LABO ERIC
EQUIPE DE RECHERCHE EN INGENIERIE DES CONNAISSANCES

Projet scientifique Laboratoire ERIC

2011-2014

Université de Lyon Lumière Lyon 2
5, avenue Pierre Mendès-France
69676 Bron Cedex

Tel. +33 4 78 77 23 76
Fax. +33 4 78 77 23 75

Web. http://eric.univ-lyon2.fr

Université de Lyon, C. Bernard Lyon 1
43, bd du 11 novembre 1918
69622 Villeurbanne cedex

Tél. +33 4 72 43 16 54
Fax +33 4 72 43 10 44

Sommaire

1 Contexte 5

2 Projet et objectifs scientifiques 8

2.1 Cadre scientifique : Fouille de données complexes et processus de décision associés 8

2.2 Caractéristiques des données complexes 8

2.3 Défis scientifiques dans la FDC 9

2.4 Défis technologiques de la prise de décision dans la cadre de la FDC 10

2.5 Projet scientifique 11

2.5.1 Axes théoriques fondamentaux 12

2.5.2 Domaines d'application transversaux 12

2.5.3 Implication technologique 13

2.5.4 Synthèse 13

3 Organisation du laboratoire 14

3.1 Gouvernance 14

3.1.1 Correspondants de sites 15

3.1.2 Conseil de laboratoire 15

3.1.3 Conseil de direction 16

3.1.4 Direction 16

3.2 Organisation Scientifique 16

4 Adequation des moyens humains et financiers de l'unité avec le projet 18

4.1 Politique d'animation scientifique 18

4.2 Analyse prospective à moyen et long terme des moyens et des compétences 18

4.3 Politique de construction de partenariats 19

4.4 Schéma de financement du projet 19

4.5 Capacité de l'unité à valoriser ses travaux de recherche 19

4.6 Implication de l'équipe en matière de diffusion de l'information scientifique et technique. 19

5 Annexe : descriptions des axes théoriques et domaines transversaux de recherche 21

5.1 ENtrepôts et Analyse en ligne de Données Complexes (ENA-DC) 21

5.2 FOuille de Données et Apprentissage (FODA) 23

5.3 DECision et COmplexité (DECCO) 27

5.4 Santé et Environnement 29

5.5 Sciences Humaines et Sociales (SHS) 31

5.6 Logiciels libres 34

1 CONTEXTE

Au cours de l'année 2009, le laboratoire ERIC souhaitait accueillir de nouveaux chercheurs issus de deux autres établissements : l'Université Claude Bernard Lyon 1 et L'École Pratique des Hautes Études de Paris. Le but était de réunir les forces de trois équipes de recherche :

- EA 3083 ERIC de l'université Lyon 2, dirigée par le Pr. Zighed ;
- Equipe LaISC de l'EPHE de Paris (issue de l'EA 4004), dirigée par le Pr. Bui ;
- Equipe MA^{2}D de l'université Lyon 1, dirigée par le Pr. Lamure et issue du LIRIS[1] UMR 5205 CNRS-Lyon1-Lyon2-INSA-ECL.

Cependant, compte tenu de la réoroganisation en cours dans les universités parisiennes, seuls les collègues de l'université Lyon 1 ont pu intégrer le laboratoire ERIC.

Pour le prochain contrat quadriennal, le laboratoire ERIC, ainsi nouvellement constitué, sera positionné sur des thématiques originales et porteuses. Il sera capable de relever des défis scientifiques et technologiques dans ses domaines d'expertise tout en atteignant une visibilité internationale.

Le regroupement du laboratoire ERIC avec l'équipe MA^{2}D de l'université Lyon 1 se justifie par :

- Les complémentarités thématiques des deux équipes qui, mises en commun, permettent d'envisager la prise en main de problèmes scientifiques complexes notamment pour la fouille de données volumineuses et non structurées, la modélisation et la simulation de processus complexes en particulier dans le domaine de la décision ;
- La forte sensibilité de chacune des équipes aux applications dans les domaines des SHS (Sciences Humaines et Sociales) et de la santé ainsi que leur expertise acquise et reconnue dans ces domaines ;
- L'existence de longues et nombreuses collaborations à la fois scientifiques et pédagogiques entre les équipes.

[1] LaISC: Laboratoire d'Informatisation des Système Complexes; MA2D Méthodes & Algorithmes pour l'Aide à la Décision; LIRIS Laboratoire Informatique des Réseaux de l'Image et des Systèmes.

