ENTREPÔTS, REPRÉSENTATION & INGÉNIERIE des CONNAISSANCES

Laboratoire ERIC
Equipe d'Accueil 3083

Rapport d'activité
2009-2012

http://eric.univ-lyon2.fr

Lyon 1

UNIVERSITÉ LUMIÈRE LYON 2

UNIVERSITÉ DE LYON

RÉSUMÉ

SUMMARY

Résumé

Les activités de recherche du laboratoire ERIC visent à valoriser les grandes bases de données complexes, notamment dans les domaines des sciences humaines et sociales. Les champs d'expertise d'ERIC couvrent les problématiques liées à la modélisation et l'exploitation des entrepôts de données complexes, la fouille de données hétérogènes, massives et peu structurées et les processus d'aide à la décision.
Le laboratoire ERIC est composé de 52 membres : 22 enseignants-chercheurs, 1 BIATOSS, 23 doctorants, 3 post-doctorants et 3 membres associés. Il est structuré en deux équipes de recherche : Systèmes d'Information Décisionnels (SID) et Data Mining et Décision (DMD).

- La production scientifique globale sur la période 2009-2012 couvre 42 articles dans des revues internationales (dont 8 de rang A selon l'ERA), 84 articles dans des conférences internationales (dont 17 de rang A selon l'ERA), 11 ouvrages directions d'ouvrages ou de revues, 43 articles dans des revues et des conférences nationales et 76 autres publications. 4 logiciels ont également été produits. 10 thèses et 2 habilitations à diriger des recherches ont été soutenues.
- Les ressources financières du laboratoire sont en constante progression (+30 % par an en moyenne depuis 2009.
- ERIC participe à plusieurs programmes internationaux (dont 2 européens), invite régulièrement des professeurs et des chercheurs étrangers (18 de 2009 à 2012), accueille 8 étudiants en thèse dans le cadre de cotutelles, 2 postdoctorants, et collabore avec une trentaine d'universités du monde entier.
- ERIC participe à plusieurs projets nationaux et locaux (dont 1 ANR dont nous sommes porteurs) et collabore avec une vingtaine d'autres laboratoires français (dont 18 UMR).
- Localement, les membres d'ERIC sont impliqués dans plusieurs structures fédératives (Institut des Sciences de l'Homme, Communauté de Recherche Académique régionale ARC6). Ils contribuent également aux formations d'enseignement supérieur des Universités Lyon 1 et Lyon 2.

Summary

Research activities at ERIC aim at extracting value from large, complex databases, especially in the domains of humanities. ERIC's fields of expertise include the problematics related to modeling and exploiting complex data warehouses, mining heterogeneous, massive and loosely-structured data, as well as decision-support processes.
The ERIC lab is constituted of 52 members: 22 professors and associate professors, 1 administrative staff members, 23 Ph.D. students, 3 postdoctoral fellows and 3 associate members. It is composed of two research teams: Decision-support Information Systems (DIS) and Data Mining and Decision (DMD).

- ERIC's global scientific production from 2009 to 2012 includes 42 papers in international journals (8 of which are ranked A by the ERA), 84 papers in international conferences (17 of which are ranked A by the ERA), 11 books, edited books or journal issues, 43 papers in domestic journals and conferences and 76 other publications. 4 software prototypes have also been designed. 10 Ph.D. theses and 2 qualifications for supervising research have been defended.
- The lab's budget is continuously growing (+30% per year on an average since 2009).
- ERIC is involved in several international programs (including 2 European projects), regularly invites foreign professors and researchers (18 from 2009 to 2012), and houses 8 Ph.D. students in collaboration with foreign labs, 2 postdoctoral fellows, and collaborates with about 30 universities worlwide.
- ERIC is involved in many national and local projects (including 1 ANR project we lead) and collaborates with more than 20 research structures in France.
- Locally, members of the lab are involved in several federating bodies (Human Sciences Institute and regional Academic Research Community #6). They also contribute to the higher education activity of the universities Lyon 1 and Lyon 2.

SOMMAIRE

Sommaire

Section 1

Présentation du laboratoire

1 Introduction

Le laboratoire ERIC (Entrepôts, Représentation et Ingénierie des Connaissances) est une unité de recherche (Équipe d'Accueil 3083) dont les établissements de tutelle sont l'Université Lumière Lyon 2 et l'Université Claude Bernard Lyon 1.

Fondé à Lyon 2 en 1995, le laboratoire ERIC a opéré à partir de 2009 un regroupement, officialisé en 2010, avec l'équipe MA²D (Méthodes et Algorithmes pour l'Aide à la décision) de Lyon 1, sur la base de complémentarités thématiques en matière de recherche dans le domaine du décisionnel, d'une sensibilité commune au domaine des Sciences Humaines et Sociales (SHS) et de nombreuses collaborations sur les plans scientifiques et pédagogiques. La période 2009-2012 a donc été un moment charnière d'intégration de nouveaux collègues au sein d'ERIC.

Le laboratoire ERIC occupe une position originale dans le paysage informatique lyonnais, d'une part par son positionnement scientifique ciblé sur l'informatique décisionnelle, alors que les trois autres laboratoires d'informatique de la place de Lyon, le LIRIS[1], le LIP[2] et le DISP[3], sont positionnés sur des créneaux plus généraliste pour le premier (image, données, connaissances, services) et également très spécialisés (parallélisme et systèmes de production, respectivement) pour les deux autres.

Toutefois, des proximités thématiques existent avec des équipes du LIRIS et du DISP, qui se traduisent actuellement par des participations croisées à des jurys de thèse, par exemple, et sur la base desquelles nous pouvons développer de nouvelles collaborations, notamment dans le cadre de projets régionaux (Section 3 Partie 1).

D'autre part, la tutelle de l'Université Lyon 2 et l'intégration en 2012 du laboratoire à l'Institut des Sciences de l'Homme (ISH)[4] de Lyon, Unité de Service et de Recherche multitutelle dirigée par un membre d'ERIC, ouvre de nouveaux terrains d'application privilégiés au laboratoire dans les domaines des lettres et des sciences humaines et sociales. Le recrutement en 2010 par ERIC du directeur scientifique adjoint en charge des Maisons des Sciences de l'Homme (réseau dont l'ISH est membre) et des Instituts d'Etudes Avancées à l'Institut des Sciences Humaines et Sociales du CNRS a encore renforcé ces synergies.

Au plan national, ERIC est leader des communautés scientifiques qui se sont constituées autour des entrepôts de données (journées EDA) et de la fouille de données (conférence EGC).

Enfin, l'activité du laboratoire est reconnue au plan international, comme en attestent la production (revues internationales Computational Intelligence, European Journal of Combinatorics, Computational Materials Science, Medical and Biological Engineering and Computing, Computational Statistics & Data Analysis, IEEE Transactions on Knowledge and Data Engineering… ; conférences internationales FUZZ'IEEE, IEEE CEC, IDA, MICCAI, GECC, ECML/PKDD, AIME, EACL, ICDM, IJCAI, PAKDD, IJCNN, ICONIP… (Annexe 5), le rayonnement scientifique de ses membres (participation aux comités de lecture des revues internationales EEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Multimedia, European Journal of Operation Research, Data & Knowledge Engineering, Journal of Intelligent Information Systems, Journal of Decision Systems… ; des conférences internationales ECML-PKDD, PAKDD, AAMAS, DEXA, ADBIS, DOLAP, ICTAI… ; Section 1. Partie 9) et les collaborations et projets internationaux qu'ils mènent (projets européens ECHOUTCOME et FLURESP, notamment ; Section 1. Partie7).

1 http://liris.cnrs.fr
2 http://www.ens-lyon.fr/LIP/
3 http://disp-lab.fr
4 http://www.ish-lyon.cnrs.fr

2 Thèmes de recherche

Les activités de recherche du laboratoire ERIC visent à valoriser les grandes bases de données complexes, notamment dans les domaines des SHS, mais aussi en lien avec l'industrie à travers des thèses CIFRE et des contrats. Les recherches du laboratoire ERIC se situent dans les domaines suivants :

- **les entrepôts de données :** intégration intelligente de données complexes, modélisation multidimensionnelle d'objets complexes, analyse en ligne personnalisée, sécurité du processus d'entreposage ;
- **la fouille de données et la décision :** apprentissage automatique, étude et fouille de graphes, analyse de données complexes, agrégation multicritère, fouille d'opinion, logiciels de fouille de données.

3 Organisation

Compte-tenu des axes de recherche développés au sein d'ERIC, le laboratoire s'est progressivement structuré d'un point de vue scientifique, depuis le regroupement de 2009, en deux équipes :

- Systèmes d'Information Décisionnels (SID), composée de 8 membres permanents ;
- Data Mining et Décision (DMD), composée de 14 membres permanents.

Chaque équipe est gérée par un responsable et dispose d'un budget propre depuis 2010, pris sur la dotation du laboratoire et établi au prorata des effectifs. Chaque équipe renforce ces ressources par des contrats et des projets. Des collaborations existent entre les équipes, à travers des coencadrements de thèses et des projets transversaux.

D'un point de vue fonctionnel, l'organisation du laboratoire est présentée dans la Figure 1. Compte-tenu de la double tutelle d'ERIC, il est apparu indispensable que le laboratoire soit dirigé par un directeur d'un des deux établissements de tutelle et un directeur adjoint de l'autre. Le conseil de laboratoire, qui se réunit tous les mois, étant constitué de tous les membres permanents, d'un représentant du personnel administratif, d'un représentant des doctorants et post-doctorants et d'un représentant des membres associés, nous avons mis en place une instance collégiale intermédiaire : le conseil de direction, constituée du directeur, du directeur adjoint, des responsables d'équipes de recherche et du responsable administratif. Le conseil de direction se réunit tous les mois, entre les conseils de laboratoire, et a pour fonction de proposer à ce dernier des éléments d'animation scientifique et de stratégie de laboratoire.

Les statuts du laboratoire sont fournis en Annexe 1.

Sur le plan matériel, ERIC ne dispose plus de locaux à Lyon 1 depuis 2011, hormis les bureaux d'enseignants de deux collègues. Le laboratoire est donc provisoirement intégralement hébergé à Lyon 2, sur le campus de la Porte des Alpes, où un espace a été libéré pour les collègues de Lyon 1 de passage.

Ces locaux, partagés avec le Département Informatique et Statistique de l'ICOM-Lyon 2, représentent une surface utile de 400 m² et sont clairement insuffisants compte-tenu de l'effectif global du laboratoire. Nous sommes malgré tout parvenus à ménager des espaces communs pour stocker un petit fonds documentaire constitué de revues (IEEE TKDE, ACM SIGMOD Record, ACM SIGKDD Explorations, Computer, la Lettre SPECIF) auxquelles le laboratoire est abonné en propre. L'équipe SID dispose également d'une armoire faisant office de bibliothèque, hébergée dans un bureau d'enseignants-chercheurs.

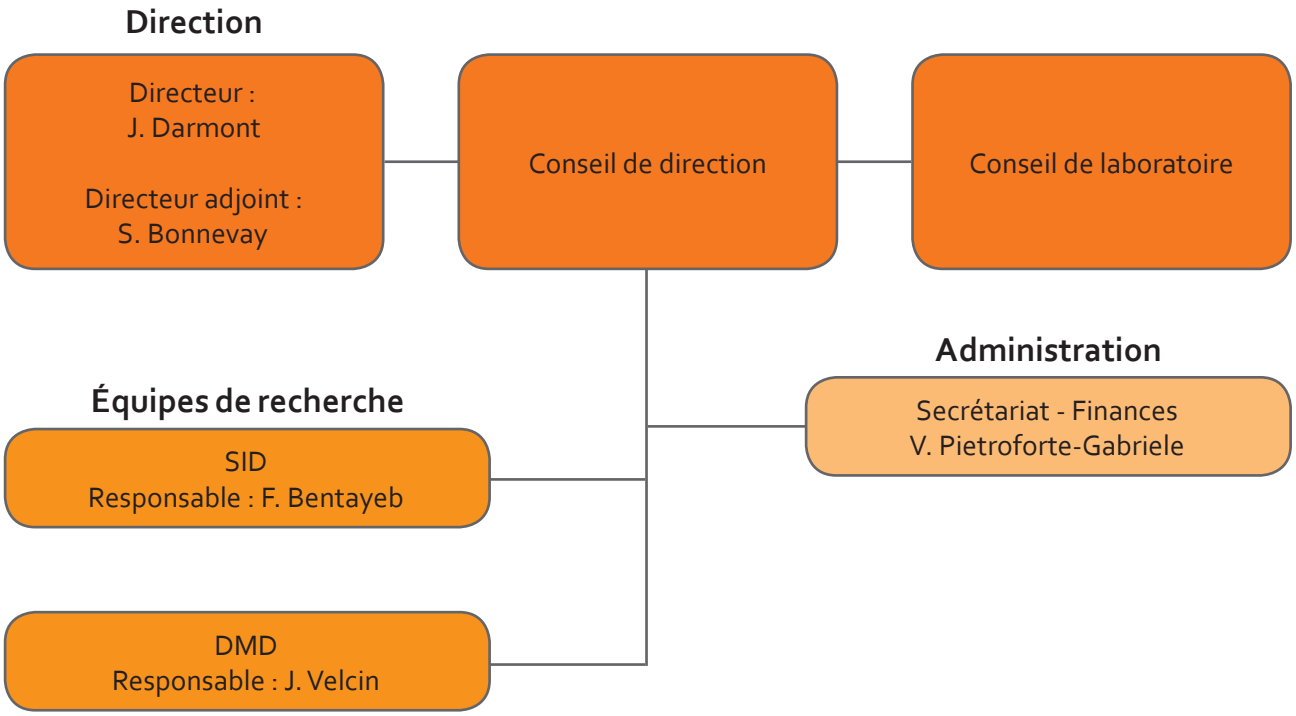

Figure 1 : Organigramme d'ERIC

4 Personnel

52 membres

22 enseignants-chercheurs

Au 1er juillet 2012, le laboratoire ERIC est composé de 52 membres, dont 22 enseignants-chercheurs, 1 personnel BIATOSS, 23 doctorants, 3 post-doctorants et 3 membres associés.

Sur les 22 enseignants-chercheurs permanents, 4 bénéficient de la Prime d'Encadrement Doctoral et de Recherche (PEDR) ou de la Prime d'Excellence Scientifique (PES), soit 18 % d'entre eux.

L'évolution des effectifs d'ERIC depuis 2009 est donnée dans la Figure 2.

L'augmentation significative des effectifs en 2010 correspond au regroupement avec l'équipe MA2D.

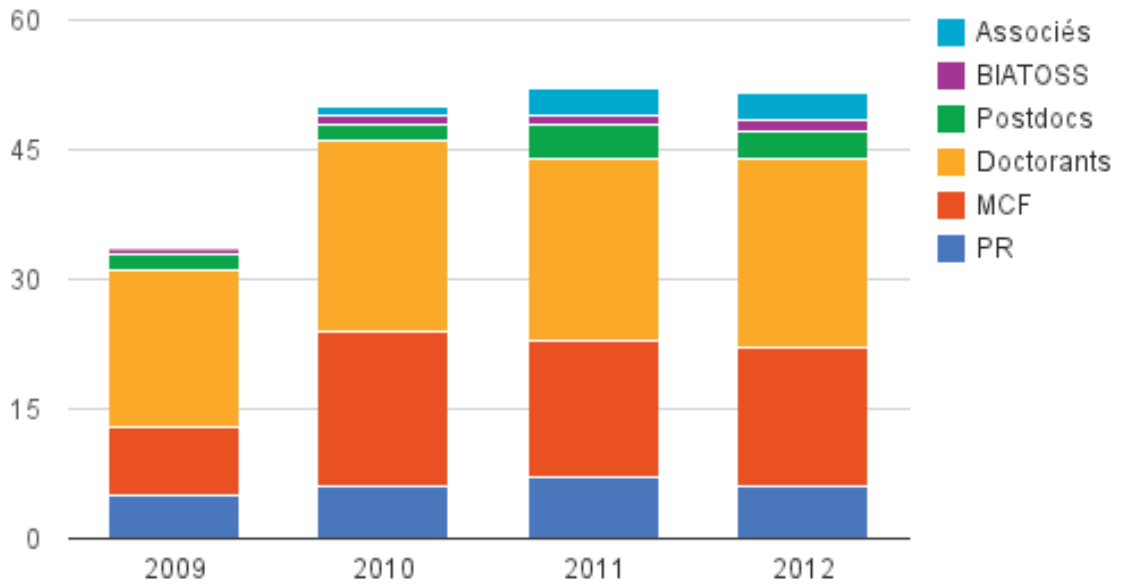

Figure 2 : Évolution des effectifs d'ERIC

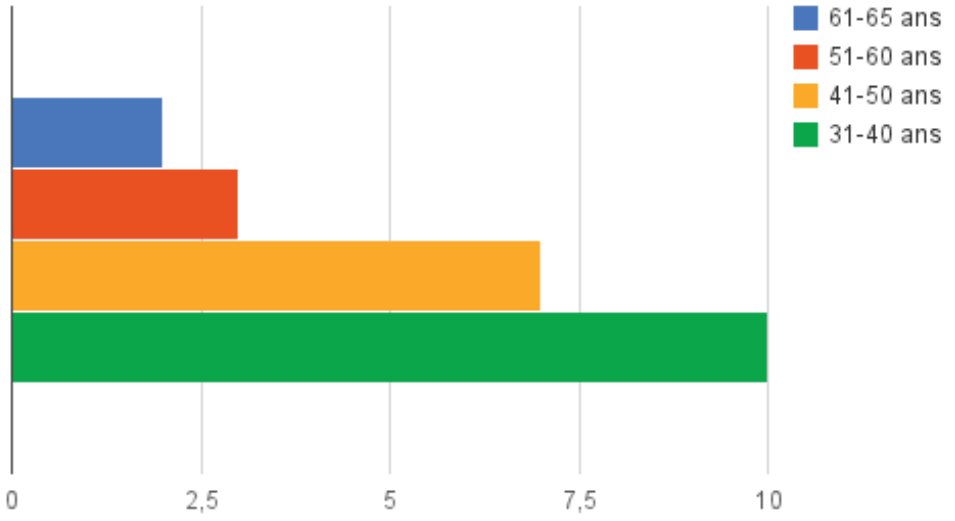

Figure 3 : Distribution des âges des enseignants-chercheurs d'ERIC

Âge moyen : 45 ans

Forte représentation des 31-40 ans

La distributions des âges des enseignants-chercheurs d'ERIC est donnée dans la Figure 3.

L'âge moyen est d'environ 45 ans et la tranche d'âge la plus représentée est celle des 31-40 ans.

