils accèdent à des analyses précalculées ciblées. Par exemple, l'acteur enseignant-chercheur procède à l'analyse bibliométrique OLAP à partir du cube **synthese** et peut exploiter un graphique par les

fonctions Drill Down[10] et Drill up[11] pour effectuer une recherche sur l'item «evenement. Ce procédé de navigation permet la fouille de données et met en évidence que les organiseurs types Prospectives en 2001 et ADBS en 2005 ont consacré le nombre le plus important de journées consacrées à l'événement Intelligence Economique.

De la même façon nous avons complété notre application avec l'analyse des fichiers de log des services de l'ENT consultés via le web. Les services concernent les actualités, les annonces, les cours déposés par les enseignants, les dossiers des étudiants, les signets, l'espace de stockage. L'analyse multidimensionnelle permet à l'acteur enseignant de constater ses cours consultés par les étudiants. Il peut varier le niveau d'analyse à partir de différents paramètres hiérarchisés autour du temps c'est-à-dire par année, par semestre, par trimestre, par mois, par semaine et par jour. Ce type d'analyse amènerait une valeur ajoutée aux statistiques proposées au sein des plate forme d'archives ouvertes où les statistiques ne présentent que des vues à deux dimensions. Les besoins d'évaluation des acteurs dépositaires ou des acteurs gestionnaires varient selon leur rôle d'où une nécessaire granularité dans les possibilités de fouilles de données.

Conclusion

Nous avons montré par ce papier qu'à côté de données chiffrées nous proposons des analyses de contenus documentaires. Pour cela la phase expérimentale a mis en relief la nécessité de travailler autour des métadonnées qui constituent les en-têtes des lignes et des colonnes que l'on retrouve dans le schéma XMLA en vue d'analyses OLAP. Les différents services d'un ENT sont en mesure de proposer des statistiques orientées «infocentre». En exposant les métadonnées de documents électroniques, nous pouvons passer à une dimension supérieure c'est à dire la possibilité de corréler des données via des cubes virtuels provenant de systèmes d'information différents (Open Archive Initiative, Système d'Information Documentaire, fichiers de log…) de façon à élaborer un système d'information stratégique orienté vers la prise de décision au service de l'évaluation d'un établissement de recherche par exemple. Notre plate forme http://perso17.bu.sciences.uhp-nancy.fr:8080/openi/ permet de tester l'analyse bibliométrique multidimensionnelle présentée ici, ainsi que des analyses autour de fichiers de log des services de l'ENT selon des analyses précalculées. Nous avons prolongé nos recherches dans la faisabilité à mener des analyses OLAP interactives après intégration de données en temps réel.

Bibliographie

[1] *Agence de mutualisation des universités*, http://www.amue.fr/Amue/Default.asp

[2] E. Annoni et F. Ravat et O. Teste et G. Zurfluh. Méthode de développement des systèmes d'information décisionnels : essai-erreur *Actes du XXIVème congrès Inforsid Hammamet*, Tunisie, 31 mai-3 juin 2006

[3] A. Belaïd, D. Besagni and N. Benet, *Bibliographic Reference Segmentation for Bibliometrics*, International Workshop on Technology Development in Indian Language (IWTDIL-2003), jan 2003, Calcutta, India

[4] M. Cadot et P. Cuxac et C. François. *Règles d'association avec une prémisse composée : mesure du gain d'information.* EGC 2006: 599-600.

[5] J. Ducloy. *ARTIST et la TEI* Séminaire TEI des 24-25 Mars 2006 au LORIA

[6] S. Dalbin *La modélisation : pourquoi l'intégrer dans les systèmes d'information documentaire ?* La revue Documentaliste - Sciences de l'information, 2003, vol. 40, n° 3, p. 226-231

[7] *Le décisionnel, clé des données structurées : les moteurs de recherche misent sur la capacité de restitution des outils de business intelligence pour remonter les données issues du monde structuré*, O1 Informatique, 2006, juin, p.43

[8] V. Duveau-Patureau. *Le Nouvel enseignant-chercheur : un pédagogue créatif autour de son expertise* http://www.formasup.education.fr/fichier_statique/campus/salon/VDPcompetenseigner.ppt

[9] *Esup portail : Environnement numérique de travail d'accès intégré aux services pour les étudiants et le personnel de l'enseignement supérieur* http://www.esup-portail.org/

[10] E. Fernandez-Medina et J. Trujillo et R. Villarroel et M. Piattini. *Access control and audit model for the multidimensional modeling of data warehouses* In Decision Support Systems vol 42, 2006, p. 1270–1289.

[11] O. Foucaut et O. Thiéry. *L'Evolution des méthodes de conception des systèmes d'information stratégiques*. Conférence invitée au Symposium sur les Systèmes d'Informations Stratégiques, Luxembourg, 1996.

[10] Drill Down : Forer vers le bas. Aller du général au particulier dans une recherche d'information dans une base de données multidimensionnelle. Détailler selon une dimension, par exemple année, Mois et Semaine.

[11] Drill up : Analyse de données à un attribut parent. Remonter dans la hiérarchie dune dimension.

[12] J.M. Franco. *Le Data Warehouse : objectifs, définitions, architectures*, Eyrolles, 1997.

[13] Giorgini P., Rizzi S., Garzetti M. *Goal-Oriented Requirement Analysis for Data Warehouse Design*, In Decision Support Systems, 2007

[14] S. Lainé-Cruzel. *Appropriation, mutualisation, expérimentations des technologies de l'information scientifique et technique*. Paru dans : Partie 1 http://ametist.inist.fr/personne.php?id=151&type=auteur.

[15] F. Peguiron. *Application de l'Intelligence Economique dans un Systeme d'Information Strategique universitaire : les apports de la modelisation des acteurs.* Thèse. Université Nancy II - 2006-11-16, Odile Thiéry (Dir.)

[16] F. Peguiron et Thiery. *Modélisation des acteurs et des ressources : application au contexte d'un SIS universitaire,* ISKO2005, Nancy,

[17] F. Peguiron et O. Thiery. *Système d'information stratégique dédié à l'environnement universitaire*, COSI2005, Bejaia.

[18] E.B. Renaud. *Google se met au service du reporting*, O1 Informatique, 2006, mars, p.17.

[19] D. Rongeat. *Intégration dans les ENT*, Esup Days 26 janvier 2007.

[20] *Schéma directeur des espaces numériques de travail, Ministère de la jeunesse, de l'éducation nationale, et de la recherche*, 2004, http://www.educnet.education.fr/chrgt/SDET-v1.doc.

[21] H. Tardieu et B. Guthmann. *Le Triangle stratégique*. Les Editions d'Organisation, 1991.

[22] Y. Toussaint. *Extraction de connaissances à partir de textes structurés* (Knowledge extraction from structured texts) In Document numérique, 2004, vol. 8, no 3

[23] M. Varandat. *Avez-vous nommé votre gouverneur de données ?* O1 Informatique, 2005, octobre, p. 44-46.