En outre, ce regroupement vise à :

- Accroître la masse critique d'enseignants chercheurs qui atteint d'ores et déjà 22 enseignants-chercheurs permanents, 5 chercheurs associés et près de 40 doctorants ;
- Favoriser la conduite de grands projets de recherche fonctionnant sur des crédits pluriannuels tels que des projets ANR, européens, ...
- Développer une recherche informatique de haut niveau à l'interface des domaines d'application en sciences humaines et sociales et en santé.

Le projet scientifique de ce nouveau laboratoire s'appuie sur les compétences des deux équipes :

- ERIC : entrepôts de données complexes, fouille de données, apprentissage automatique ;
- MA^2D : méthodes et outils pour l'aide à la décision dans un environnement complexe et incertain.

Ainsi, la thématique de recherche du futur laboratoire s'inscrira dans le domaine de la fouille de données, la modélisation et la simulation des processus de décision en environnement complexe. Nous expliciterons plus loin les enjeux et les verrous scientifiques et technologiques de cette problématique.

Les recherches seront structurées en projets de manière à développer des synergies entre les chercheurs issus des deux équipes initiales. Cela passera par des collaborations autour de problématiques théoriques, autour du développement de logiciels et autour des applications avec des partenaires universitaires ou industriels. Plus encore, toute l'activité de recherche sera articulée autour des formations de Master et de doctorat afin d'attirer de bons étudiants vers la recherche.

Le rattachement principal du futur laboratoire sera l'Université Lyon 2. Sa direction sera confiée à l'actuel directeur d'ERIC, le Pr. D.A. Zighed qui s'appuiera sur un conseil de direction composé des responsables actuels de chaque équipe et d'un conseil de laboratoire qui associera de manière large des représentants des professeurs, des maîtres de conférences, des chercheurs, des ITA et des doctorants.

La réussite de ce projet passe par la prise en compte :

- de points forts :
  - la nouvelle configuration du laboratoire en termes de complémentarités thématiques, d'effectifs et de partage d'une vision commune sur les enjeux scientifiques et technologiques,
  - une thématique sur laquelle nous bénéficions d'une reconnaissance ;
- un positionnement au sein d'un secteur d'applications (SHS et santé) en forte mutation technologique, demandeur d'une expertise pour laquelle nous disposons de compétences ;

- des points faibles qui sont :
    - la multilocalisation du laboratoire sur deux sites,
    - la nécessité de moyens de gestions nouveaux pour faire face à l'évolution de l'équipe,
    - les faibles moyens en termes de personnel administratif et technique ;
- des opportunités qui sont :
    - une thématique porteuse sur laquelle il existe une forte demande sociétale en termes d'applications industrielles, d'emplois et de formations,
    - la proximité d'un vivier de doctorants pouvant contribuer aux travaux de recherche,
    - la connaissance et les liens que nous avons avec la communauté internationale travaillant sur nos thématiques,
    - notre présence au sein des structures d'animation et de valorisation de la recherche, notamment à travers notre participation aux comités de programmes des principales conférences nationales et internationales du domaine,
- des risques qui sont liés à :
    - la non affectation, par nos tutelles, des moyens humains, administratifs et techniques nécessaires, pour atteindre nos objectifs,
    - l'insuffisance des moyens financiers dont nous pourrions disposer pour attirer de bons doctorants, post-docs, et conférenciers,
    - la non affectation, par les établissements, de nouveaux postes d'enseignants chercheurs au laboratoire.

2 PROJET ET OBJECTIFS SCIENTIFIQUES

2.1 Cadre scientifique : Fouille de données complexes (FDC) et processus de décision associés

Depuis toujours, le rôle des données complexes (image, vidéo, texte non structuré ou combinaison de ces médias) n'a cessé de croître, pour être aujourd'hui le principal véhicule d'information. La problématique d'une diffusion massive et de qualité est quasi-réglée grâce aux technologies de l'entreposage massif des données et aux réseaux à haut débit. Nous disposons de volumineuses bases de données complexes dont la croissance est exponentielle mais dont la valorisation reste encore très faible.

Le défi qu'il faut relever est de tirer profit, dans tous les sens du terme, de ces données : recherche d'information, extraction de connaissances, création de valeur économique etc. La Fouille de Données Complexes tente de répondre à ce besoin. Elle propose de définir un cadre méthodologique et des outils pour structurer les données complexes, les analyser en vue d'extraire des connaissances ou des informations non accessibles par des moyens classiques. La finalité de ce processus est d'accroître les connaissances des acteurs-décideurs dans des domaines précis. Le projet scientifique d'ERIC s'inscrit dans la perspective d'analyser, de modéliser et de fournir des systèmes d'aide à la décision aux acteurs travaillant à partir de corpus de données complexes notamment dans les SHS et la santé.