La Figure 4 précise l'origine des recrutements effectués au laboratoire depuis 2009. Sont considérés comme recrutements internes les recrutements de collègues ayant préparé leur thèse ou leur HDR au sein d'ERIC. L'apport d'enseignants-chercheurs externes lié au regroupement avec l'équipe MA2D est également dissocié des recrutements effectifs, tous effectués à Lyon 2.

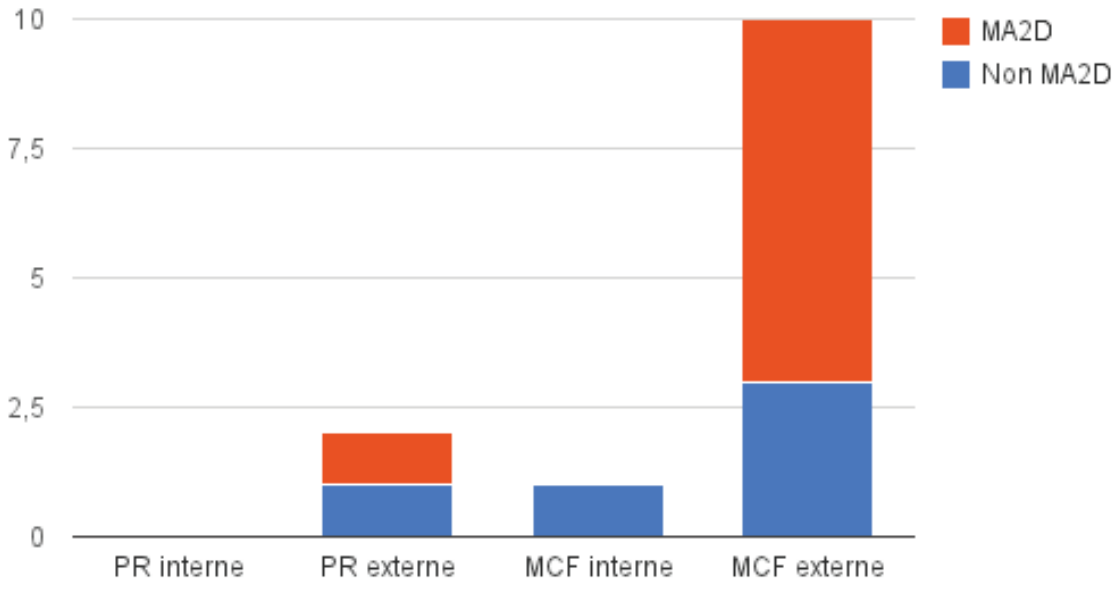

Figure 4 : Origine des recrutements

Depuis 2010, la responsabilité administrative d'ERIC est assurée par un personnel BIATOSS affecté à temps plein par l'Université Lyon 2. Occasionnellement, le laboratoire recrute sur ses fonds propres des personnels en CDD pour mener à bien des missions précises.

Par exemple, un programmeur a été embauché (six mois à temps partiel en 2011-2012) pour mettre au point un outil de reporting des publications du laboratoire. Enfin, ERIC fait également ponctuellement appel aux services du technicien informatique du Département Informatique et Statistique de la Faculté de Sciences Économiques et de Gestion de Lyon 2, dont il partage les locaux. Ces prestations sont facturées à ERIC par le département de formation.

Enfin, le laboratoire ERIC est majoritairement alimenté en doctorants par le biais de son réseau international, grâce à une bonne attractivité de boursiers d'excellence financés par leur gouvernement et à de nombreuses cotutelles (Figure 5). Toutefois, malgré un environnement très concurrentiel, ERIC est un membre actif de l'École doctorale InfoMaths[5] de Lyon et dispose de divers contacts industriels, ce qui permet de fournir un débouché aux étudiants du Master pro-recherche ECD (Extraction de Connaissances à partir des Données) de Lyon 2 qui souhaitent poursuivre en thèse avec un financement CDU ou CIFRE. Enfin, quelques doctorants travaillant dans le domaine de la santé et relevant de l'École doctorale EDISS[6] sont salariés des secteurs pharmaceutique et médical.

Les profils de tous les membres du laboratoire sont présentés de façon synthétique dans l'Annexe 6.

5 http://infomaths.univ-lyon1.fr
6 http://www.ediss-lyon.fr

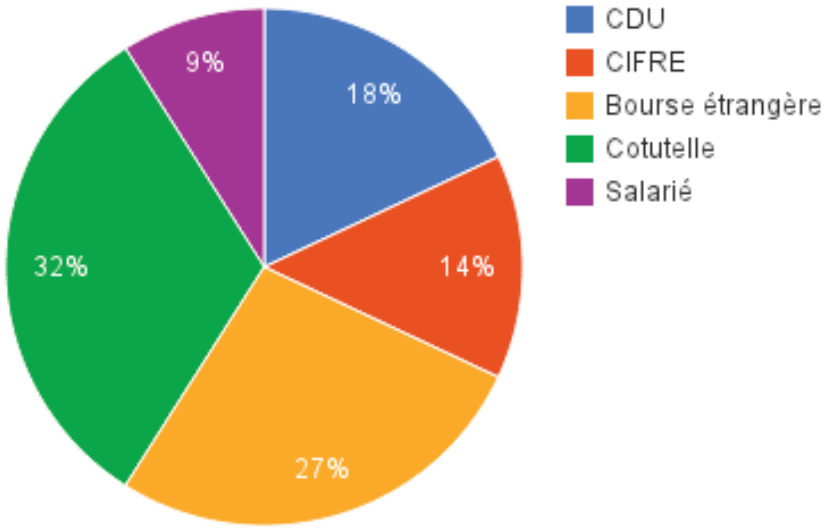

Figure 5 : Mode de financement des doctorants

5 Production scientifique

Un effort important consenti pour augmenter le nombre de publications internationales du laboratoire

+44 % par rapport à la période 2005-2008

En termes de production scientifique, ERIC a accentué depuis 2011 sa politique de promotion des publications :

1) dans les revues internationales ;

2) dans les conférences internationales classées par l'ERA (Excellence in Research for Australia)[7], seule référence dont nous disposons à l'heure actuelle.

Les Figures 6 et 7 comparent la production scientifique d'ERIC pendant les périodes 2005-2008 et 2009-2012, en termes de type de publication et de rang des publications internationales, respectivement.

La Figure 6 montre une augmentation du volume de publication du laboratoire de 20 %, essentiellement due au regroupement avec l'équipe MA2D en 2010.

7 http://www.arc.gov.au/era/

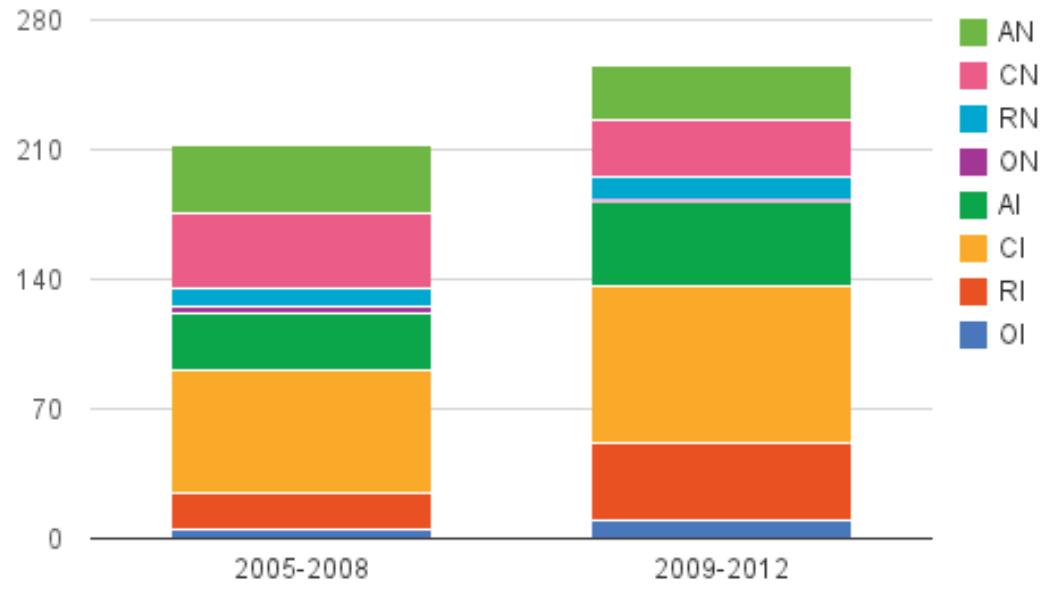

Figure 6 : Évolution de la production scientifique par type de publication

Codification employée :

O - Ouvrages et direction d'ouvrages
R - Revues
C - Conférences avec comité de lecture et actes

A - Autres publications
I - Portée internationale
N - Portée nationale

Ce phénomène masque toutefois une baisse de 20 % du nombre de publications nationales, du fait de l'effort consenti par ailleurs sur les publications internationales, qui ont augmenté de 50 % en volume.

Notre volonté :

Soutenir de nouveaux supports ou des thématiques émergentes

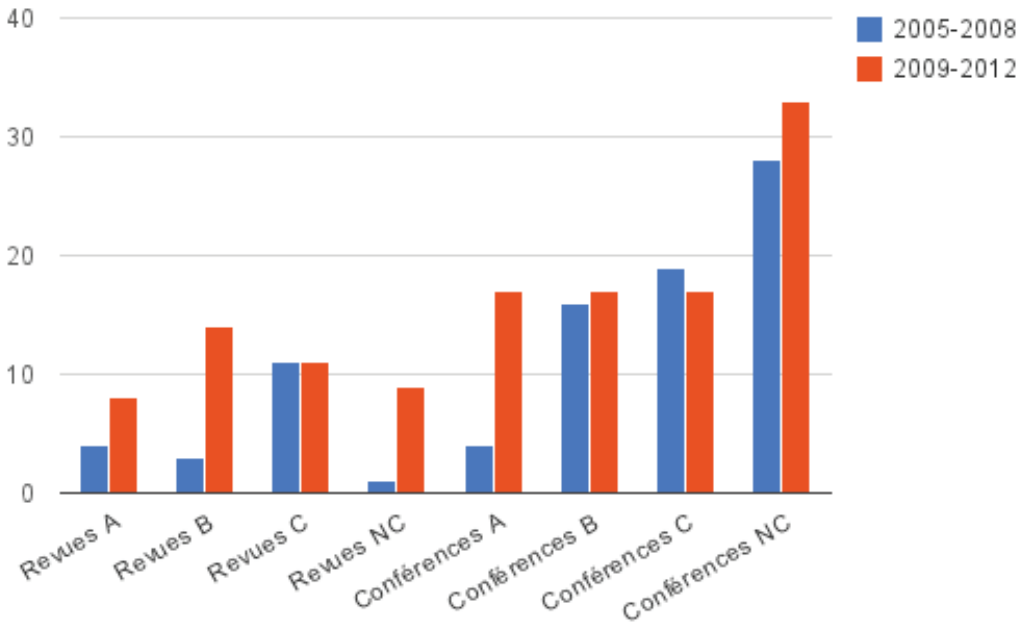

Figure 7 : Évolution de la production scientifique internationale par rang

D'un point de vue qualitatif, la Figure 7 illustre les effets de la politique de publication mise en oeuvre au laboratoire, avec notamment un doublement du nombre de publications dans des revues internationales, dû entre autres au report de publications dans les chapitres d'ouvrages sur des revues.

L'effort effectué sur les publications dans des supports de qualité (revues de rang A et B et conférences de rang A du classement ERA) apparaît également nettement.

Par ailleurs, le nombre de revues et surtout de conférences non classées s'explique dans notre volonté d'implication dans les communautés scientifiques auxquelles nous appartenons. Ce sont souvent des supports nouveaux ou des thématiques émergentes qu'il nous paraît important de soutenir en plus des revues et conférences établies. Nous pouvons citer à titre d'exemples la revue International Journal of Data Warehousing and Mining ou la conférence internationale Advances in Social Networks Analysis and Mining (ASONAM).

La Figure 8 présente une synthèse chiffrée de la production d'ERIC par membre permanent sur la période 2009-2012. La moyenne est d'environ 11 publications par membre.

Par ailleurs, les membres d'ERIC cosignent des publications avec de nombreux autres chercheurs, que se soit au sein du laboratoire (dans l'autre équipe que la leur), sur la place de Lyon (notamment à l'occasion de collaborations pluridisciplinaires) ou aux niveaux national et international (Figure 9).

L'intégralité des références des publications des membres d'ERIC sur la période 2009-2012 est fournie dans l'Annexe 7.

Enfin, 10 thèses et 2 habilitations à diriger des recherches ont été soutenues à ERIC sur la période 2009-2012.

En moyenne, 11 publications par membre

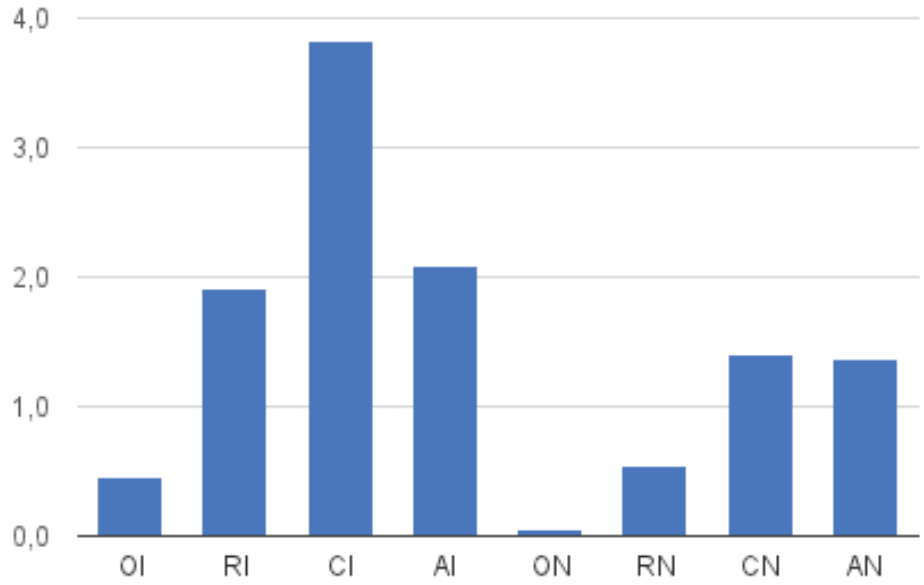

Figure 8 : Synthèse de la production scientifique par membre permanent

Codification employée

O - Ouvrages et direction d'ouvrages
C - Conférences avec comité de lecture et actes
R - Revues
A - Autres publications
I - Portée internationale
N - Portée nationale

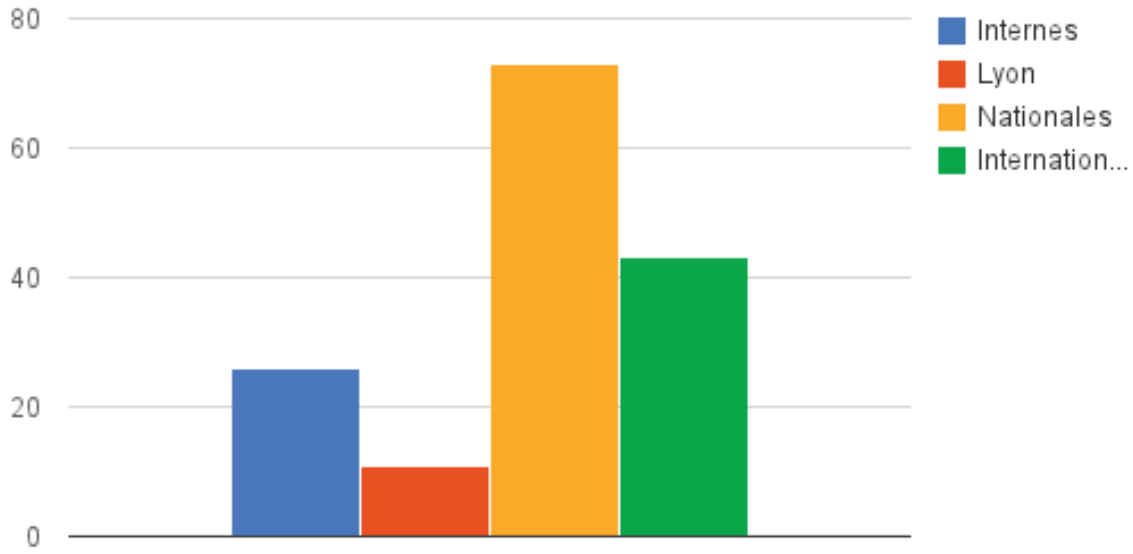

Figure 9 : Publications en collaboration

De plus, 5 membres d'ERIC ont demandé et obtenu leur qualification (3 PR, 2 MCF), soit la totalité des candidats. La durée moyenne des thèses sur la période 2009-2012 est d'environ quatre ans. Cette durée supérieure à la moyenne est due aux nombreuses cotutelles gérées par le laboratoire, dont la durée est typiquement supérieure à trois ans, ainsi qu'aux boursiers étrangers, dont certains sont d'emblée financés sur quatre ans (par exemple, les bourses du gouvernement égyptien) afin de permettre l'apprentissage du Français et l'adaptation des doctorants.

Pour finir, nous effectuons un suivi systématique du devenir de nos docteurs. 30 % d'entre eux ont obtenu un poste dans l'enseignement supérieur et la recherche. Les autres sont en poste dans l'industrie (tous en CDI sauf un, récemment embauché) à l'exception d'une recherche d'emploi en cours (Section 2 Parties 1.4.2 et 2.4.2).

Enfin, un maître de conférences HDR a obtenu un poste de professeur des universités dans un autre laboratoire à compter d'octobre 2012.

6 Finances

Le budget d'ERIC provient de différentes sources de financement que nous classons en deux grandes familles :

1. les Dotations Générales de Fonctionnement (DGF) allouées par les universités de tutelle Lyon 1 et Lyon 2 ;
2. les contrats académiques (Europe, ANR, région Rhône-Alpes) ou industriels (dont CIFRE) et autres subventions, sur lesquels est opéré un prélèvement de 10 % (hors ressources fléchées) par le laboratoire.

Ces recettes, dont l'évolution (hors taxe) depuis 2009 est présentée dans la Figure 10, sont allouées, d'une part, aux dépenses mutualisées du laboratoire (fonctionnement courant, investissement) et, d'autre part, aux budgets propres des équipes de recherche, établis au prorata des effectifs (permanents et doctorants).

Ce graphique montre la croissance importante des recettes du laboratoire, grâce à une augmentation de la DGF en 2010 suite à la fusion avec l'équipe MA2D, puis, la DGF diminuant lentement, mais régulièrement ensuite, surtout grâce à l'effort porté sur l'obtention de contrats de recherche européens et ANR.

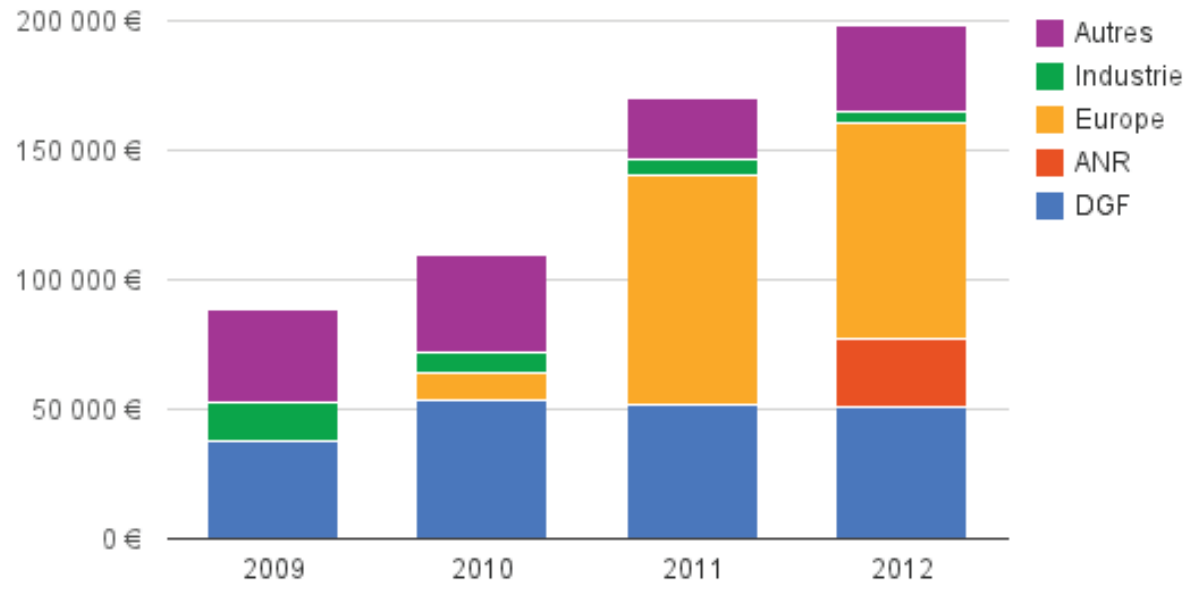

Figure 10 : Évolution des recettes

L'évolution des dépenses globales (TTC) du laboratoire depuis 2009 est récapitulée dans la Figure 11. Elle suit bien sûr celle des recettes et montre clairement que le poste principal de dépense est le fonctionnement (dont les missions).

Les dépenses de personnel correspondent en 2009-2010 au financement à temps partiel (20 %) d'un technicien chargé de la maintenance du parc informatique, et en 2011-2012 à l'embauche en CDD pendant six mois d'un développeur informatique qui a travaillé sur les outils Intranet du laboratoire, ainsi qu'aux salaires des deux postdoctorants accueillis au laboratoire et de contractuels qui ont travaillé sur nos projets européens.

L'exercice 2012 n'étant pas clos, les dépenses ne sont pas encore totalement reportées sur la figure 11 pour cette année. Elles incluent, dans la masse personnel, le salaire d'un doctorant rémunéré sur le contrat ANR ImagiWeb.

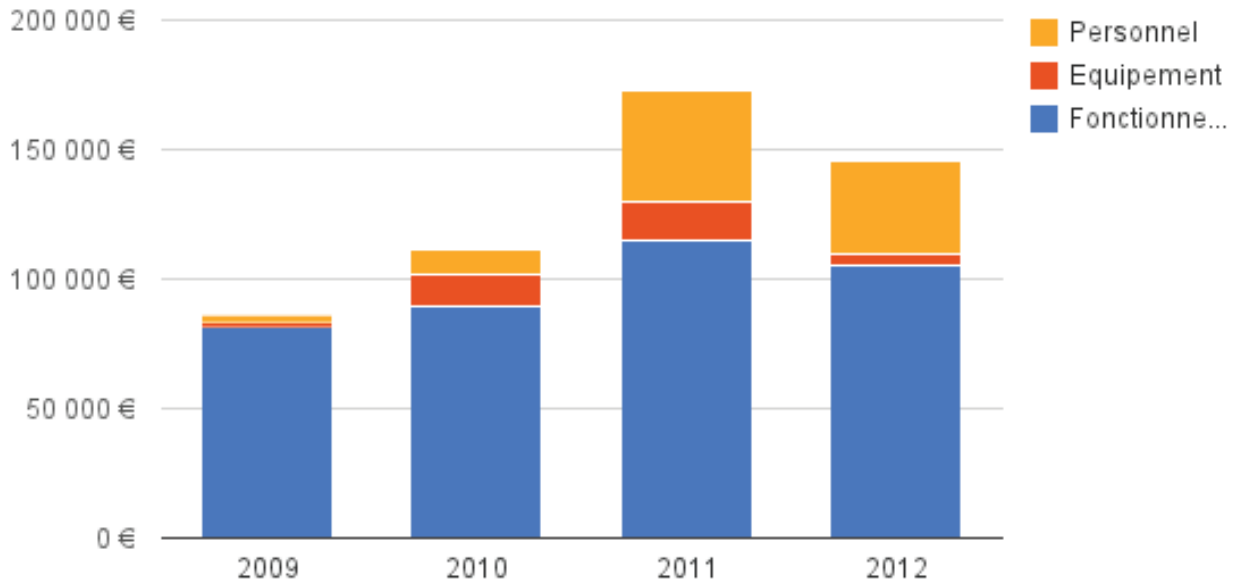

Figure 11 : Évolution des dépenses

7 Projets et coopérations

7.1. Au plan international

- 2011-2014 : FLURESP[8] (responsable scientifique). Définition des principaux scénarios d'une pandémie humaine au niveau européen, description des stratégies de réponses possibles et évaluation de ces stratégies d'intervention dans un cadre d'analyse multi-critères et des analyses coût-efficacité - Financement (Union européenne) : 201000 €.
- 2010-2013 : ECHOUTCOME[9] (responsable scientifique). Étude des systèmes de santé européens dans le but d'évaluer la prise de décision dans le cadre des critères des besoins nationaux et les attentes dans tous les états membres concernant les résultats des soins de santé et analyses coûts-utilité - Financement (Union européenne) : 282000 €.
- 2011-2012 : TASSILI (porteur). Mobilité entrante dans le cadre du partenariat Hubert Curien franco-algérien - Financement (Egide) : 19000 €.
- 2011-2012 : CMIRA. Accueil d'une doctorante en cotutelle pendant six mois (Elena Orobinska, Ukraine) - Financement (Région Rhône-Alpes) : 4300 €.
- 2010-2014 : Subvention formation doctorale (deux thèses encadrées à ERIC) - Financement (gouvernement algérien) : 49000 €.
- Thèses en cotutelle ou en coencadrement : Université Nationale d'Economie de Kharkov (Ukraine), Universités de Sfax et de Tunis (Tunisie), ENSA Agadir (Maroc), Universités d'Alger, de Laghouat, de Jigel, de Blida et d'Oran (Algérie)

8 http://www.fluresp.eu
9 http://www.echoutcome.eu

7.2. Au plan national

- 2012-2015 : ImagiWeb (porteur). Etude de l'image (au sens de leur représentation) d'entités de diverses natures (entreprises, hommes politiques, etc.) telle qu'elle est émise et perçue sur Internet - Financement (ANR et pôles de compétitivité) : 872000 € dont 160000 € pour ERIC.
- 2012 : G-Graphs and networks (partenaire) - Financement (GDR Recherche Opérationnelle) : 2500 €.
- 2009 : MOUSSON (co-responsable scientifique). Mise en place un système d'alerte à la pollution à Ouagadougou, Burkina Faso - Financement (CNRS) : 5500 €.

7.3. Au plan local

- 2012 : Modèles de grands réseaux, jeux de poursuite et décomposition modulaire, Application au WWW (porteur). ARC6 - Financement (Région Rhône-Alpes) : 3000 €.
- 2011-2012 : DocNet (porteur). Conception d'un portail social sémantique pour l'exposition des compétences et la veille scientifique en SHS - Financement (BQR Lyon 2) : 25000 €.
- 2011-2012 : Accueil post-doctorante sur le projet de recherche «XML On Line Analysis Processing» (responsable scientifique) - Financement (BQR Lyon 2) : 31500 €.
- 2010-2011 : "Le rôle des forums citoyens dans le débat public" (co-porteur). Construction et test d'outils semi-automatiques pour l'étude de la dynamique des discours - Financement (BQR Lyon 2) : 20000 €.
- 2009-2010 : "Construction de nouveaux outils de fouille de données pour les SHS" (porteur) - Financement (BQR Lyon 2) : 25000 €.
- 2009-2010 : "Détection de phénomènes complexes dans les corpus linguistiques oraux" (porteur) - Financement (BQR Lyon 2) : 20000 €.
- 2010 : "Web Intelligence" (membre). Cluster n° 2 Informatique, signal, logiciel embarqués - Financement (Région Rhône-Alpes) : 2000 €.
- 2008-2009 : ProxAn (responsable scientifique). Évaluation des zones de chalandise avec un outil d'aide à l'implémentation des commerces (incubation d'entreprise innovante) - Financement (Région Rhône-Alpes) : 30500 €.
- 2007-2009 : BETWEEN (responsable scientifique). Modèle de représentation et d'analyse des débats en ligne sur Internet (incubation d'entreprise innovante) - Financement (Région Rhône-Alpes) : 20500 €.

8 Collaborations industrielles et création de logiciels

ERIC développe des collaborations avec des partenaires industriels variés.

Ces collaborations s'effectuent dans plusieurs contextes :

- contrats d'études/consulting : Aéroport de Lyon, Buzzinbees, CERTIRA Lyon, Eureval, Rithme, Hôpital du Vinatier (montant global : 27000 €) ;
- thèses sous convention CIFRE (3) : AID, AMI Software, Creative Research (montant global des conventions hors bourses de thèse : 45000 €) ;
- projets ANR (Section 1. Partie 7.2) : AMI Software, EDF, Xerox ;
- autres collaborations : Alcatel Lucent Bell Labs, Technicolor, We Are Cloud, Visioglobe.

Certaines de ces collaborations donnent lieu à des transferts de technologie sous la forme de logiciels prototypes. D'autres logiciels sont également développés au laboratoire pour un public plus académique.

La plupart sont des logiciels libres distribués sous des licences ad-hoc ou Creative Commons[10]. Tous sont téléchargeables depuis le site Web d'ERIC.

- DWEB (Data Warehouse Engineering Benchmark) est un banc d'essais décisionnel permettant d'évaluer les performances des entrepôts de données.
- SMAIDoC est une plate-forme d'intégration de données complexes qui met en oeuvre à la fois les principes de l'entreposage de données classique et la technologie multi-agents.
- TANAGRA est un logiciel libre de data mining à destination des étudiants, des enseignants et des chercheurs. Il implémente une série de méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'analyse de données, de l'apprentissage automatique et des bases de données (178 opérateurs). C'est le plus téléchargé de nos logiciels, avec une moyenne de 20000 accès par mois au site (Figure 12) depuis le monde entier (dont 65 % depuis des pays francophones).
- XWeB (XML Warehouse Benchmark) est le premier (et à notre connaissance le seul) banc d'essais pour entrepôts de données XML.

Le laboratoire contribue également au projet international de développement collaboratif open source Decision Deck[11], qui implémente des méthodes d'aide à la décision multicritères.

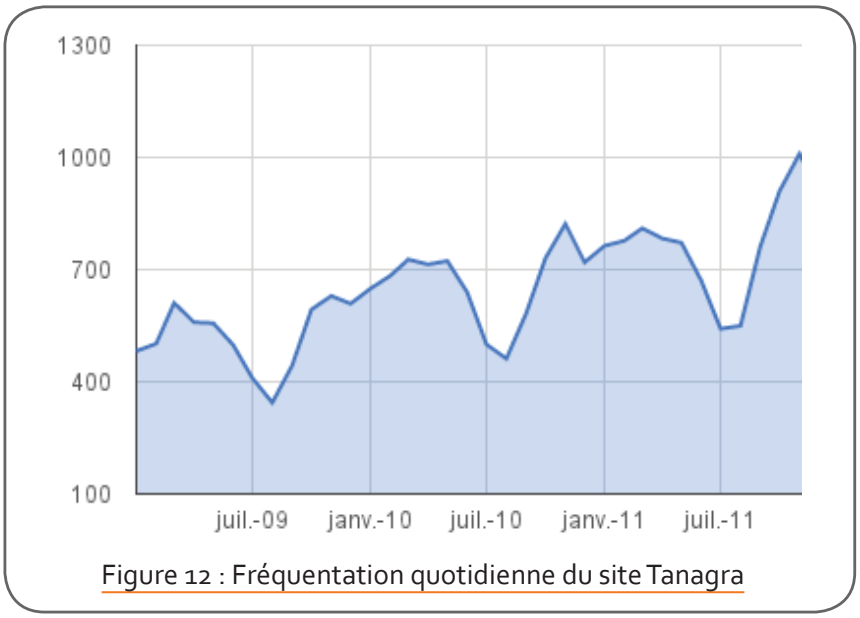

Figure 12 : Fréquentation quotidienne du site Tanagra

10 http://creativecommons.org/licenses/
11 http://ww.decision-desk.org

9 Rayonnement scientifique

Les membres d'ERIC sont impliqués dans de nombreux groupes de travail et ont une activité éditoriale soutenue, résumée dans le Tableau 1 (le comptage est donné pour chaque chercheur).

On y notera, au niveau international, l'évaluation d'articles pour des revues et des conférences majeures du domaine, et au niveau national, l'animation de la communauté scientifique avec la vice-présidence de l'association EGC (Extraction et Gestion de Connaissances), la présidence de l'association RNTI (Revue des Nouvelles Technologies de l'Information ; plus d'une trentaine de numéros édités depuis la fondation de la revue) et le comité de pilotage des journées EDA (Entrepôts de Données et Analyse en ligne).

Le laboratoire est également la cheville ouvrière de l'organisation des conférences internationales Algorithmic Learning Theory (ALT, de rang A dans le classement ERA) et Discovery Science (DS) en 2012 à Lyon.

	Internationaux	Nationaux	Exemples les plus significatifs
Comités éditoriaux	11	12	International Journal of Data Mining, Modelling and Management, International Journal of Data Analysis Techniques and Strategies, International Journal of Social Network Mining, Revue des Nouvelles Technologies de l'Information, conférences QIMIE, EGC, EDA
Comités de lecture* *Évaluation d'articles de revues et de chapitres d'ouvrages	60	14	IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Multimedia, European Journal of Operation Research, Data & Knowledge Engineering, Journal of Intelligent Information Systems, Journal of Decision Systems
Comités de programme	96	85	ECML-PKDD, PAKDD, AAMAS, DEXA, ADBIS, DOLAP, ICTAI
Comités d'organisation	26	16	ALT-DS 2012, WWW 2012, WI-AT 2011, QIMIE 2011, PRETOPOLOGICS 2010, VLDB 2009, MEDES 2009
Jurys de thèse et HDR		55	15 comme rapporteur, 2 HDR
Expertises	10	58	AUF (3), COS (25), ANR (6), ANRT (18)
Invitations	16	19	Séminaire Dagstuhl, Entretiens Jacques Cartier, Ecole d'été Web Intelligence, Professeurs invités : Cantho (Vietnam), Zagreb (Croatie), Oran (Algérie), Luxembourg, Avignon

Tableau 1 : Animation et expertise scientifique des membres d'ERIC

De plus, le laboratoire organise des séminaires réguliers (en moyenne une dizaine par an).

Les intervenants sollicités sont à la fois des universitaires et des industriels (EADS, Orange, XEROX, Masa Group, AMI Software...) venant de toute la France et de l'étranger (Canada, Suisse, Royaume Uni, Italie, Roumanie...). Les séminaires sont ouverts à tous, étudiants de Master compris. Les intervenants de la période 2009-2012 sont listés dans l'Annexe 6.

Un membre du laboratoire a également été impliqué dans l'organisation de l'école d'été Web Intelligence 2010 "Le Web centré sur l'utilisateur".

Enfin, ERIC attire de nombreux collègues et jeunes chercheurs étrangers.

- Professeurs invités (10) : Tomas Aluja (Université Polytechnique de Catalogne), Jean Dumais (Statistique Canada), Stan Matwin (Université d'Ottawa), Rokia Missaoui (Université du Québec en Outaouais), Jan Rauch (Université d'Economie de Prague), Gilbert Ritschard (Université de Genève), Lorenza Saita (Université du Piémont Oriental), Ivan Sidorenko (Université Nationale d'Economie de Kharkov), Stefan Trausan (Université Polytechnique de Bucarest), Iryna Zolotarieva (Université Nationale d'Economie de Kharkov)
- Séjours de recherche (une semaine à six mois) : 8 doctorants étrangers dans le cadre des programmes PROFAS, TASSILI et Erasmus Mundus eLink.

10 Contribution à l'enseignement et à la formation par la recherche

Les membres du laboratoire ERIC sont tous des enseignants-chercheurs relevant des sections 26 (Mathématiques appliquées) et 27 (Informatique) du CNU. Ils sont très impliqués dans la vie des universités Lyon 1 et Lyon 2 et dans leurs enseignements d'informatique, à la fois pour des publics de spécialistes (Départements d'Informatique et de Mathématique de la Faculté des Sciences et Technologies et école d'ingénieurs Polytech à Lyon 1, Département Informatique et Statistique de l'Institut de la Communication[12] et IUT Lumière à Lyon 2) et pour des non-spécialistes (Facultés de Sciences Économiques et de Gestion, de Sociologie et de Droit, Institut de la Communication et IUT Lumière à Lyon 2).

Compte-tenu du sous-encadrement en informatique à l'Université Lyon 2, la majorité des membres d'ERIC effectue plus que son service statutaire minimum.

Au niveau de la vie de ses universités de tutelle, ERIC est représenté à Lyon 1 au conseil d'administration de l'université (jusqu'en 2012) et à celui de Polytech' Lyon ; et à Lyon 2, au conseil de la Faculté de Sciences Économiques et de Gestion (jusqu'en 2012) et au conseil scientifique (depuis 2012).

Nous sommes également représentés à l'école doctorale multi-établissement InfoMaths2 dont nous dépendons et dans le comité des directeurs de laboratoires de l'ISH1, unité de service et de recherche à laquelle nous appartenons.