2.2 Caractéristiques des données complexes

La plupart des données réelles disponibles, issues de la vie de tous les jours, sont complexes. Elles sont généralement :

- Volumineuses : plusieurs téraoctets. Par exemple le Dossier Médical Personnalisé (DMP) peut atteindre plusieurs dizaines de giga-octets par patient ;
- Distribuées : le DMP, pour rester sur cet exemple, peut être stocké dans différentes bases de données distribuées selon les services médicaux où le patient a séjourné ;
- Hétérogènes : les données peuvent être de différente nature. Dans le cas toujours du DMP, on aura des images radiologiques, des comptes-rendus textuels, des tableaux de chiffres de mesures biologiques, des courbes d'électrocardiogramme, des enregistrements vidéo d'échographie, etc.

- Evolutives : différents enregistrements avec des contenus différents. Par exemple, le DMP contient divers examens réalisés à des instants différents et qui ne portent pas nécessairement sur les mêmes tests médicaux ;
- Non structurées : elles ne sont généralement pas modélisées dans le cadre d'un schéma de base de données mais stockées quasiment en vrac et dans le meilleur des cas dans des formats ad hoc. Elles échappent souvent au classique format attribut-valeur.

2.3 Défis scientifiques dans la FDC

Les défis scientifiques que soulèvent ces particularités des données complexes sont multiples. Parmi ceux que l'on peut assez facilement identifier, on peut lister :

- La grande dimensionnalité. Outre les problèmes liés à l'optimisation du stockage des données volumineuses, les attributs issus des images, des textes, des graphes-réseaux et des autres peuvent atteindre plusieurs centaines, voire des milliers de variables. Comment évaluer la pertinence de cet espace de représentation par rapport aux tâches que l'on souhaite effectuer comme l'apprentissage supervisé ou la classification ? comment réduire l'espace si tant est que cela soit possible ?
- Absence de structure mathématique. Généralement les codages effectués sur les données complexes sont faits de sorte que les tableaux qui en résultent soient assimilés à des points plongés dans des espaces multidimensionnels et, très souvent, dans des espaces vectoriels. Dans ce cadre, l'outillage mathématique, issu notamment de l'algèbre linéaire et de la programmation mathématique, permet de traiter ces données. Or ce codage n'est pas toujours possible notamment en présence de données hétérogènes qualitatives, quantitatives, symboliques ou de données non structurées comme des graphes. Comment alors fouiller ces ensembles de données sans trop d'altérations ? autrement dit, quel codage faut-il adopter ? comment construire des indices de proximités qui sont des outils indispensables pour les tâches d'apprentissage notamment non supervisé ? Quelles propriétés mathématiques résultent de ces choix pour savoir si oui ou non des algorithmes classiques de fouille peuvent être utilisés ?
- Différence de niveau sémantique. Les données qui se rapportent à des objets complexes ne se situent pas toujours, toutes, sur le même niveau d'abstraction. Ce phénomène bien connu dans le domaine de la représentation des connaissances et notamment dans les ontologies prend une nouvelle dimension encore plus difficile à maîtriser. Par exemple, un compte-rendu médical écrit par un médecin sur un patient peut être le résultat d'une interprétation de clichés et d'examens biologiques. Par conséquent, le niveau sémantique du texte peut être différent de celui des images car il contient une partie de la sémantique de l'image. Dans ce cas, les attributs issus des comptes-

rendus textuels auront du mal à être alignés sur ceux issus des images radiologiques ou des examens biologiques. Comment alors intégrer ces niveaux d'abstraction sémantique pour ensuite pouvoir décrire des patients ou les comparer ? Le texte compte-rendu médical par exemple devrait-il être vu comme un subsumant des images et des données biologiques ? difficile d'y répondre.

- Fusion des données et intégration des connaissances du domaine. Souvent, dans nos processus d'interprétation des situations qui nous entourent, comme être humains, nous pouvons mieux inférer grâce à une contextualisation des données que nous recevons par rapport à d'autres qui leurs sont liées indirectement. Par exemple, les phénomènes de pollution dans les grands centres urbains ont des causes dues à l'activité humaine, mais cette même activité humaine ne produit pas les mêmes effets selon la région. Par exemple, un contexte climatique ou géographique particulier peut affecter significativement et de manière différente le phénomène de pollution. Dans un autre cadre, la santé, où les connaissances formelles servent à renforcer ou à rejeter certaines hypothèses de diagnostic issues de l'observation, les connaissances du domaine, formalisées ou non jouent un rôle crucial. Ce procédé de contextualisation est particulièrement développé dans la fouille de données textuelles et est généralement destiné à améliorer, par exemple, la désambigüisation.