De plus, deux membres d'ERIC, Djamel Zighed et Bertrand Jouve, sont respectivement directeur de l'ISH et directeur scientifique adjoint en charge des Maisons des Sciences de l'Homme et des Instituts d'Etudes Avancées à l'Institut des Sciences Humaines et Sociales du CNRS. Enfin, les membres d'ERIC sont impliqués dans de nombreuses fonctions pédagogiques et administratives dont le détail est fourni en Annexe 3.

12 Le Département Informatique et Statistique a quitté en 2012 la Faculté de Sciences Economiques et de Gestion pour former un nouveau pôle d'excellence «Communication-Informatique» sur le campus Porte des Alpes avec l'Institut de la Communication, dont le nouveau nom contiendra le terme «informatique».

Les formations animées par des membres d'ERIC et susceptibles d'offrir la possibilité d'une thèse au laboratoire dépendent essentiellement de la spécialité "pro-recherche" Fouille de Données et Gestion des Connaissances du Master d'Informatique de Lyon 2 :

- parcours Extraction des Connaissances à partir des Données (ECD) en partenariat avec Polytech' Nantes (une trentaine d'étudiants en M2) ;
- parcours Erasmus Mundus Data Mining & Knowledge Management (DMKM) en partenariat avec Polytech' Nantes, l'Université Paris 6, l'Université Polytechnique de Catalogne (Barcelone, Espagne) l'Université du Piémont Oriental (Alessandria, Italie) et l'Université Polytechique de Bucarest (Roumanie) (également une trentaine d'étudiants en M2).

Un projet de Doctorat Erasmus Mundus (Eramus Mundus Joint Doctorate) DMKM a également été déposé en 2012 auprès de la Commission Européenne avec les mêmes partenaires que le Master DMKM.

De plus, les étudiants des formations à vocation professionnelles dans lesquelles sont impliqués les membres d'ERIC participent également à la recherche au laboratoire, par des projets ou des stages :

- spécialité Informatique Décisionnelle et Statistique (IDS) du Master d'Informatique de Lyon 2 (entre 130 et 150 étudiants en M2, formation continue incluse) ;
- spécialité e-Miage du Master d'Informatique de Lyon 1 ;
- Licences Informatique Décisionnelle et Statistique (IDS) et Mathématiques et Informatique Appliquées aux Sciences Humaines et Sociales (MIASHS) de Lyon 2 ;
- Licence pro Chargé(e) d'Etudes Statistiques (CESTAT) de l'IUT Lumière.

Les membres d'ERIC sont également très impliqués dans des formations internationales : doubles diplômes de Licence et de Master en informatique à l'université de Ho Chi Minh (Saigon, Vietnam) à Lyon 1, Master Erasmus Mundus DMKM et double-diplôme de Master franco-ukrainien Informatique Décisionnelle et Statistique pour le Management (IDSM-Kharkov) à Lyon 2 ; ce qui contribue à l'attractivité du laboratoire. Un membre d'ERIC et notamment représentant de l'Université Lyon 1 au consortium de l'Agence Universitaire de la Francophonie jusqu'en 2014.

Enfin, un des membres d'ERIC porte un effort particulier sur la diffusion de matériel pédagogique, avec un portail *data mining*[13] hébergeant plus d'une douzaine de e-books et de tutoriels très téléchargés dans le monde universitaire francophone. Un autre membre du laboratoire met également à disposition sur son site Web un tutoriel pour l'apprentissage du langage SQL[14] également apprécié en dehors de nos universités de tutelle.

13 http://eric.univ-lyon2.fr/~ricco/data-minig/
14 http://eric.univ-lyon2.fr/~jdarmont/tutoriel-sql/

11 Formation du personnel, hygiène et sécurité

Le détail des formations suivies par les membres du personnel BIATOSS est fourni en Annexe 4.

Les activités de recherche d'ERIC et les matériels dont nous disposons ne génèrent pas de risque particulier pour le personnel et ne requièrent donc pas de précaution particulière en termes d'hygiène et de sécurité.

En revanche, les locaux dans lesquels est hébergé le laboratoire présentent des défauts d'isolation et d'étanchéité. De l'amiante est également présent dans la colle des dalles du sol et dans certains plafonds mal identifiés.

Ce constat, dressé en 2008, n'a été pris en charge que très partiellement (installation d'une climatisation dans trois salles ou bureaux en 2011, sur fonds propres du laboratoire). Toutefois, aucun accident n'a heureusement été à déplorer.

Enfin, certains membres du laboratoire (direction, administration) ont suivi des formations de secourisme et de gestion des alertes incendies.

12 Ethique

Conformément au règlement de l'École doctorale InfoMaths2 de Lyon dont nous dépendons, tous les doctorants inscrits à ERIC doivent bénéficier d'un financement mensuel net minimum de 1300 € pendant toute la durée de leur thèse. La charte des thèses est également en vigueur et doit être signée par le doctorant et son ou ses encadrants.

Nous distribuerons également dès la rentrée 2012 un document à l'intention des nouveaux doctorants, qui inclura, outre la charte des thèses, un règlement intérieur du laboratoire (en cours d'élaboration), la charte de sécurité de l'Université Lumière Lyon 2 pour la bonne utilisation des outils mis à disposition dans l'environnement numérique de travail, ainsi qu'un texte destiné à sensibiliser les étudiants chercheurs aux règles éthiques à respecter lors de la soumission d'un article (originalité du matériel publié, absence de soumission multiple, référencement des sources, propriété intellectuelle…), basé sur Ethical guidelines for journal publication[15].

Enfin, au niveau de l'encadrement des thèses, nous avons mis en place en 2010 une procédure de soutenance à mi-parcours pour les doctorants de deuxième année, qui permet de faire un point sur l'avancée des travaux et d'avoir l'avis de deux rapporteurs extérieurs à ERIC. Cette soutenance fait l'objet d'un rapport écrit transmis à l'école doctorale. Le taux d'encadrement moyen des enseignants-chercheurs du laboratoire est de 1,2 pour les HDR et 0,4 pour les non-HDR.

15 http://www.elseevier.com/wps/find/authorsview.authors/rights

13 Synthèse des objectifs du projet précédent et des résultats obtenus

L'Annexe 2 présente le rapport des experts qui ont évalué le laboratoire pour le compte de l'AERES sur la période 2005-2008.

Le Tableau 2 résume les objectifs que nous nous sommes donnés suite à cette évaluation, ainsi les actions que nous avons mises en oeuvre pour les atteindre et, le cas échéant, les résultats que nous avons obtenus.

Objectifs	Actions/Résultats
Publications : améliorer la qualité globale de la production scientifique	Incitation à la publication dans les revues internationales, ainsi que dans les conférences notées A, B ou C dans le classement ERA → augmentation de 95 % des articles dans des revues internationales et de 18 % dans des conférences internationales A, B, C
Projets financés : augmenter les ressources du laboratoire et sa visibilité par des projets sélectifs de type ANR et PCRD	Incitation à porter des projets ANR → 2 projets européens en cours et 1 ANR acceptée en 2012, pour un budget global de 405 000 €
Situation du laboratoire : clarifier le positionnement stratégique de l'unité dans son environnement immédiat	Engagement clair sur le terrain d'application des SHS →Intégration à l'ISH ; Recrutement de Bertrand Jouve, ex-directeur de la MSH de Toulouse, qui collabore avec Institut Rhône-Alpin des Systèmes Complexes (IXXI- http://www.ixxi.fr) ; participation au projet de LabEx (non retenu) Humanités et Humanités Numériques (H2N)
Organisation scientifique : adopter une structure plus simple à gérer que la matrice projets-axes de recherche proposée, avec rattachement principal et secondaire des enseignants-chercheurs	Restructuration du laboratoire en deux équipes de recherche (contre trois axes précédemment) avec un seul rattachement par enseignant-chercheur

Tableau 2 : Objectifs du précédent projet et résultats obtenus

Section 2

Bilan des équipes de recherche

1 Équipe SID

1.1 Membres de l'équipe

Responsable : Fadila BENTAYEB

Nom	Prénom	Statut
AKNOUCHE	Rachid	Doctorant (Bourse Algérie)
ASFARI	Ounas	Post-doctorant (ATER)
ATTASENA	Varunya	Doctorante (Bourse Thaïlande)
BEN HASSINE-GUETARI	Soumaya	Doctorante (CIFRE) - Co-direction SID-DMD
BENTAYEB	Fadila	MCF HDR (Lyon 2)
BOUATTOUR	Sonia	Doctorante (co-tutelle Tunisie)
BOUSSAID	Omar	PR (Lyon 2)
DARMONT	Jérôme	PR (Lyon 2)
FAVRE	Cécile	MCF (Lyon 2)
GAVIN	Gérald	MCF (Lyon 2)
GRABOVA	Oksana	Doctorante (co-tutelle Ukraine) - Co-direction SID-DMD
HACHICHA	Marouane	Doctorant (ATER)
HARBI	Nouria	MCF (Lyon 2)
KABACHI	Nadia	MCF (Lyon 1)
KHEMIRI	Rym	Doctorante (co-tutelle Tunisie)
KIT	Chantola	Post-doctorante (Lyon 2)
LOUDCHER	Sabine	MCF HDR (Lyon 2)
NGUYEN	Huu-Hoa	Doctorant (Bourse Vietnam)
SELMANE	Sid Ali	Doctorant (Bourse Algérie)
TRIKI	Salah	Doctorant (co-tutelle Tunisie)
YOUNSI	Fatima-Zohra	Doctorante (co-tutelle Algérie)

Tableau 3 : Membres de l'équipe SID au 01/07/2012

1.2 Thématique et objectifs scientifiques

Les entrepôts de données répondent à un fort besoin en matière d'accès à une information résumée. Cependant, en suivant le processus classique d'entreposage et d'analyse en ligne (OLAP) de données, les systèmes d'information décisionnels exploitent très peu le contenu informationnel des données. En effet, avec l'avènement des données complexes (par exemple composées de textes, d'images, de son ou de vidéo), l'analyse en ligne doit s'adapter à la nature spécifique de ces données tout en gardant l'esprit de l'OLAP. Les opérateurs OLAP sont définis pour des données classiques et sont souvent inadaptés quand il s'agit de données complexes.

Les données complexes véhiculent souvent plus de sémantique que les données classiques, sémantique qu'il est nécessaire de prendre en compte dans leur modélisation et leur analyse en ligne.

Les activités de recherche de l'équipe SID visent deux objectifs primordiaux :

1. proposer de nouveaux modèles d'entrepôts de données et d'analyse en ligne pour répondre aux nouveaux défis liés à l'avènement des données complexes, à la prise en compte de l'utilisateur et à la sécurité des données et des résultats ;
2. concevoir et développer des architectures de systèmes d'information décisionnels dans différents domaines d'application comme la linguistique, l'histoire, la médecine, etc.

La spécificité de nos recherches réside dans la combinaison de l'OLAP avec plusieurs méthodes de traitement de données comme la fouille de données, la statistique et la recherche d'information, afin de franchir le fossé sémantique qui existe entre des données complexes riches en information et des analyses OLAP simples.

Le travail de recherche de l'équipe SID couvre un large champ de méthodologies pour apporter des solutions à des problèmes liés aux quatre thèmes principaux suivants :

1. intégration et représentation des objets complexes,
2. analyse en ligne (OLAP) de données complexes,
3. entrepôts centrés utilisateur,
4. sécurité et qualité des données.

1.3. Contributions majeures

Intégration et représentation des objets complexes

L'intégration des données complexes dans un processus d'entreposage a remis en cause le caractère mécanique du processus d'ETL (Extract-Transform-Load).

Ce dernier, devenant par conséquent plus complexe, nécessite de nouvelles méthodes pour mieux concevoir et effectuer les différentes tâches du processus d'intégration de données.

Les travaux existants dans la littérature portent plutôt sur la modélisation des flux de tâches et de données.

L'approche que nous avons proposée se focalise sur le pilotage des différentes tâches et s'appuie sur différentes technologies [11-213, 12-202, 12-234].

Nous stockons simultanément les données et des appels à des services Web sur un même support, en l'occurrence un document XML actif servant de repository (entrepôt de contenu).

Un moteur ETL permet d'accomplir les tâches d'extraction, de transformation et de chargement.

L'originalité de cette méthode est que ces tâches s'exécutent à l'évocation de services Web.

La détection des changements au niveau des sources de données est également monitorée à l'aide de services Web. Le traitement des événements, tels que les changements dans les sources de données, à l'aide de technique de fouille de données (découverte de règles d'association), permet d'effectuer les tâches d'intégration en temps réel et de tendre vers l'autonomisation du processus d'ETL.

Analyse en ligne (OLAP) de données complexes

Une des caractéristiques de l'analyse en ligne est de se restreindre à des aspects exploratoires et navigationnels. De plus, les données complexes sont souvent décrites dans des documents XML, pour lesquels les opérateurs OLAP classiques ne sont pas adaptés. Les limites de l'OLAP, ainsi que la spécificité des données complexes, nécessitent une évolution ou une adaptation de l'OLAP.

Dans cette vaste problématique, nous avons travaillé :

1) à prendre en compte la sémantique contenue dans les données complexes lors de leur analyse ;

2) à créer des opérateurs OLAP adaptés aux données complexes et allant au-delà de la simple exploration ou navigation ;

3) à étendre l'OLAP à l'analyse de documents XML.

Pour apporter des premières solutions, nous nous sommes orientés vers une combinaison des principes de l'OLAP, de la fouille de données et de la recherche d'information. Les arbres de régression nous permettent de proposer à l'utilisateur de faire de la prédiction dans un cube et d'avoir ainsi une démarche de type What If Analysis.

Pour pouvoir visualiser (en ligne) l'information contenue dans les données complexes, nous utilisons une méthode factorielle (AFC) et proposons ainsi un nouvel opérateur de visualisation [11-209, 12-207].

Nous avons également étudié l'analyse en ligne de cubes de données XML. Modéliser des données multidimensionnelles au niveau logique et physique à l'aide du langage XML permet en effet de prendre en compte des structures hétérogènes au niveau des hiérarchies de dimensions (hiérarchies multiples, imbriquées et/ou incomplètes, que nous appelons hiérarchies complexes), qui existent en réalité, mais sont difficiles à représenter dans les modèles multidimensionnels et relationnels classiques.

Toutefois, la prise en compte de hiérarchies complexes pose des problèmes d'additivité dans les requêtes OLAP.

C'est pourquoi nous avons proposé une approche XML-OLAP (ou XOLAP) efficace, centrée autour d'un opérateur de forage vers le haut (rollup), qui permet de gérer les problèmes d'additivité automatiquement, au moment du requêtage [10-15, 10-224], contrairement aux approches antérieures qui nécessitaient une coûteuse normalisation a priori du schéma ou des données par un expert.

Entrepôts de données centrés utilisateurs

En s'appuyant sur l'intuition que des connaissances sur le métier, sur les données entreposées, leur usage (requêtes) et leur contexte d'utilisation peuvent contribuer à aider l'utilisateur dans son exploration et sa navigation dans les données, nous avons proposé des solutions originales pour la prise en compte de l'utilisateur dans un système d'information décisionnel (SID).

La prise en compte de l'utilisateur dans le processus d'entreposage de données (de la conception de l'entrepôt à l'analyse en ligne des données) est devenue un enjeu crucial. En effet, alors même que les SID sont censés être centrés utilisateur, l'OLAP classique ne dispose pas d'outils permettant de guider l'utilisateur vers les faits les plus pertinents d'un cube de données.

Pour remettre l'utilisateur au cœur d'un SID, nous avons apporté des solutions selon trois axes : **modélisation, évolution de modèle et aide à l'analyse**.

Du point de vue de la modélisation, il s'agit de représenter la réalité des données au travers de hiérarchies de mesures et de dimensions contextuelles, grâce aux connaissances de domaine des experts.

Nous avons introduit pour cela le concept de satellite [10-206, 12-240].

Parallèlement, pour envisager davantage de richesse dans l'évolution des possibilités d'analyse, nous avons proposé une approche OLAP centrée utilisateur qui permet d'étendre les hiérarchies de dimension en créant de nouveaux niveaux d'analyse [09-227].

Cela passe par l'intégration de connaissances dans l'entrepôt, qui peuvent être exprimées par l'utilisateur lui-même ou extraites à partir de données grâce à un couplage OLAP/data mining (avec un paramétrage opéré par l'utilisateur lui-même dans ce cas).

Enfin, afin d'aider l'utilisateur dans sa démarche d'exploration des données pour découvrir des informations cachées et potentiellement pertinentes pour lui, nous avons défini un nouvel opérateur d'agrégation et de structuration, RoK (Roll-up with K-means), qui facilite la découverte d'un bon regroupement des instances d'un niveau d'analyse existant choisi par l'utilisateur à partir duquel un nouveau niveau d'analyse peut être créé [09-38].

Ce nouveau niveau de hiérarchie permet de créer de nouvelles requêtes décisionnelles que le SID pourra recommander à l'utilisateur.

Sécurité et qualité des données

Les entrepôts de données visent à avoir une vue commune de l'ensemble des données du système opérationnel, permettant ainsi la prise de décision.

Cependant, cette approche crée un conflit. D'une part, les entrepôts de données doivent permettre un accès facile aux données et, d'autre part, les organisations doivent s'assurer que leurs données ne sont pas divulguées sans contrôle. La sécurisation des entrepôts de données peut alors être abordée à deux niveaux : **au niveau conceptuel**, pour concevoir un entrepôt de données sécurisé ; et **au niveau exploitation**, afin de renforcer les droits d'accès/habilitations des utilisateurs, et à interdire tout utilisateur mal intentionné d'inférer des données interdites à partir des données auxquelles il a accès.

Dans ce cadre, nous avons travaillé à trois niveaux : **sécurité, intégrité et confidentialité des données**.

Du point de vue de la **sécurité des données**, nous avons travaillé à améliorer la détection des attaques en utilisant des techniques de fouille de données dans les systèmes de détection d'intrusions. Il s'est agit de trouver la meilleure combinaison de classifieurs pour caractériser les attaques en temps réel [11-228, 11-229, 12-238].

Afin de garantir **l'intégrité** (cohérence, disponibilité, fraicheur...) des données dans les systèmes d'information décisionnels, nous avons travaillé au niveau du processus d'ETL pour traiter le problème avant leur stockage dans un entrepôt et leur utilisation pour des analyses.

Nous avons étudié cette problématique et apporté des réponses dans le cadre des entrepôts classiques [10-229, 11-225] et nous l'étudions actuellement dans le cas où les entrepôts sont stockés dans les nuages.

Enfin, pour assurer **la confidentialité des données sensibles** et faire face aux exigences de sécurité des données multidimensionnelles destinées à l'analyse, nous avons défini les autorisations et les restrictions d'accès aux données sous formes de profils qui rassemblent les rôles et les responsabilités des utilisateurs [12-201].

Lorsque les données sont stockées dans les nuages, nous avons proposé une première approche n'accordant pas de confiance aux fournisseurs de services.

Il s'agit de stocker les données chez plusieurs fournisseurs à l'aide l'algorithme de clés secrètes de Shamir, ce qui permet d'assurer à la fois la sécurité des données vis-à-vis des fournisseurs et de minimiser le risque de non disponibilité des données [12-205].

Finalement, des résultats importants de cryptographie ont montré que tout calcul effectué sur des données distribuées peut se faire en garantissant leur confidentialité. Néanmoins, ces résultats généraux n'ont qu'un intérêt théorique et beaucoup d'efforts restent à faire pour obtenir des protocoles efficaces pour des applications particulières.

Nous nous sommes plus particulièrement intéressés aux aspects d'intégration d'entrepôts de données distribués ainsi qu'à la préservation de la confidentialité des requêtes [10-214, 11-251] .

1.4. Production scientifique

1.4.1. Publications

Nous avons constitué une liste de revues et conférences (internationales et nationales) sélectives et/ou significatives vis-à-vis de nos thématiques de recherche[16].

Elle couvre le domaine général des bases de données, ce qui permet de cibler les revues et conférences de meilleur rang, mais aussi le sous-domaine plus spécifique et plus récent des entrepôts de données, dans la communauté duquel nous nous situons. Les membres de notre équipe sont fortement encouragés à publier principalement dans cette liste.

Les Figures 13 et 14 comparent la production scientifique de l'équipe SID pendant les périodes 2005-2008 et 2009-2012, en termes de type de publication et de rang ERA des publications internationales, respectivement.

La Figure 13, souligne l'effort effectué sur les publications internationales, dont le volume a augmenté de 73 % (le nombre de revues internationales a, notamment, plus que doublé), tandis que le volume de publications nationales a baissé de 26 % du fait de cet effort.

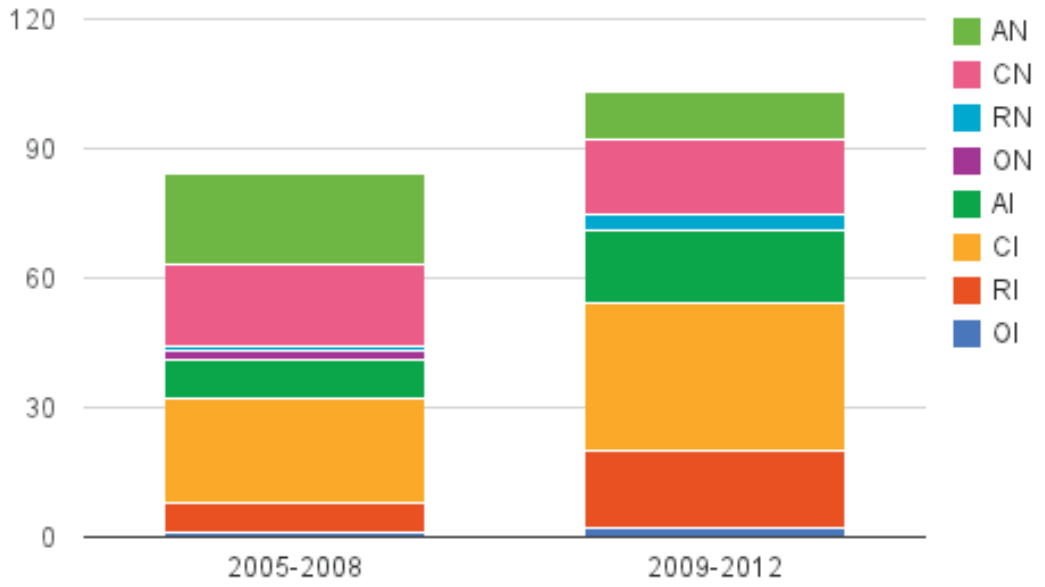

Figure 13 : Évolution de la production scientifique de SID par type de publication

Codification employée

O - Ouvrages et direction d'ouvrages	R - Revues	I - Portée internationale
C - Conférences avec comité de lecture et actes	A - Autres publications	N - Portée nationale

Ces statistiques sont à l'image des efforts de tous les chercheurs de l'équipe SID pour une recherche de qualité.

Concernant la publication dans des conférences et des revues internationales sélectives, l'équipe SID a suivi les recommandations du comité d'évaluation de 2009, en privilégiant la qualité et non la quantité, même si des efforts restent à faire pour publier dans les revues et conférences de rang A (Figure 14).

16 http://eric.univ-lyon2.fr/programme_de_recherche/axe_ena_dc/sdoc-24-cibles-publis-enadc.xlsx

un effort important sur les publications internationales,

dont le volume a augmenté de 68 %

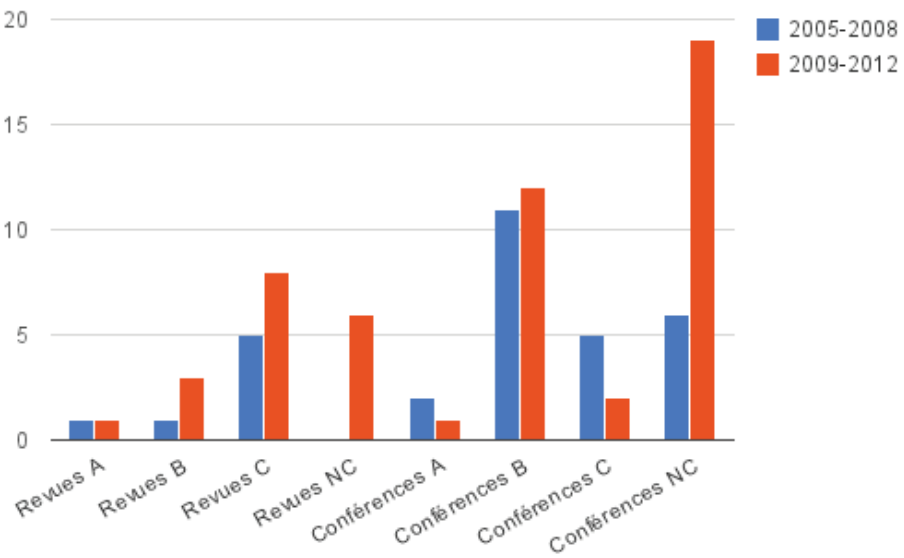

Figure 14 : Évolution de la production scientifique internationale de SID par rang de publication

La Figure 15 présente une synthèse chiffrée de la production de l'équipe SID par membre permanent sur la période 2009-2012. La moyenne est d'environ 12 publications par membre.

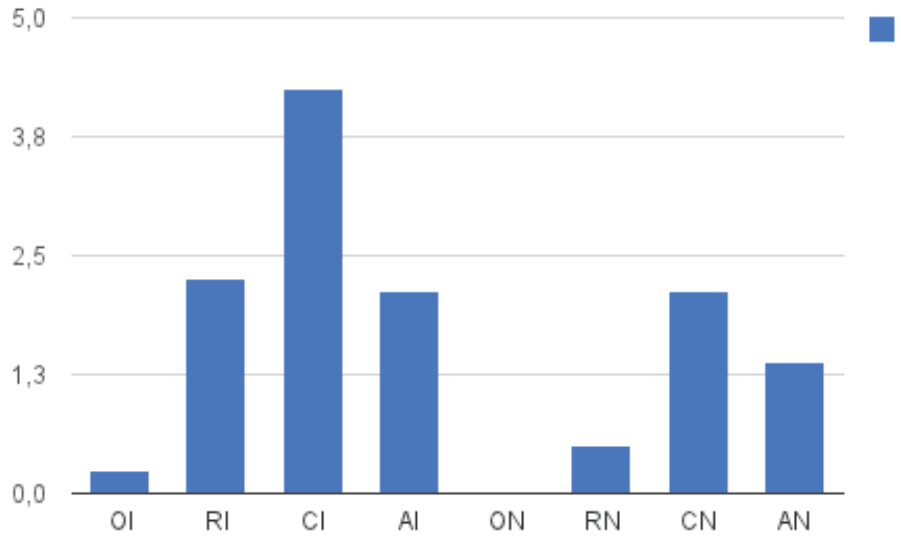

Figure 15 : Synthèse de la production scientifique par membre permanent de SID

Une moyenne de 12 publications par membre

Codification employée

O - Ouvrages et direction d'ouvrages
C - Conférences avec comité de lecture et actes
R - Revues
A - Autres publications
I - Portée internationale
N - Portée nationale

des collaborations internationales riches

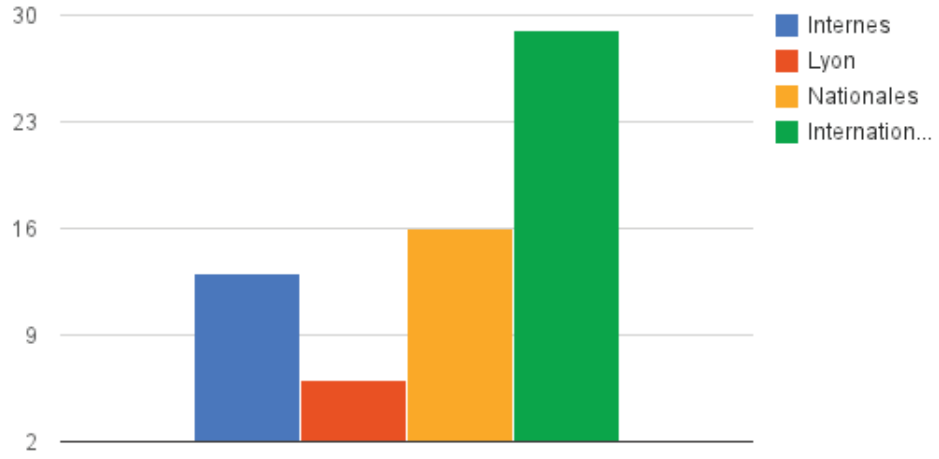

Figure 16 : Publications en collaboration de l'équipe SID

Enfin, les membres de l'équipe cosignent des publications avec de nombreux autres chercheurs (Figure 16), notamment au niveau international grâce à des bonnes relations avec les collègues de la communauté DOLAP/DaWaK (les deux conférences phares du domaine des entrepôts de données), et de nombreux co-encadrements de thèses (cotutelles, partenariats).

1.4.2. Thèses et HDR

Nom	Prénom	Soutenance	Encadrant	Devenir
MAIZ	Zora	2010	F. Bentayeb, O. Boussaïd	Recherche emploi suite à deux congés maternité
SALEM	Rashed	2012	O. Boussaïd, J. Darmont	Enseignant-chercheur (Egypte)

Tableau 4 : Thèses soutenues au sein de l'équipe SID

Nom	Prénom	Soutenance	Coordonnateur
BENTAYEB	Fadila	2011	D.A. Zighed
LOUDCHER	Sabine	2011	O. Boussaïd

Tableau 5 : HDR soutenues au sein de l'équipe SID

1.5. Animation, vie de l'équipe

1.5.1. Réunions et séminaires

Notre équipe se réunit chaque vendredi après-midi à raison de deux à trois heures.

Nous consacrons un tiers du temps à l'organisation, aux aspects administratifs et logistiques de l'équipe. Le reste du temps est consacré à des exposés, des discussions et des échanges scientifiques.

1.5.2. Organisation de journées thématiques

Journée sur le traitement automatique des données linguistiques (Lyon 2, 2011)

Journée sur le décisionnel dans les nuages (Lyon 2, 2012)

1.5.3. Stagiaires de recherche

Chaque année, nous accueillons dans notre équipe des étudiants pour un stage de recherche. Ces étudiants peuvent provenir de nos formations de Master recherche ou professionnel ou aussi de l'étranger. Nous accueillons également plusieurs chercheurs étrangers pour des séjours scientifiques plus ou moins longs (de 1 à 18 mois).

1.6. Partenariats, projets

Nom	Partenaires	Années	Financement	Financeur	Implication
DPCCLO [1]	ICAR	2009-2010	20 000 €	Lyon 2 (BQR)	Porteur
ProxAn [2]	CREALYS	2008-2009	30 000 €	Région Rhône-Alpes	Porteur
Qualité [3]	AID	2009-2012	12 000 €	AID	Thèse CIFRE
Tassili	Université de Blida (Algérie)	2011-2014	39 600 €	EGIDE	Co-porteur

Tableau 6 : Projets financés menés par l'équipe SID

[1] DPCCLO : Détection de Phénomènes Complexes dans les Corpus Linguistiques Oraux
[2] ProxAn : Proximity Analysis
[3] Analyse et amélioration de la qualité des données dans un environnement multi-sources

Des membres de l'équipe SID sont également impliqués dans les projets DocNet, Web et ImagiWeb portés par l'équipe DMD (Section 2. Partie 2.6).

1.7. Visibilité nationale et internationale

1.7.1. Positionnement sur la scène nationale et internationale

Dans la thématique des entrepôts de données et de l'analyse en ligne, notre équipe est bien placée parmi les laboratoires nationaux (IRIT-Toulouse, LI-Tours, LIRMM-Montpellier) qui travaillent sur les mêmes thèmes, notamment autour de la modélisation et l'analyse multidimensionnelle de données complexes, de la personnalisation dans les entrepôts de données et de la recommandation de requêtes décisionnelles.

Notre équipe est fondatrice en 2005 et leader de la conférence francophone EDA (Entrepôts de Données et Analyse en ligne) et, depuis quelques années, nous organisons et animons des ateliers internationaux dans le domaine des entrepôts de données dans des conférences reconnues : International Workshop on Warehousing and Mining Complex Data (WMCD@EDBT 2011), User-Centric Information System for Decision Systems (UC4DS@ADBIS 2012) et International Workshop on Cloud Intelligence (Cloud-I@VLDB 2012). L'équipe SID a établi des relations internationales de qualité (R. Missaoui, L. Gruenwald, T. Pedersen, D. Lemire, A. Cuzzocrea, I. Zolotariova, F. Dewan).

Notre attractivité est importante si nous la mesurons en termes de professeurs invités, d'étudiants intéressés pour poursuivre leurs études dans nos thématiques, de postdocs ou de chercheurs étrangers venant pour des séjours scientifiques.

1.7.2. Collaborations

Internationales

- Algérie : Co-encadrement de thèses et accueil de chercheurs permanents et doctorants
    - Université Saad Dahleb de Blida (projet Tassili)
    - Université des Sciences et Technologie d'Oran
    - Ecole Nationale Supérieure d'Informatique d'Alger
    - Université Supérieure Houari Boumediene, Alger
    - Université de Laghouat
- Allemagne
    - Hasso Plattner Institute : Publication commune
    - Universität Münster : Publication commune
- Bangladesh
    - Jahangirnagar University : Publication d'articles communs
- Canada
    - Université du Québec en Outaouais : Séjour scientifique, thèse en co-encadrement et accueil de chercheurs, publications communes
    - Université du Québec à Montréal : Accueil de stagiaires recherche
- Danemark
    - Aalborg University : Co-organisation d'un atelier international
- Espagne
    - Universitat Politècnica de Catalunya, Barcelona : Publication commune
    - Universitat de Alicante : Publication commune

- Grèce
    - University of Ioannina : Publication commune
- Italie
    - CNR Calabre : Publication d'articles communs
    - Università di Bologna : Publication commune
- Madagascar
    - Université de Fianarantsoa : Publication d'articles communs
- Maroc
    - Ecole Nationale des Sciences Appliquées, Agadir : Co-encadrement de thèse et accueil de chercheurs
- Tunisie : Thèses en cotutelle
    - Université de Sfax
    - Université de Tunis
- Ukraine
    - Université Nationale d'Economie de Kharkov : Thèse en cotutelle
- Uruguay
    - Universidad de la Republica, Montevideo : Publication commune
- USA
    - Université d'Oklahoma : Accueil régulier de stagiaires recherche de niveau Master

Nationales et locales

- ETIS (Cergy-Pontoise) : projet ANR en cours d'élaboration
- ICAR (ENS-Lyon 2, linguistique) : projet DPCCLO
- IMAG (Grenoble) : projet ANR en cours d'élaboration
- IRIT (Toulouse) : publications, projet ANR en cours d'élaboration
- IRSTEA (Clermont-Ferrand) : publications, projet ANR en cours d'élaboration
- LARHRA (Lyon 2, histoire) : projet ANR non sélectionné
- LI (Tours) : projet ANR non sélectionné
- LIMOS (Clermont-Ferrand) : publications, projets ANR en cours d'élaboration
- LIPADE (Paris 6) : projet ANR en cours d'élaboration
- LIRIS (Lyon 1 et INSA Lyon) : co-organisation de la conférence INFORSID 2014
- LIRMM (Montpellier) : publications, projet ANR en cours d'élaboration
- LRI (Paris Sud) : projet ANR en cours d'élaboration
- Centre de recherche Magellan, équipe MODEME (Lyon 3) : co-organisation de la conférence INFORSID 2014

Industrielles

- AID
- Alcatel Lucent Bell Labs
- AMI Software
- Buzzinbees

1.7.3. Coencadrement de thèses

Nom	Prénom	Type thèse	Etablissement
BALA	Mahfoud	Bourse Algérie	Université Alger
BOUAKKAZ	Moustapha	Bourse Algérie	Université Laghouat, Algérie
BOUKRAA	Doulkifli	Bourse Algérie	Université de Jijel, Algérie
DERRAR	Hacene	Bourse Algérie	Université d'Alger
HANACHI	Lilia	Bourse Algérie	Université Saad Dahleb, Blida, Algérie
MEDDAH	Maamar Yacine	Bourse Algérie	USTO, Algerie
SAIR	Abdellah	Bourse Maroc	ENSA Agadir, Maroc

Tableau 7 : Thèses coencadrées par des membres de l'équipe SID

1.8. Dix principales publications

- [12-234] R. Salem, J. Darmont, O. Boussaid, «Active XML-based Web Data Integration», Information Systems Frontiers, 2013 (Special issue on Business Intelligence and the Web; à paraître).
- [12-230] D. Boukraâ, O. Boussaid, F. Bentayeb, "Complex Object-Based Multidimensional Modeling and Cube Construction", Fundamenta Informaticae Journal. À paraître.
- [11-221] M. Hachicha, J. Darmont, «A Survey of XML Tree Patterns», IEEE Transactions on Knowledge and Data Engineering, 2012 (in preprint).
- [12-207] S. Loudcher, O. Boussaïd. OLAP on Complex Data: Visualization Operator Based on Correspondance Analysis. Selected extended paper of CAISE Forum 2011. Lecture Notes in Business Information Processing series (LNBIP). Vol. 107. Pages 172-185. Springer, 2012.
- [09-3] K. Aouiche, J. Darmont, «Data Mining-based Materialized View and Index Selection in Data Warehouses», Journal of Intelligent Information Systems, Vol. 33, No. 1, 2009, 65-93.
- [12-240] Y. Pitarch, C. Favre, A. Laurent, P. Poncelet, Enhancing Flexibility and Expressivity of Contextual Hierarchies, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2012) (to appear).
- [12-238] Bahri E., Harbi N. and Nguyen Huu H., "A Multiple Classifier System Using an Adaptive Strategy for Intrusion Detection", ICICS'2012 (International Conference on Intelligent Computational System), 7-9 janvier 2012, Dubai (Emirats Arabes Unis).
- [11-207] D. Boukraâ, O. Boussaid, F. Bentayeb, Vertical Fragmentation of XML Data Warehouses Using Frequent Path Sets, DAWAK, August, 2011, Toulouse, France (DAWAK 2011), 196-207.
- [11-229] H.H. Nguyen, N. Harbi, J. Darmont, «An Efficient Fuzzy Clustering-Based Approach for Intrusion Detection», 15th East-European Conference on Advances and Databases and Information Systems (ADBIS 11), Vienna, Austria, September 2011; Research Communications, Austrian Computer Society, Vienna, Austria, 117-127.
- [09-38] F. Bentayeb, C. Favre, RoK: Roll-Up with the K-Means Clustering Method for Recommending OLAP Queries, 20th International Conference on Database and Expert Systems Applications (DEXA 09), Linz, Austria, September 2009 ; LNCS, Vol. 5690, 501-515.