2.4 Défis technologiques de la prise de décision dans la cadre de la FDC

Les défis scientifiques et technologiques sont étroitement liés. Parmi les défis technologiques identifiés, nous pensons apporter les nouvelles contributions suivantes.

- *Passage à l'échelle (scalabilité).* Nous pouvons disposer d'une solution formelle et même opérationnelle sans pour autant être en mesure de l'utiliser sur des corpus de données réelles, soit pour des raisons de temps de calcul, soit pour des raisons d'espace mémoire insuffisant. Par exemple, dans le cas de la fouille dans les Dossiers Médicaux Personnalisés (DMP), la classification d'une population de patients s'avère quasi impossible de façon directe si l'on prend en compte la totalité des informations. Outre le problème de l'alignement sémantique des données-attributs, le mélange de types de données, la dimension élevée de l'espace de représentation, le grand nombre d'observations, notamment, rendent cette opération impossible à réaliser en pratique de façon directe. On peut alors s'interroger sur les aspects méthodologiques permettant de développer des algorithmes incrémentaux appropriés, par exemple, ou encore sur la meilleure façon d'exploiter les ressources physiques des machines. Les grilles de calcul (*grid computing*) est l'une des réponses possibles. Une autre approche consisterait à travailler sur des

vues partielles des objets, comme en fouille de données multi-tables, et à introduire de nouvelles technologies de fouille distribuée.

- *Processus de fouille de données destiné à produire des connaissances à partir de données en perpétuelle évolution.* Dans ce contexte, le processus de fouille doit alors être continu pour identifier à temps les éventuelles modifications majeures au niveau des connaissances qui pourraient survenir sur un phénomène modélisé. Comment alors assurer le couplage fort entre inférence sur des cas et amélioration incrémentale des connaissances qui sous-tendent cette inférence ?

2.5 Projet scientifique

D'une façon générale, notre projet porte sur une méthodologie pour une approche dynamique et intégrée de la fouille des données complexes (hétérogènes, distribuées, volumineuses etc.) et le déploiement des connaissances dans des applications réelles de prise de décision. Le plus souvent, les défis technologiques et scientifiques sont étroitement liés. Au sein des équipes qui forment le laboratoire ERIC, nous avons toujours été soucieux de ce va-et-vient entre le scientifique et le technologique. Nous avons souhaité fédérer nos activités de recherche, qu'elles soient, théoriques et/ou appliquées, autour d'une méthodologie intégrée dont l'objectif est de déboucher sur des méthodes et des outils informatiques permettant d'assurer :

- L'acquisition et la gestion des données complexes avec l'ensemble des problèmes sous-jacents de stockage, d'accès, de mise à jour, de sécurité, d'anonymat le cas échéant etc.
- L'organisation et la représentation de ces données pour assurer une fouille efficace avec toutes les questions liées au codage, à l'indexation, à la mise en forme etc.
- Le traitement des données complexes par des algorithmes de fouille capables de couvrir les besoins en description, en structuration ou en explication-prédiction avec un intérêt majeur pour les questions liées à la prise en compte des phénomènes non linéaires ;
- L'évaluation et la validation des connaissances (modèles) produites que cela soit en termes de reproductibilité des modèles (statistique) ou de cohérence des modèles (logique).
- L'intégration automatique des connaissances dans un système à base de connaissances, lequel pouvant également recevoir des connaissances en provenance de l'expert. Cela conduit à imaginer des systèmes plus ouverts pour la gestion des connaissances, dotés de formalismes de représentation unifiés.
- La mise au point de moteur d'inférence capable de fournir une réponse à une requête d'utilisateur. La requête peut se faire dans un cadre de recherche d'information ou d'une prise de décision. Il convient en outre d'imaginer des stratégies d'inférence combinant le symbolique, le numérique et le multiexpert.

Le spectre de compétences des chercheurs d'ERIC (entrepôts de données, apprentissage, statistique, décision, systèmes distribués etc.) permet en effet d'affecter nos ressources pour couvrir de façon transversale les différents besoins évoqués pour notre projet scientifique.

Les thèses en cours au laboratoire illustrent de manière concrète les projets de recherche à court et moyen terme et, qui tous, d'une façon ou d'une autre, trouvent leur place dans les perspectives de recherche du laboratoire telles qu'elles sont esquissées. Par conséquent, nous allons surtout définir les grands axes d'orientation majeurs pour le futur : moyen et long terme.

Pour répondre à ces besoins et s'attaquer à ces défis, le laboratoire a décidé de mettre en place une organisation scientifique structurée selon des axes des thèmes de recherche et des projets. A terme, nous envisageons de structurer l'activité du laboratoire en Equipe-Projet. Le travail de réflexion sur les projets et leur structuration autour de groupes d'enseignants chercheurs est en cours. Mais, dans l'immédiat, nous détaillons ci-dessous la structuration de l'activité telle qu'elle va se présenter pour ces deux à trois prochaines années :

La recherche d'ERIC s'effectuera autour de trois axes théoriques fondamentaux et aura pour cadre deux domaines d'application transversaux que nous présentons brièvement ci-dessous et qui sont décrit de façon plus détaillée en annexe (section 5).

2.5.1 Axes théoriques fondamentaux

Il s'agit des travaux relevant des spécialités au sens des disciplines représentées au sein du laboratoire, à savoir, l'informatique et les mathématiques appliquées. Dans ce cadre, les principaux axes sont :

- ENtrepôts et Analyse en ligne de Données Complexes (ENA-DC) : Il s'agira notamment de poursuivre les travaux en cours dans le domaine de l'intégration, de la modélisation multidimensionnelle et de la navigation OLAP dans les entrepôts de données complexes.
- FOuille de Données et Apprentissage (FODA) : il s'agira de poursuivre des travaux à la fois théoriques et algorithmiques pour analyser de grands corpus de données complexes, les résumer ou les modéliser dans un but de prédiction.
- DECision et COmplexité (DECCO) : il s'agira d'appréhender le rapport aux réalités de terrain et notamment d'élaborer de nouveaux modèles dans des environnements où de multiples agents sont en interaction au sein du processus de décision.

2.5.2 Domaines d'application transversaux

Les deux domaines d'application regroupent des projets fédérateurs faisant intervenir des compétences pluridisciplinaires. Ces projets, qui sont généralement menés avec des partenaires dans les domaines des SHS ou de la Santé, ont une durée de vie de 3 à 5 ans :

- Santé et Environnemant : le laboratoire ERIC compte capitaliser sur son expertise et ses relations avec le monde de la santé. Les activités du laboratoire porteront sur la modélisation des processus de décision dans le secteur de la santé, tant dans le domaine médical pur (clinique, biologique, radiologique), que le domaine médico-économique.
- Sciences Humaines et Sociales (SHS) : l'autre particularité de l'environnement d'ERIC est qu'il relève des SHS. Nous pensons que cette proximité est très avantageuse pour nous comme pour les SHS et qu'elle conduira nécessairement à une fertilisation croisée.

2.5.3 Implication technologique

Le laboratoire ERIC maintient une tradition de diffusion de logiciels libres dans les domaines de la fouille de données complexes ou de la décision.