2 Équipe DMD

2.1. Membres de l'équipe

Responsable : Julien VELCIN

Nom	Prénom	Statut
AAZI	Fatima Zahra	Doctorante (cotutelle Maroc)
ABDESSELAM	Rafik	MCF HDR (Lyon 2)
AH-PINE	Julien	MCF (Lyon 2)
BONNEVAY	Stéphane	MCF HDR (Lyon 1)
BOUNEKKAR	Ahmed	MCF (Lyon 1)
CHAUCHAT	Jean-Hugues	PR émérite (Lyon 2)
DERMOUCHE	Mohamed	Doctorant (CIFRE) Co-direction SID-DMD
DESROZIERS	Katia	Doctorante (CIFRE)
EZZEDINE	Diala	Doctorante (bourse Liban)
FORESTIER	Mathilde	Doctorante (ATER)
FRENKIEL	Jérôme	Doctorant (salarié)
GUILHAUME	Chantal	Doctorante (salariée)
GUILLE	Adrien	Doctorant (CDU) Co-direction SID-DMD
KAFAFY	Ahmed	Doctorant (bourse Egypte)
JOUVE	Bertrand	PR (Lyon 2)
LALLICH	Stéphane	PR (Lyon 2)
LUST	Thibaut	Postdoctorant (projet européen)
OROBINSKA	Olena	Doctorante (cotutelle Ukraine)
RAKOTOMALALA	Ricco	MCF (Lyon 2)
RICO	Agnès	MCF (Lyon 1)
RICO	Fabien	MCF (Lyon 1)
RIZOIU	Marian-Andrei	Doctorant (CDU)
ROLLAND	Antoine	MCF (Lyon 2)
SIANI	Carole	MCF HDR (Lyon 1)
VELCIN	Julien	MCF (Lyon 2)
ZIGHED	Djamel Abdelkhader	PR (Lyon 2)

Tableau 8 : Membres de l'équipe DMD au 01/07/2012

2.2. Thématique et objectifs scientifiques

L'objectif de l'équipe DMD est de concevoir de nouveaux systèmes, modèles et algorithmes pour la fouille de données complexes et l'aide à la décision. Les données complexes sont des données structurées (par exemple sous forme de graphes), hétérogènes (descriptions attributs-valeurs, textes, images, etc.), dynamiques (qui évoluent au fil du temps), imprécises, volumineuses. Pour manipuler ces données, l'équipe s'appuie sur des approches principalement statistiques (analyse des données, inférence) et inspirées de l'intelligence artificielle : apprentissage automatique, représentations floues, raisonnement dans l'incertain, etc.

L'activité de l'équipe peut se découper en quatre axes thématiques, même si les chercheurs travaillent dans ces axes sont souvent amenés à collaborer :

- apprentissage automatique pour la fouille de données ;
- modélisation, caractérisation, fouille dans les graphes ;
- modèles d'aide à la décision multicritère ;
- analyse des données complexes et fouille d'opinions.

Ces quatre axes sont détaillés dans la suite de ce document.

2.3. Contributions majeures

Apprentissage automatique pour la fouille de données

Une partie des travaux de l'équipe porte sur le développement de nouvelles techniques d'apprentissage automatique (machine learning).

Tout d'abord, des nouvelles méthodes d'ensemble, méthodes qui consistent à agréger plusieurs classifieurs afin d'améliorer les résultats de l'apprentissage supervisé, ont été proposés.

En particulier, l'équipe a réalisé des contributions théoriques sur des ensembles composés de forêts aléatoires [10-210].

Dans une autre optique, nous nous sommes plutôt intéressés à l'extraction de règles d'association de classe et à leur évaluation à l'aide de mesures d'intérêt.

En collaboration avec Telecom Bretagne, nous avons identifié et généralisé un certain nombre de propriétés d'antimonotonie et nous avons ensuite établi des conditions nécessaires et/ou suffisantes pour qu'une mesure d'intérêt possède ces propriétés [12-222].

L'équipe a également étudié le problème de la classification non supervisée dans son ensemble : critères de classification, mesures de similarités, modélisation sous forme de problèmes d'optimisation et algorithmes qui passent à l'échelle.

Un algorithme de complexité linéaire et ne dépendant pas de l'ordre des individus a été proposé [09-220].

D'autre travaux ont concerné les modèles d'apprentissage non supervisé dédiés à l'analyse de données textuelles (topic models).

En particulier, l'équipe a proposé des algorithmes d'étiquetage des catégories thématiques (clusters) et d'évaluation de la qualité des thématiques obtenues. Réalisé en collaboration avec des chercheurs de l'Ecole Polytechnique de Bucarest, l'algorithme d'évaluation permet d'émuler le jugement humain en se basant sur une recherche de correspondance (mapping) entre les thématiques et une base de connaissance lexicale [11-214].

Notons enfin des travaux sur l'apprentissage automatique de la structure de réseaux bayésiens lorsque le nombre de variable est très grand.

L'idée consiste à réaliser l'apprentissage sur des sous-ensembles de variables de taille raisonnable, puis de recombiner les différents résultats [09-60].

Pour finir, des méthodes numériques basés sur des algorithmes de type réseaux de neurones ont été appliqués pour la modélisation économique et financière [09-41].

Modélisation, caractérisation, fouille dans les graphes

Dans ce deuxième axe, les chercheurs de l'équipe ont travaillé sur l'analyse de graphes, avec des contributions sur les voisinages dans les graphes, sur la recherche d'information dans ces graphes, et sur la caractérisation de famille de graphes.

En particulier, l'équipe a récemment développé une approche d'apprentissage qui utilise des graphes de voisinage pour la comparaison, le regroupement et l'équivalence topologique de mesures de proximité [12-232].

D'autres travaux ont permis d'étudier les propriétés structurelles des graphes.

Ainsi, l'équipe a étudié la décomposition de graphes (graphes orientés et 2-structures) en 2-clans, qui sont des modules (ou clans) à deux éléments, ainsi que leurs caractérisation par l'utilisation d'homotopies [09-229].

Des travaux récents ont également été menés sur la recherche d'information au sein de graphes, dans une approche de fouille du Web (web mining).

L'approche a consisté à prendre en compte à la fois la structure du graphe (degrés, composantes connexes, etc.) et le contenu textuel des noeuds.

Nous avons montré qu'il était possible d'extraire le réseau social sous-jacent à des discussions en ligne [11-203], d'en extraire les messages les plus intéressants [09-57] et les acteurs jouant des rôles clefs [11-218].

Enfin, un algorithme d'extraction des communautés, basé sur une approche prétopologique, a été proposé et testé efficacement sur des données bibliographiques, en collaboration avec le laboratoire Hubert Curien de Saint-Etienne [09-202].

Modèles pour l'aide à la décision

Les membres de l'équipe qui travaillent sur cet axe de recherche se sont intéressés aux méthodes et aux modèles d'aide à la prise de décision collective, multicritères et multi-objectifs.

Des recherches théoriques ont été menées afin d'étudier les propriétés de nouvelles méthodes dans le cadre de l'analyse multicritères : propriétés des intégrales de Sugeno [09-213], modèles à base de points de référence, modèles à base de bicapacités [12-214].

D'autres travaux ont été plus particulièrement axés sur la théorie de l'évidence et des possibilités [11-219], mais aussi sur l'étude des liens et interactions possibles entre l'aide à la décision multicritères et les autres champs de la théorie des processus décisionnels et de l'apprentissage supervisé [10-239], ou encore sur l'agrégation par programmation linéaire [10-216].

Des applications ont été développées en collaboration avec d'autres partenaires : utilisation de l'intégrale de Choquet pour la reconnaissance d'image en collaboration avec le LIRMM [09-215], aide à la sélection de fruits en collaboration avec l'INRA d'Avignon, adaptation de méthodes d'analyse multicritère en santé.

Du point de vue de la prise de décision dans le cadre d'une optimisation multiobjectif, les travaux développés proposent un nouvel algorithme évolutionnaire hybride HEMH combinant différentes techniques (DM-GRASP, Path-Relinking et recherche locale) pour résoudre le problème du sac à dos multi-objectif [12-221].

Des recherches ont été menées spécifiquement sur l'intégration de la prise en compte du décideur pour aider au choix des bonnes solutions.

Analyse des données complexes et fouilles d'opinions

En suivant cet axe, l'équipe privilégie l'élaboration de nouvelles techniques d'analyse des données et leur utilisation sur des données complexes : données temporelles/évolutives, textuelles, mixtes (variables quantitatives et qualitatives).

De nouvelles méthodes ont été proposées : analyse des associations dissymétriques dans le cadre de l'analyse des correspondances et de l'analyse de la variance, traitement des données mixtes (quantitatives et qualitatives) dans le contexte de l'analyse en composantes principales et enfin, des techniques de classement et de prédiction par l'analyse discriminante sur données évolutives, sur données mixtes ou encore à plusieurs variables cibles-groupes [10-5].

Récemment, l'équipe s'est attaquée au cas des données textuelles complexes car elles contiennent des opinions parfois difficiles à déceler, elles sont produites par des acteurs eux-mêmes inscrits dans un réseau de relations, et elles évoluent au fil du temps.

Ces données posent de nombreux problèmes d'analyse descriptive et de visualisation, d'une part, et de validation (comment découper un corpus ainsi structuré en échantillons d'apprentissage et de validation ?) d'autre part.

Sur ces thèmes, l'équipe a notamment proposé une nouvelle visualisation pour le suivi de données textuelles évolutives, en collaboration avec l'IRISA (Rennes) et l'Université de Zagreb [09-12]

2.4. Production scientifique

2.4.1. Publications

Les Figures 17 et 18 comparent la production scientifique de l'équipe DMD pendant les périodes 2005-2008 et 2009-2012, en termes de type de publication8 et de rang ERA des publications internationales, respectivement.

La Figure 17 montre une augmentation de 44 % du nombre de publications internationales (le nombre des ouvrages et articles de revue a notamment doublé).

D'un point de vue qualitatif, la Figure 18 montre clairement un effort de l'équipe DMD ces dernières années pour publier dans des revues et des conférences à fort taux de sélection reconnues internationalement (Computational Intelligence, IDA, ECML-PKDD, IJCAI, etc.). Dans le même temps, le taux de global de publications non classées a chuté, ce qui est surtout le cas des publications dans des conférences non ou peu reconnues internationalement.

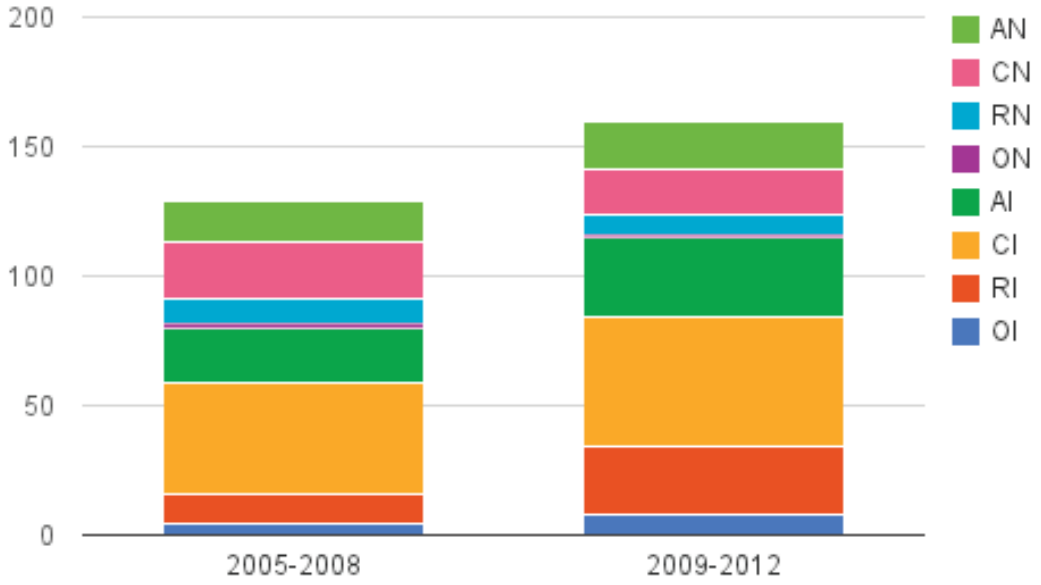

Figure 17 : Évolution de la production scientifique de DMD par type de publication

Codification employée

O - Ouvrages et direction d'ouvrages
C - Conférences avec comité de lecture et actes
R - Revues
A - Autres publications
I - Portée internationale
N - Portée nationale

un effort pour publier dans des revues et des conférences à fort taux de sélection

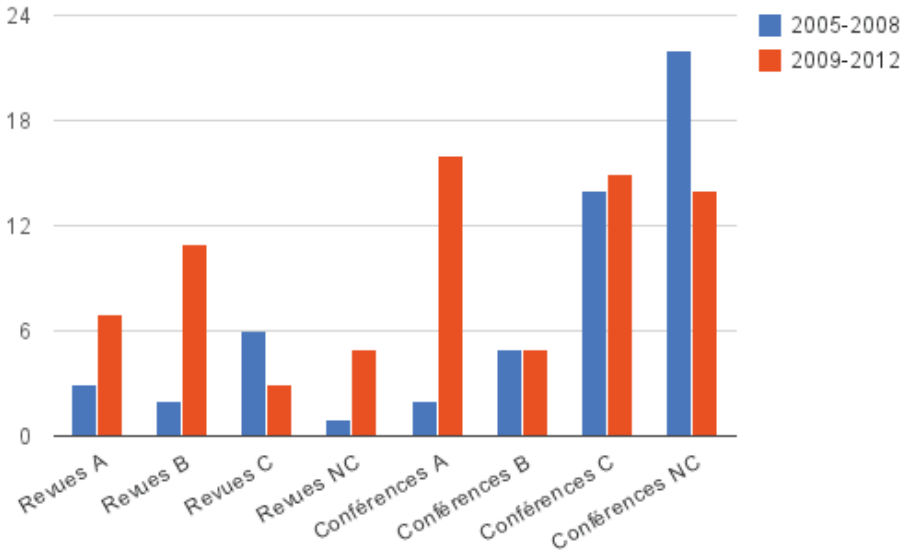

Figure 18 : Évolution de la production scientifique internationale de DMD par rang de publication

La Figure 19 présente une synthèse chiffrée de la production de l'équipe DMD par membre permanent sur la période 2009-2012. La moyenne est d'environ 11 publications par membre.

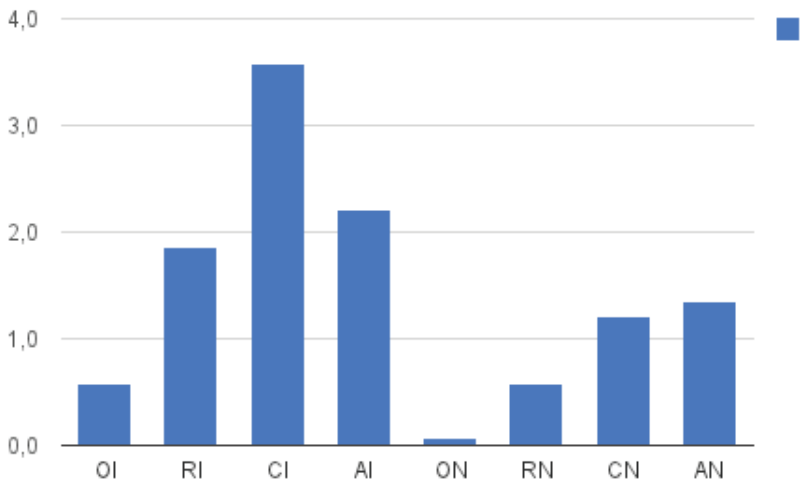

Figure 19 : Synthèse de la production scientifique par membre permanent de DMD

Une moyenne de 11 publications par membre

Codification employée

O - Ouvrages et direction d'ouvrages
C - Conférences avec comité de lecture et actes
R - Revues
A - Autres publications
I - Portée internationale
N - Portée nationale

Enfin, les membres de l'équipe DMD cosignent des publications avec de nombreux autres chercheurs, notamment au niveau national (Figure 20).