2.5.4 Synthèse

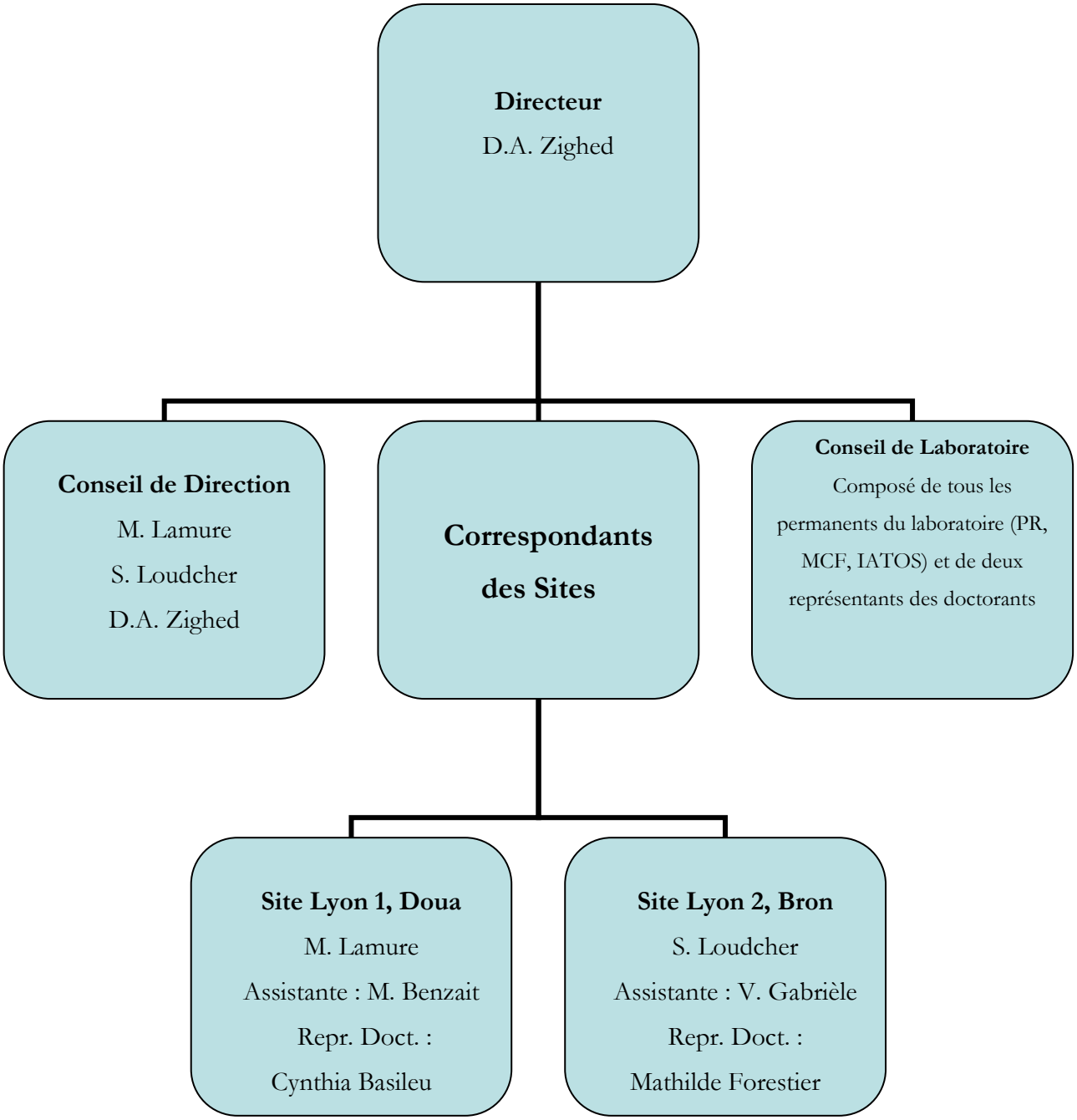
Le tableau ci-dessous résume l'organisation scientifique proposée.

			Axes de recherche fondamentaux		
			ENA DC	**FODA**	**DECCO**
			Resp. Jérôme Darmont	Resp. Stéphane Lallich	Resp. Stéphane Bonnevay
			- Modélisation des objets complexes - Analyse en ligne des objets complexes	- Caractérisation des espaces de représentation - Analyse et évaluation des structures topologiques sous-jacentes	- Modélisation de l'agrégation des préférences - Modélisation des dynamiques complexes
Domaines d'application transversaux	**Santé et Environnement**	Resp. Michel Lamure	Base de données médico-économiques : structuration en entrepôts, navigation	Extraction et représentation des connaissances médicales, aide au diagnostic, visualisation de trajectoires patients,...	Modélisation des systèmes complexes, décision et simulation par rapport au fonctionnement des systèmes de santé, la gestion des risques sanitaires, ...
			bases hétérogènes et complexes sur des données de climatologie, de santé, de sociologie, Géographiques,...	Navigation dans les données, visualisation interactive des données, construction d'une ontologie de la pluridisciplinarité	Modélisation des systèmes complexes, gestion des risques pour la mise en place d'un système d'alerte. Systèmes multiagents,...
	SHS	Resp. Jean-Hugues Chauchat	Base de données historiques : structuration en entrepôts, navigation	Visualisation interactive des données complexes, Apprentissage semi-supervisé	Interrogation des données

3 ORGANISATION DU LABORATOIRE

3.1 Gouvernance

Le laboratoire s'appuiera sur l'organigramme suivant :



Liste des enseignants-chercheurs et chercheurs associés par site	
Site Lyon 1	**Site Lyon 2**
- Stéphane Bonnevay, MCF HDR	- Rafik Abdesselam, MCF HDR
- Ahmed Bounekkar, MCF	- Fadila Bentayeb, MCF
- Denis Bourgeois, Pr CE (associé)	- Omar Boussaid, Pr
- Denis Clot, MCF	- Jean-Hugues Chauchat, Pr1
- Michel Dubois (associé)	- Jérôme Darmont, Pr
- Claude Dussart, Pharm. HDR (associé)	- Cécile Favre, MCF
- Bruno Fantino, Med HDR (associé)	- Nouria Harbi, MCF
- Gérald Gavin, MCF	- Stéphane Lallich, Pr
- Nadia Kabachi, MCF	- Sabine Loudcher, MCF
- Michel Lamure, Pr 1	- Ricco Rakotomalala, MCF
- Fabien Rico, MCF	- Julien Velcin, MCF
- Carole Siani, MCF HDR	- Jacques Viallaneix, MCF
- Mondher Toumi, Pr Chaire (associé)	- Abdelkader Zighed, Pr1

3.1.1 Correspondants de sites

Sur chaque site, un correspondant assure la représentation officielle du laboratoire auprès de l'établissement concerné. Le correspondant de chaque site assure également la gestion administrative des enseignants chercheurs de son site. Une assistance locale à chaque site aidera le correspondant dans les tâches administratives, financières et d'échanges avec les autres sites du laboratoire.

3.1.2 Conseil de laboratoire

Le conseil de laboratoire est l'instance de validation des choix stratégiques de l'unité. Il valide le budget, le classement des candidats aux allocations de recherche, l'attribution de moyens aux différents enseignants chercheurs, le choix des projets portés par le laboratoire et toutes les actions relatives à politique scientifique d'animation. Le conseil de laboratoire est composé de tous les enseignants chercheurs et personnels IATOS permanents ainsi que de trois représentants élus chaque année parmi les doctorants.

Si la taille du laboratoire venait à croître de manière significative, un mode de représentation par élus serait mis en place.

Le conseil de laboratoire délibère à bulletin secret, sauf accord unanime, à la majorité des présents. Le vote par procuration est valable.

Il se réunit de façon ordinaire tous les mois et peut être convoqué par le conseil de direction à tout moment.

Les réunions sont animées par le directeur ou l'un des membres du conseil de direction si le directeur est absent. L'ordre du jour est communiqué à l'avance et peut être complété par tout enseignant chercheur. Les réunions donnent lieu à un compte rendu qui est diffusé auprès de tous les membres du laboratoire.

3.1.3 Conseil de direction

Il est composé des correspondants de site et présidé par le directeur du laboratoire. Son rôle est de veiller à l'exécution des décisions du conseil de laboratoire et de coordonner les activités des deux sites. Il prépare les choix stratégiques à soumettre au conseil de laboratoire et assure le suivi du budget. Les membres du conseil de direction jouissent d'une délégation de signature du directeur pour toutes les affaires propres à chaque site.

3.1.4 Direction

La direction est assurée par un directeur disposant d'une assistance administrative et financière. La direction assure le suivi des projets scientifiques et le déroulement global de la vie du laboratoire : personnel, finance, etc. En cas d'absence du directeur, l'un des membres du conseil de direction assumera l'ensemble des tâches jusqu'à son retour.

3.2 Organisation Scientifique

La recherche est organisée autour de trois axes fondamentaux de recherche (ENA-DC, FODA, DECCO) et deux domaines d'application (Santé et Environnement, SHS). Afin d'assurer une meilleure interaction entre chercheurs nous avons ventilé les membres d'ERIC sur l'ensemble des trois axes et des deux domaines. Chaque chercheur se positionne dans un axe de recherche principal où se concentra la majeure partie de son activité et un axe secondaire enfin d'éviter les cloisonnements et assurer une meilleure circulation des idées. Et nous avons procédé de même en ce qui concerne les domaines d'application. Le tableau suivant indique pour chaque chercheur son positionnement.

Répartition des chercheurs selon les axes théoriques et les domaines transversaux			
Enseignants-Chercheurs	**Axes de rattachement**		**Domaines d'application**
	Principal	Secondaire	
Adbesslam Rafik	FODA	DECCO	Santé et Environnement
Bentayeb Fadila	ENA-DC	FODA	SHS
Bonnevay Stéphane	DECCO	FODA	Santé et Environnement
Bounnekar Ahmed	DECCO	FODA	Santé et Environnement
Bourgeois Denis (associé)			Santé et Environnement
Boussaid Omar	ENA-DC	FODA	SHS
Chauchat Jean-Hugues	FODA	DECCO	SHS
Clot Denis	FODA	DECCO	Santé et Environnement
Darmont Jérôme	ENA-DC	DECCO	SHS
Dubois Michel (associé)			Santé et Environnement
Dussart Claude (associé)			Santé et Environnement
Fantino Bruno (associé)			Santé et Environnement
Favre Cécile	ENA-DC	FODA	SHS
Gavin Gérald	FODA	ENA-DC	Santé et Environnement
Harbi Nouria	ENA-DC		SHS
Kabachi Nadia	DECCO	ENA-DC	Santé et Environnement
Lallich Stephane	FODA	DECCO	SHS
Lamure Michel	DECCO	FODA	Santé et Environnement
Loudcher Sabine	ENA-DC	FODA	SHS
Rakotomalala Ricco	FODA		SHS
Rico Fabien	DECCO		Santé et Environnement
Siani Carole	DECCO	FODA	Santé et Environnement
Toumi Mondher (associé)			Santé et Environnement
Velcin Julien	FODA	ENA-DC	SHS
Viallaneix Jacques			SHS
Zighed Abdelkader	FODA	DECCO	SHS