Le détail des collaborations est donné dans la partie 2.7.2 ci-dessous.

des collaborations nationales riches

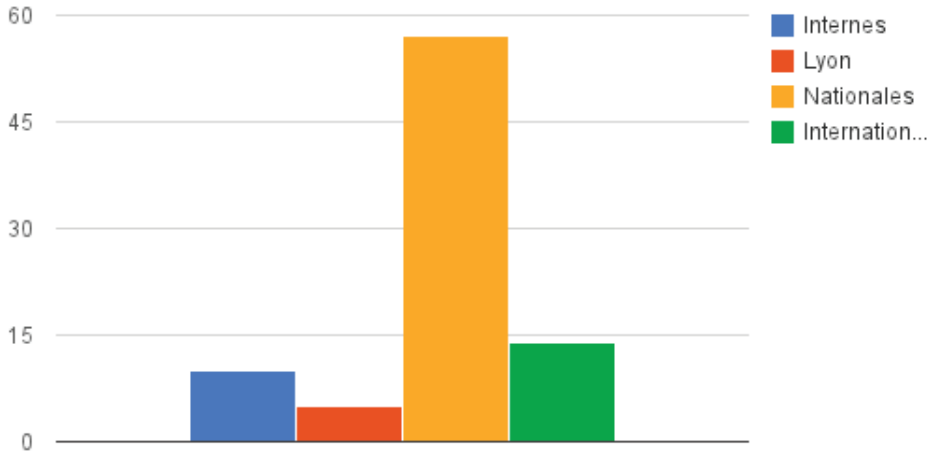

Figure 20 : Publications en collaboration de l'équipe DMD

2.4.2. Thèses

Nom	Prénom	Soutenance	Encadrant	Devenir
PRUDHOMME	Elie	2009	S. Lallich	Ingénieur R&D
THOMAS	Julien	2009	D. A. Zighed	Ingénieur d'étude
BAHRI	Emna	2010	S. Lallich	Ingénieur d'étude
STAVRIANOU	Anna	2010	J. H. Chauchat, J. Velcin	Ingénieur de recherche
QHRESHI	Taimur	2010	D. A. Zighed	Enseignant-chercheur (Pakistan)
WEI	Zhihua	2010	J. H. Chauchat	Enseignant-chercheur (Chine)
MAVRIKAS	Efhtymios	2010	D. A. Zighed	Chef d'entreprise (Grèce)
PISETTA	Vincent	2012	D. A. Zighed	Ingénieur

Tableau 9 : Thèses soutenues dans l'équipe DMD

2.5. Animation, vie de l'équipe

2.5.1. Réunions et séminaires

Notre équipe se réunit au rythme d'une fois tous les mois ou tous les mois et demi, à raison de deux à trois heures. Nous consacrons un tiers du temps à l'organisation, aux aspects administratifs et logistiques de l'équipe. Le reste du temps est consacré à des exposés, des discussions et des échanges scientifiques.

Nous organisons régulièrement des journées (ou demi-journées) thématiques afin de faciliter la communication à la fois à l'intérieur de l'équipe, mais également au sein du laboratoire et en invitant des chercheurs de laboratoires du PRES de Lyon.

2.5.2. Organisation de journées thématiques

- Journée sur la "Modélisation et fouille de données historiques" (mars 2009, 8 présentations, 36 participants).
- 7ème journées "Prétopologie et Modélisation" (mai 2010).
- Demi-journée sur la "Fouille de données dans les réseaux sociaux et sur le Web" (février 2012, 14 participants).
- Demi-journée sur la "Fouille de textes et fouille d'opinions" (24 février 2012, 20 participants).

2.5.3. Stagiaires de recherche

Chaque année, nous accueillons dans notre équipe des étudiants pour un stage de recherche. Ces étudiants peuvent provenir de nos formations de Master recherche ou professionnel, mais peuvent également venir de l'étranger. Nous accueillons également régulièrement des chercheurs étrangers pour des séjours scientifiques plus ou moins longs (de 1 mois à 18 mois).

2.6. Partenariats, projets

Nom	Partenaires	Années	Financement	Financeur	Implication
BETWEEN [1]		2007-2009	20 500 €	Région Rhône-Alpes	Responsable scientifique
NOFDSHS [2]	ISH	2008-2010	25 000 €	Lyon 2 (BQR)	Porteur
Dynamic	Visioglobe	2010	5000 €	Visioglobe	Porteur
RFCDP [3]	ELICO	2010-2011	20 000 €	Lyon 2 (BQR)	Co-porteur
SHS-DOC-NET [4]	ISH	2011-2012	20 000 €	Lyon 2	Porteur
G-Graphs and networks	ESSEC, Univ. Martinique, Univ. Bucarest	2012	2 500 €	GDR RO	Partenaire
Web [5]	AMI Software	2012-2014	27 000 €	AMI Software	Thèse CIFRE
ImagiWeb [6]	XRCE, AMI S., EDF, LIA, CEPEL	2012-2015	840 000 €	ANR CONTINT	Porteur
Réseaux [7]	ARC6	2012	3000 €	Région Rhône-Alpes	Porteur

Tableau 10 : Projets financés menés par l'équipe DMD

[1] Modèle de représentation et d'analyse des débats en ligne sur Internet (incubation d'entreprise innovante).

[2] Nouveaux Outils de Fouille de Données pour les SHS.

[3] Le rôle des forums citoyens dans le débat public. Construction et test d'outils semi-automatiques pour l'étude de la dynamique des discours.

[4] Mise en place d'un portail internet de type Web social sémantique pour l'exposition des compétences et la veille scientifique en SHS.

[5] Modélisation des controverses sur le Web et les médias publics par le contenu de la structure du réseau.

[6] Images sur le Web : analyse de la dynamique des images sur le Web 2.0.

[7] Modèles de grands réseaux : jeux de poursuite et décomposition modulaire. Applications au WWW.

Des membres de l'équipe DMD sont également impliqués dans le projet "Qualité" porté par l'équipe SID (Section 2. Partie 1.6).

2.7. Visibilité nationale et internationale

2.7.1Positionnement sur la scène nationale et internationale

L'équipe DMD est clairement positionnée sur les domaines de la fouille des données (data mining) et de l'aide à la décision, domaines sur lesquels elle a acquis une solide reconnaissance.

Avec plusieurs laboratoires d'Informatique (LINA, LRI, LIRMM, LIRIS, TELECOM-ParisTech), elle fait partie des membres fondateurs de l'association Extraction et Gestion des Connaissances (EGC[17]), qui organise régulièrement la conférence du même nom. Les membres de l'équipe continuent à animer la communauté francophone de fouille de données par le biais de l'association.

L'un des membres de l'équipe est le co-directeur de la Revue des Nouvelles Technologies de l'Information (RNTI), publiée chez Hermann. Dans le domaine de la décision, l'équipe DMD a également animé le groupe SCDD (Systèmes Complexes et Décision Distribuée) associé à la ROADEF et le GDR MACS. Elle a fondé et co-animé l'association PretopologiCS qui a pour but la promotion, la valorisation et la diffusion de la recherche en prétopologie et modélisation des systèmes complexes.

Au niveau international, les membres de l'équipe animent des événements scientifiques internationaux (pilotage des workshops QIMIE@PAKDD[18] et MSND@WWW[19] 2012, organisation des conférences ALT et DS en 2012[20]).

2.7.2. Collaborations

Internationales

- Arabie Saoudite
  - King Saud University : publications en commun
- Canada
  - Université d'Ottawa : collaborations régulières avec le Prof. S. Matwin, séjour scientifique prochainement prévu d'un membre du laboratoire
- Chine
  - Université Tondji (Shangaï) : thèse en cotutelle
- Croatie
  - Université de Zagreb : séjours scientifiques, publications en commun, dépôt de projets EGIDE
- Espagne
  - Université Polytechnique de Catalogne : séjours scientifiques du Prof. T. Aluja et du Prof. M. Becue
- Italie
  - Université du Piémont Oriental : séjours scientifiques du Prof. L. Saitta
- Maroc
  - Université de Casablanca : publications en commun
  - Université Hassan Ier Settat : thèse en cotutelle

17 http://www.egc.asso.fr
18 http://conferences.telecom-bretagne.eu/qimie2011/
19 http://eric.univ-lyon2.fr/msnd/
20 http://eric.univ-lyon2.fr/alt-ds-2012/

- Roumanie
    - Ecole polytechnique de Bucarest : séjours scientifiques du Prof. S. Trausan-Matu, accueil régulier de stagiaires et de doctorants
    - ESSEC : projet GDR RO en commun
- Slovénie
    - Université de Ljubjlana : publications communes
- Tunisie
    - Institut Supérieur de Gestion de Tunis : thèse en cotutelle
    - Université de Sfax : publications communes
- Ukraine
    - Université Nationale d'Economie de Kharkov : Thèse en cotutelle, séjours scientifiques du Prof. I. Zolotariova et d'enseignants-chercheurs
    - Université Polytechnique de Kharkov : Thèse en cotutelle

Nationales

- IRISA (Rennes), équipe Texmex : publications communes, collaborations régulières
- IRIT (Toulouse) : publications communes (équipe ADRIA), projet RTRA Aéronautique commun, membre de comité de thèses
- IMT (Toulouse) : projet ANR en commun, co-encadrement de thèse
- LabSTICC (Bretagne sud) : publications en commun, organisation d'événements scientifiques
- LBBE (Lyon 1), équipe Epidémiologie et Santé Publique : collaborations récurrentes
- LHC (Saint-Etienne) : publications en commun, organisation d'événements scientifiques
- LIA (Avignon) : séjour scientifique au LIA (1 semaine en 2010), projet ANR ImagiWeb (2012-2015) en commun
- LIP6 (Paris), équipe ACASA : collaborations régulières (co-encadrement de thèses, dépôt de projets ANR)
- LIRIS (Lyon), équipe DM2L : publications en commun et co-encadrement de thèse
- LIRMM (Montpellier), équipe ICAR : un chercheur est associé au LIRMM, publications communes

Industrielles

- Alcatel-Lucent Bell Labs : accueil de stages de recherche, collaborations régulières
- AID (Paris) : thèse CIFRE
- AMI Software (Montpellier) : thèse CIFRE, projet ANR ImagiWeb
- EDF (Paris) : projet ANR ImagiWeb
- Creative Research : thèse CIFRE
- Technicolor (Rennes) : co-encadrement de stages de recherche et mise en place prochaine d'une thèse CIFRE
- Xerox Research Center Europe (Grenoble) : projet ANR ImagiWeb
- Visioglobe : accompagnement scientifique

2.8. Dix principales publications

- [09-229] Culus, J. F. and Jouve B. (2009), Convex circuit free coloration of an oriented graph. European Journal of Combinatorics 30(1): pp. 43-52.
- [09-41] De Peretti, C., Siani, C. and Cerrato, M. (2009), An artificial neural network based heterogeneous panel unit root test in case of cross sectional independence. International Joint Conference on Neural Networks (IJCNN), pp. 2487-2493.
- [11-224] Kafafy, A., Bounekkar, A. and Bonnevay, S. (2011), A Hybrid Evolutionary Metaheuristics (HEMH) Applied On 0/1 Multiobjective. Knapsack Problems, Genetic and Evolutionary Computation Conference (GECCO), Dublin, Irlande, pp. 497-504.
- [12-222] Le Bras Y., Lenca P., Lallich S. (2012), Optimonotone Measures for Optimal Rule Discovery. Computational Intelligence (CI), Wiley. Article first published online : 2 MAY 2012 | DOI: 10.1111/j.1467-8640.2012.00422.x
- [11-215] Musat, C., Velcin, J., Rizoiu, M. A. and Trausan-Matu, S. (2011), Improving Topic Evaluation Using Conceptual Knowledge. Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI). Barcelona, Spain. July 2011.
- [10-210] Pisetta, V., Jouve, P. E. and Zighed, D. A. (2010), Learning with Ensembles of Randomized Trees : New Insights. ECML/PKDD, pp. 67-82.
- [12-228] Kasparian Jérôme and Rolland, A. (2012). OECD's "Better life index": can any country be well ranked ?. Journal of Applied Statistics, 39, 10, pp. 2223-2230.
- [12-235] Silic, A., Morin, A., Chauchat, J. H. and Basic, B. D. (2012), Visualization of Temporal Text Collections Based on Correspondence Analysis. Expert Systems with Applications, vol. 39, no. 15.
- [09-60] Thibault, G., Aussem, A. and Bonnevay, S. (2009), Incremental Bayesian Network Learning for Scalable Feature Selection, The 8th International Symposium on Intelligent Data Analysis (IDA), Lyon (France), vol.5772, pp. 202-212.
- [11-236] Zighed, D. A., Ezzeddine, D. and Rico, F. (2012), Neighborhood Random Classification. The 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Kuala Lumpur : Malaisie. Springer-Verlag.

Section 3

Projet
2013 / 2017

1 Stratégie globale

Suite à l'évaluation par l'AERES du laboratoire ERIC en 2010 au début de sa réorganisation, nous avons souhaité mettre en place une évaluation à mi-parcours, comme c'est le cas depuis la création du laboratoire, selon les mêmes modalités.

Notre objectif est de :

1) préciser notre stratégie globale par rapport aux observations des évaluateurs en 2010 ;

2) repositionner notre projet scientifique à cinq ans en fonction de la recomposition en deux équipes de recherche.

1.1. Positionnement scientifique

Depuis sa création, le laboratoire ERIC s'attache à maintenir une activité de recherche à trois niveaux : travaux à caractère théorique, développement de logiciels et recherche de terrains d'application.

Dans ce dernier domaine, nous allons poursuivre la politique amorcée vis-à-vis des SHS, avec notamment l'intégration à l'Institut des Sciences de l'Homme.

Le projet SHS-Doc-Net, financé par Lyon 2 et réalisé en collaboration avec l'ISH, devrait nous permettre d'être plus proactifs dans cette optique. Le projet SHS-Doc-Net a en effet permis la mise en ligne d'un portail Internet mettant en oeuvre des technologies innovantes du Web social sémantique et dont l'objet est la mise en valeur des compétences en SHS développées par les acteurs de la recherche de l'Université de Lyon. Le portail est une application collaborative qui permet aux acteurs de la recherche en SHS d'exposer leurs expertises, compétences et savoir-faire, de gérer leur identité numérique professionnelle sur le Web, de bénéficier de services et outils pour le développement de collaborations ainsi que pour la veille scientifique.

1.2. Positionnement dans l'environnement

Malgré une spécificité thématique (entrepôts et fouille de données) sur laquelle ERIC est leader en France et un terrain d'application privilégié (les SHS) original sur la place de Lyon, les intersections et les complémentarités scientifiques avec les autres laboratoires d'informatique lyonnais existent. Des liens existent d'ores et déjà avec le LIRIS (équipe Bases de données et équipe Data mining et machine learning en cours de création), le DISP (équipe Modélisation, Intégration, Système d'Information) et le Centre Magellan (équipe MODEles et MEthodes de conception des systèmes d'information avancés) de Lyon 3. Au-delà de Lyon, ERIC entretient également des liens avec le Laboratoire Hubert Curien (équipe Machine learning) de Saint-Etienne et commence à collaborer avec l'IMAG de Grenoble (Section 3. Partie 1.3). Ces liens se matérialisent notamment par des participations croisées à des jurys de thèse, l'organisation d'événements comme la conférence internationale ALT-DS 2012[21] et la collaboration au sein groupes comme l'ARC6[22] de la Région Rhône-Alpes (projet Web Intelligence), la plateforme eTechSanté WEBIMATICS[23] ou l'Institut rhône-alpin des systèmes complexes (IXXI)[18]. Notre objectif est de renforcer ces liens et de développer des synergies plus importantes, notamment dans le cadre de projets régionaux ou nationaux.

21 http://eric.univ-lyon2.fr/alt-ds-2012
22 http://arc.rhonealpes.fr/spip.php?rubrique36#27
23 http://sfrsantelyonest.univ-lyon1.fr/plateau23-etechsante.html

1.3. Politique scientifique

L'effort consenti sur la qualité de nos publications ayant commencé à porter ses fruits, il s'agit maintenant de continuer à renforcer le niveau de qualité de la production scientifique du laboratoire.

Par ailleurs, la dotation de nos tutelles tendant à décroître régulièrement au profit des financements par projets et l'opportunité de participer à des projets européens futurs n'étant pas garantie (absence de structure de valorisation adaptée à Lyon 2), l'accent doit plus que jamais être mis sur la soumission de projets nationaux afin que le budget global du laboratoire puisse continuer de se développer favorablement

Quatre projets de ce type sont d'ores et déjà en cours d'élaboration :

- Portal of Objects in the Cloud (POC) : intégration de données hétérogènes, analyse et visualisation, optimisation des performances et sécurité ; avec des champs d'application dans l'environnement et le patrimoine - Partenaires : ETIS Cergy-Pontoise, IRSTEA Clermont-Ferrand, LIMOS Clermont-Ferrand, LIRMM Montpellier, LRI Paris Sud, BNF, Musée Rodin - Programme visé : ANR INS - Porteur : ERIC.
- Sécurité et Cloud, applications dans le domaine de la santé - Partenaires : IMAG Grenoble, LAMSADE Paris 5, LIMOS Clermont-Ferrand, société Yansys Medical - Programme visé : ANR CSOSG - Porteur : LAMSADE.
- REcursive QUEry and Scalable Technology (REQUEST) : création, structuration et animation de la communauté française "Big Data" ; traitement du problème de la dualité big data/ big analytics - Partenaires : Thalès, Orange Labs, Alcatel Lucent, LIP6 Paris, LABRI Bordeaux, UTT Troyes, CNRS, INRIA, 7 PME, Police et Gendarmerie Nationale - Programme visé : Investissements d'avenir/Fonds national pour la société numérique (Cloud Computing/Big Data) - Porteur : Thalès.
- Détection et suivi des rôles dans les communautés en ligne : à la suite de la mise en place d'une thèse CIFRE fin 2012 avec l'entreprise Technicolor, qui constitue actuellement une nouvelle équipe autour du data mining sur le Web, il est prévu de répondre ensemble au prochain appel ANR CONTINT avec un consortium en cours de constitution. Le porteur n'est pas encore désigné.

1.4. Politique de recrutement

Le laboratoire va avoir à gérer deux départs à très court et court termes : Rafik Abdesselam (maître de conférences section 26), promu professeur des universités en 2012 dans un autre laboratoire et Stéphane Lallich (professeur section 27 mais statisticien d'origine), dont le départ à la retraite est prévu en 2013 ou 2014.

ERIC s'étant toujours nourri de la complémentarité de l'informatique et des mathématiques appliquées, et notamment de la statistique, il est vital pour le laboratoire de pouvoir compenser ces départs et recruter un maître de conférences et un professeur de statistique (section 26).

Toutefois, dans une optique de développement de notre laboratoire d'informatique et pour faire diminuer la pression des heures complémentaires d'enseignement, le recrutement de collègues de la section 27 (informatique) est également très important. Il est donc essentiel de pouvoir recruter un maître de conférences et un professeur dans cette section pour accompagner nos efforts dans le développement des humanités numériques (poursuite du projet SHS-Doc-Net, projet de formation en humanités numériques). Pour tous ces recrutements, ERIC continuera de favoriser les candidatures extérieures.

1.5. Synergie enseignement-recherche

Bien que l'activité d'enseignement et d'animation pédagogique des membres d'ERIC soit très importante et chronophage, elle est un élément essentiel pour la vitalité de la recherche au sein du laboratoire et sa visibilité des collègues de SHS avec qui nous développons des collaborations.

Il est donc indispensable, non seulement de la maintenir, mais aussi de la renforcer.

Nous avons dans cette optique un projet de formation pluridisciplinaire en humanités numériques au niveau master, qui pourrait être déployé au niveau du PRES Lyon-Saint-Etienne. Il nous semble également important de proposer au niveau licence une "majeure associée" à toutes les composantes de Lyon 2, en collaboration étroite avec les formations bidisciplinaires déjà en place (licences MIASHS et IDEA), afin de proposer un cursus SHS-informatique-mathématiques appliquées complet, de la licence au doctorat, à tous les étudiants intéressés.

1.6. Analyse de la stratégie

Nous synthétisons dans la Figure 21 sous forme de matrice SWOT les atouts et les handicaps liés à la stratégie que nous venons de présenter.

Forces	**Faiblesses**
• Thématique de recherche sur laquelle ERIC bénéficie d'une reconnaissance nationale et internationale. • Positionnement au sein d'un secteur d'application (SHS) en demande d'expertise dans notre domaine.	• Bilocalisation du laboratoire. • Moyens matériels (locaux à Lyon 2 et à Lyon 1) et humains (personnel administratif et technique) insuffisants, comme déjà souligné par le précédent rapport des experts de l'AERES (Annexe 2)
Opportunités	**Menaces**
• Demande sociale pour la valorisation des "big data" • Afflux d'étudiants dans nos formations de master, notamment internationales • Développement de collaborations plus étroites avec les laboratoires lyonnais et rhône-alpins aux thématiques voisines	• Non affectation par nos établissements de tutelle des moyens matériels et humains nécessaires pour atteindre nos objectifs • Non affectation par nos établissements de tutelle de nouveaux postes d'enseignants-chercheurs permettant le développement du laboratoire

Figure 21 : Matrice SWOT du projet d'ERIC

2 Projet scientifique

Le projet scientifique du laboratoire, bien que décliné selon ses deux équipes, souligne clairement des synergies, en particulier autour de la prise en compte des données textuelles et des réseaux sociaux dans l'analyse décisionnelle.

Ces synergies se concrétisent aujourd'hui dans la codirection de quatre thèses et la participation conjointe à des projets de recherche (ImagiWeb, CNAF, Qualité).

La conception de nouvelles méthodes d'aide à la décision pour l'analyse des médias sociaux nous semble d'une importance stratégique majeure. Ces méthodes pourront être facilement mises en oeuvre avec l'aide de spécialistes en science de la communication, voire de journalistes, collaborations qui seront facilitées par l'écosystème scientifique dans lequel se situe ERIC.

De plus, ces travaux pourront être aisément appliqués à d'autres domaines des SHS. Par exemple, face au volume de plus en plus important des bases de données historiques, un système de recommandation pourrait aider l'historien à résoudre des problématiques de son domaine. Connaître les tendances ou résumer un vaste corpus de données hétérogènes, en particulier textuelles, pourraient également apporter de nouveaux outils aux linguistes ou aux spécialistes en sciences sociales.

2.1. Équipe SID

Les entrepôts de données constituent un terrain fertile pour effectuer de nouvelles recherches. Aussi, les perspectives associées à ce domaine de recherche sont nombreuses. Certaines font d'ores et déjà partie de nos prospections. D'autres sont des perspectives à plus long terme.

2.1.1. Entrepôts et analyse en ligne d'informations

Dans la continuité de nos travaux sur l'entreposage et l'analyse en ligne de données complexes, nous abordons la problématique du décisionnel dans le cadre spécifique des données textuelles, et plus particulièrement des réseaux sociaux. Un de nos objectifs scientifiques est de concevoir des méthodes d'analyse en ligne adaptées aux réseaux sociaux, à leur structure (le plus souvent représentée sous forme de graphes) et aux types d'analyse ayant cours dans le domaine (détection de communautés d'intérêts, par exemple). Il faut alors définir la notion de cube de graphes, d'agrégation, etc.

De manière générale, nous assistons de plus en plus à une profusion de données, d'informations et de connaissances (obtenues à partir de plusieurs outils d'analyse et d'extraction de connaissances) dans les entreprises et sur le Web sémantique. Ces informations sont souvent peu ou pas exploitées. Par exemple, des travaux récents ont abordé l'analyse multidimensionnelle de résultats de simulation. Il s'agit de produire, à partir d'une analyse en ligne sur les résultats de simulation, de nouvelles connaissances pouvant aider le décideur à mieux évaluer l'impact d'une décision.

L'idée plus globale que nous souhaitons développer est la construction d'entrepôts d'informations et/ou de connaissances pouvant élargir la notion d'analyse en ligne des données à celle de l'analyse en ligne à partir d'informations. Nous pensons que cette nouvelle approche ouvre des perspectives intéressantes pour l'aide à la décision. Il s'agit de concevoir les modèles d'entrepôts les plus adaptés pour intégrer des sources d'informations et de connaissances et de définir de nouveaux opérateurs OLAP de navigation et d'exploration des informations.

Dans ce contexte, nous nous intéressons particulièrement à deux pistes de recherche qui nous paraissent importantes : les entrepôts de données ouverts et l'analyse en ligne collaborative.

Entrepôts de données ouverts

Notre objectif est de construire des entrepôts de données qui peuvent être enrichis par des informations / connaissances issues d'autres applications ou plus largement provenant du Web.

En effet, plusieurs applications (de santé publique, par exemple) nécessitent en plus des données issues de sources locales ou de sources externes des informations complémentaires qui peuvent provenir du Web, l'objectif étant par exemple de prévenir une pandémie au lieu de la constater a posteriori.

Pour atteindre ce but, différents verrous scientifiques doivent être levés :

1) intégrer dans le processus d'entreposage une ontologie du domaine étudié afin d'alimenter l'entrepôt avec des informations pertinentes (instances et relations entre instances dans l'ontologie) ;

2) alimenter l'entrepôt avec des données de simulation afin d'affiner des résultats d'analyse ;

3) de manière un peu plus orientée vers les données du Web (réseaux sociaux, blogs, etc.), étudier les entrepôts Web (data webhouses) en utilisant la technologie du Web sémantique combinée avec l'OLAP.

Analyse en ligne collaborative

À l'image du Web, qui est un lieu d'informations et d'échanges qualifié de social, participatif et collaboratif, les applications décisionnelles se doivent de fournir une analyse en ligne pouvant être partagée par plusieurs utilisateurs selon leurs profils. Les utilisateurs ont alors la possibilité d'annoter, de donner leur avis, etc. sur les analyses partagées.

Dans ce contexte, nous nous intéressons particulièrement à la problématique posée par ce que l'on peut qualifier d'analyse en ligne collaborative. Plusieurs verrous scientifiques sont alors soulevés dans le but de faciliter le processus de personnalisation et de recommandation d'analyses en ligne collaborative :

1) comment enrichir les modèles d'entrepôt de données et des cubes OLAP par des connaissances utilisateurs (profils, annotations, etc.) ;

2) comment intégrer les annotations personnalisées des différents utilisateurs dans le processus d'analyse en ligne ;

3) comment évaluer la qualité d'une analyse personnalisée dans un tel contexte ?

Pour répondre à ces questions, il est nécessaire de définir un environnement collaboratif dédié à un groupe d'utilisateurs désirant effectuer un ensemble d'analyses OLAP répondant à un objectif donné. Cet environnement doit aider l'utilisateur à obtenir une meilleure qualité d'analyse et à bénéficier efficacement de l'expérience des autres membres du groupe, sachant que la pertinence de l'information délivrée et son adaptation aux usages et préférences des utilisateurs constituent des facteurs clés du succès ou du rejet de ces systèmes.

Une des mesures de la qualité d'un processus d'analyse est le temps de réponse, qui est le temps nécessaire au processus pour exécuter la requête et afficher le résultat.

Concernant la qualité des données délivrées, il s'agit d'évaluer la qualité d'une analyse proposée en réponse à un besoin, notamment la proximité des résultats calculés avec les résultats attendus. C'est dans ce cadre que nous envisageons d'engager nos recherches à venir pour améliorer le processus de personnalisation dans les systèmes décisionnels collaboratifs.

2.1.2. Informatique décisionnelle dans les nuages

Un autre problème auquel les entreprises sont confrontées est que leurs équipes passent beaucoup de temps à collecter de l'information, à la transformer puis à l'intégrer dans un système décisionnel au détriment de l'analyse, qui est bien sûr l'activité à plus forte valeur ajoutée.

Afin de gagner en performance et en compétitivité, il est essentiel de diffuser l'usage de la Business Intelligence (BI) au plus grand nombre d'utilisateurs, afin que chaque décideur (expert ou non) soit en position de prendre les meilleures décisions. Les informations pourraient dans ce cas être accessibles et partagées par le plus grand nombre de collaborateurs, en fonction des rôles de chacun (besoins d'analyse spécifiques, besoins de métadonnées propres à chaque application décisionnelle, contraintes de visualisation, mais aussi droits d'accès, etc.).

C'est pourquoi nous nous intéressons à la BI à la demande (BI collaborative), qui doit être capable de fournir un outil d'analyse en ligne pour tous. Au-delà des applications d'entreprise, ce concept doit aussi permettre aux citoyens (ONG, coopératives, associations, particuliers) de s'approprier les outils décisionnels.

En plus de partager des données et des informations, nous étudions la question du partage des résultats d'analyse. Cette problématique est nouvelle dans le cadre des entrepôts et de l'analyse en ligne et s'inscrit dans le paradigme de l'informatique dans les nuages (Cloud Computing). L'analyse de données au sein des nuages est d'ailleurs devenu alors un enjeu majeur (Big Data / Big Analytics).

Les problèmes à aborder regroupent les problèmes classiques des systèmes largement distribués, mais aussi de nouveaux problèmes liés spécificités des nuages informatiques : facturation à l'utilisation, élasticité et facilité d'utilisation (partage d'infrastructure, de programmes, d'applications, de données, d'espace de stockage, d'applications, etc.). Plus spécifiquement, les verrous scientifiques relatifs à l'informatique décisionnelle dans ce contexte incluent la définition de modèles pour le stockage de données multidimensionnelles dans les nuages et de stratégies d'analyse en ligne à la volée prenant en compte des aspects de personnalisation basés sur les profils utilisateurs. Enfin, il nous paraît nécessaire, d'une part, d'élaborer des stratégies d'optimisation de performance permettant de minimiser les coûts d'exploitation au lieu de ne miser que sur l'élasticité et, d'autre part, de répondre de façon fiable aux problématiques de confidentialité, d'intégrité et de disponibilité des données stockées dans les nuages, mais aussi des résultats d'analyse susceptibles d'être partagés.

Nous envisageons quelques pistes de recherche afin de lever ces verrous scientifiques :

- définition de structures et de langages pour modéliser les données multidimensionnelles dans les nuages ;
- garantie de l'aspect en ligne de l'OLAP par l'utilisation de techniques d'optimisation de performance telles que la matérialisation de vues ;
- personnalisation des analyses selon les profils utilisateurs ;
- intégration des données de niveaux de protection différents dans une même infrastructure (utilisation de la cryptographie pour résoudre les problèmes de confidentialité des données) ;
- étude de la qualité des données (fraîcheur, intégrité, etc.) ;
- utilisation d'agents intelligents pour l'intégration, l'analyse en ligne et la sécurité des données dans un environnement collaboratif dans le nuage.

2.2. Équipe DMD

2.2.1. Contexte scientifique

Depuis une dizaine d'années, les données complexes ont été au cœur des préoccupations des chercheurs d'ERIC spécialisés en fouille des données. Alors que les travaux de l'équipe DMD concernaient des verrous aujourd'hui bien connus de ces données (hétérogénéité, volume important, données textuelles, présence de métadonnées), l'objectif consiste pour les années à venir à se concentrer sur deux caractéristiques essentielles qui remettent en cause les modèles et les algorithmes existants : les données sont en relation les unes avec les autres et elles évoluent dans le temps.

En effet, il est bien clair aujourd'hui que les observations qui servent de base aux techniques de recherche d'information, d'apprentissage automatique ou de modélisation de la décision ne peuvent plus être considérées comme indépendantes les unes des autres, mais reliées (interconnectées) les unes aux autres. On parle de fouille de liens (link mining) et de données relationnelles, qui dépassent le cadre habituel d'observations indépendantes (paradigme IID). C'est le cas par exemple lorsque les internautes qui produisent des opinions sont inscrits dans un ou plusieurs réseaux sociaux, alors que dans le même temps les messages sont eux-mêmes reliés entre eux et déposés sur des sites Web connectés les uns aux autres. Deuxièmement, il n'est plus possible d'aborder les données uniquement de manière statique, mais il est de plus en plus nécessaire de prendre en considération la dimension temporelle et la dynamique d'évolution de ces données.

Ces deux dimensions que nous souhaitons privilégier au sein de l'équipe sont bien présentes dans la plupart des applications en SHS. Les bases de données historiques, par exemple, nécessitent de traiter des objets de natures différentes (individu, communauté, association, entreprise, pays, etc.), reliés entre eux et où l'aspect chronologique est fondamental.

Une autre application que nous souhaitons mettre en avant dans ce projet scientifique est l'analyse des nouveaux médias sociaux, pour laquelle la fouille de données complexes constitue une approche privilégiée et encore peu explorée.

Qu'il s'agisse de partage d'informations sur l'actualité (digg, huffingtonpost, mediapart), de partage de photos et de vidéos (youtube, flickr, instagram), d'opinions exprimées sur des articles ou des films (epinions, myopinionnow), des blogs et microblogs (twitter, weibo), des réseaux sociaux (facebook, google+, linked'in), ce terrain d'application centré sur la création et le partage de contenu et d'information va nous permettre d'illustrer la présentation plus détaillée qui suit du projet scientifique de l'équipe .

2.2.2. Projet scientifique

Dans le cadre des données complexes (en particulier interconnectées et évolutives) pour lesquels l'aspect social (création, partage, diffusion) est de plus en plus présent, nous souhaitons étudier plusieurs problèmes importants en tirant profit de l'expertise complémentaire des membres de l'équipe DMD.

Tout d'abord, nous souhaitons développer de nouveaux systèmes de recommandation automatisés adaptés aux données complexes. Face à des données qui n'ont jamais été aussi volumineuses (des milliards de tweets par jour, des milliers d'articles de presse qui provoquent chacun des milliers de réactions, etc.), il est en effet indispensable de guider l'utilisateur en lui proposant quel objet pourrait l'intéresser, quel article ou quel commentaire il pourrait lire, quel internaute serait en mesure de répondre à sa question, etc.

Répondre à cette problématique implique de lever des verrous scientifiques importants. En effet, les algorithmes actuels doivent être considérablement adaptés afin de prendre en compte de manière pertinente le caractère dynamique et surtout contextuel des données : qui les a produit, à quel moment, sur quel site Web, etc. ?

La notion de contexte est entendue au sens large. Par exemple, comment peut-on prendre en compte les événements extérieurs (ex. : coupe du monde de football, cérémonie des Oscars, etc.) dans l'évolution des opinions ?

Pour attaquer ces verrous, nous projetons de développer de nouveaux modèles et algorithmes de recherche d'information et d'apprentissage automatique qui prennent en compte la nature évolutive des données, ainsi que le contexte local des objets.

Nous souhaitons continuer à travailler en particulier sur la représentation des objets, au-delà du classique espace vectoriel. Il faut également s'interroger sur le meilleur espace qui permet de les comparer, et sur la notion de voisinage, par exemple en employant des modèles issus de la théorie des graphes et de l'apprentissage topologique. Il nous semble important de prendre en compte conjointement les informations qui caractérisent les objets (variables, métadonnées, textes) mais également la structure qui connecte ces objets.

Il est nécessaire aujourd'hui que les techniques issues de la fouille de données (en particulier la fouille de textes et d'opinion) tirent partie de recherches théoriques sur la caractérisation de familles de graphes, telles que celles qui ont déjà été initiées dans l'équipe.

De manière complémentaire, les modèles d'aide à la décision nécessiteraient d'être revisités afin de pouvoir traiter ce type de données. A ce titre, l'agrégation de critères basés sur des données interconnectées et évolutives constitue un verrou tout à fait nouveau et adapté à l'expertise des chercheurs de l'équipe.

Les travaux récents sur les fonctions d'agrégation de préférence devraient permettre l'émergence de nouveaux modèles d'apprentissage automatique, en particulier supervisé.

Nous souhaitons également construire de nouvelles représentations synthétiques des données complexes, comme des cartes ou des résumés qui sont indispensables de nos jours dans un cadre de veille d'information.

Il s'agit de privilégier des méthodes d'apprentissage automatique non supervisé (approche bottom-up) dans le but d'extraire des tendances avec peu de connaissance a priori sur le résultat attendu, ou de structurer les objets (documents, individus, etc.) en groupes/communautés.

Cet objectif rencontre là encore des verrous scientifiques importants, comme par exemple de déterminer le meilleur espace de représentation de ces résumés, bien au-delà d'une simple catégorisation, ou d'être capable de prendre en compte la dimension temporelle sans découper systématiquement le temps en "tranches", approche la plus usitée mais qui présente de sérieux désavantages.

En particulier, l'extraction de rôles à partir de réseaux (y compris latents) d'internautes peut constituer un excellent cas d'étude où nous pourrions réconcilier une approche de l'apprentissage supervisée et non supervisée en explorant un clustering semi-supervisé (à ne pas confondre avec l'apprentissage semi-supervisé).

Plus précisément, l'idée est de structurer les données avec le minimum d'a priori, mais en prenant en compte un certain nombre de connaissances, par exemple formalisées sous la forme de contraintes.

La prise en compte de données temporelles et contextuelles permettrait alors de revisiter les modèles existants. Formellement, nous envisageons de développer de nouveaux modèles graphiques probabilistes afin de mieux prendre en compte les relations de dépendance (spatiale, temporelle, etc.) entre les données.

Là encore, les liens entre l'apprentissage automatique, en particulier non supervisé, et l'agrégation des préférences constituent un champ d'investigation qui nous paraît original et fécond.

Les problèmes évoqués ci-dessus ne constituent que deux exemples pour lesquels les chercheurs de l'équipe pourraient développer de nouveaux modèles et algorithmes.

Compte tenu des enjeux que représentent la fouille de données complexes, l'équipe pourra tirer pleinement partie de la complémentarité de compétence de ses membres.

En effet, l'équipe DMD a pour particularité d'intégrer des chercheurs en informatique, en statistique et en mathématiques appliquées.

Conception graphique : ACTESSENS - SIRET 52812431600019