Nous pensons que cette organisation tant administrative que scientifique aura un impact fort sur :

- l'émergence de sujets nouveaux,
- la prise de risque qui se trouve encouragée grâce au double rattachement,
- l'adaptation aux évolutions scientifiques et techniques dans le contexte local, national, européen et international,
- la réflexion sur les créneaux porteurs,
- L'évolution de l'unité à 4 ans et 8 ans vers une reconnaissance par une structure nationale comme l'INRIA.

4 ADEQUATION DES MOYENS HUMAINS ET FINANCIERS DE L'UNITE AVEC LE PROJET

4.1 Politique d'animation scientifique

- Publications Objectifs/chercheur : actuellement, les supports sélectifs et à publication rapide sont très privilégiés par les chercheurs et notamment les doctorants. Ce mode de publication sera encouragé. Cependant, un soutien financier, spécifique pour les publications dans les journaux de premier plan, sera mis en place. L'objectif est d'atteindre un niveau moyen d'une publication dans un journal par an et par permanent.

- Valorisation scientifique (Conférences, collaborations industrielles, incubation, spin off…) : l'équipe continuera sa politique d'animation scientifique et favorisera l'ouverture au monde industriel ; elle encouragera particulièrement le référencement (dépôt) de logiciels.

- Echange nationaux et internationaux (accueil de chercheurs étrangers, semestres sabbatiques) : grâce à la forte synergie entre enseignement et recherche, le laboratoire poursuivra sa politique d'accueil et d'échange de chercheurs avec des partenaires étrangers. Les maîtres de conférences du laboratoire seront particulièrement encouragés à effectuer des séjours de 1 à 6 mois dans des laboratoires étrangers.

- Formation par la Recherche : les stages de recherche seront renforcés dès la Licence afin de faire découvrir assez tôt le métier de chercheur aux étudiants.

4.2 Analyse prospective à moyen et long terme des moyens et des compétences

L'état actuel des forces en enseignants chercheurs montre un bon rapport de 1/3 entre professeurs et MCF. Ce rapport est conforme à ce qui existe dans les grandes équipes. Cependant, la création de postes en informatique doit être maintenue pour accompagner la demande croissante en matière d'enseignement et d'encadrement de la recherche.

4.3 Politique de construction de partenariats

La politique de valorisation de la recherche proposée permettra des opportunités nouvelles pour la création de partenariats tant au niveau de la recherche appliquée avec des entreprises qu'au niveau de la recherche académique. Seront privilégiés les partenariats internationaux qui pourraient accroitre la visibilité du laboratoire. Cette politique pourra également s'appuyer sur les partenariats pédagogiques qui sont souvent le point d'amorçage de collaborations en recherche.

4.4 Schéma de financement du projet

L'activité du laboratoire sera financée par les ressources attribuées par les instances de tutelle (20% à 30%) le reste étant financé dans le cadre de partenariats privés ou de projets à appels d'offres nationaux, européens ou internationaux.

4.5 Capacité de l'unité à valoriser ses travaux de recherche

La présence des chercheurs d'ERIC au niveau des comités de programmes des principales conférences du domaine, les liens que nous avons avec le tissu industriel à travers les stages d'étudiants et les diverses collaboration académiques et industrielles nous donnent les moyens de diffuser et de faire connaître les travaux et l'expertise d'ERIC.

4.6 Implication de l'équipe en matière de diffusion de l'information scientifique et technique

Les enseignants chercheurs d'ERIC continueront à s'impliquer dans les éditions d'ouvrages, de logiciels, de supports de cours en ligne etc. Avec notamment :

- Edition de la Revue RNTI, Cepaduès (D.A. Zighed, Co-directeur)
- Direction de collection Santé et Systémique, Hermès (Michel Lamure)
- Présidence de l'Association Internationale Francophone d'Extraction et de Gestion des Connaissances (EGC) (D.A. Zighed, Président)
- Présidence de l'association Pretopologics (Michel Lamure)
- Responsable de Groupe SCDD ROADEF du GDR MACS (Stéphane Bonnevay)
- Responsable de groupe de travail Fouille de données complexes (Omar Boussaid)

- Responsable de workshop « Mining Complex Data » (D.A. Zighed)
- Responsable de la conférence EDA (Fadila Bentayeb, Omar Boussaid, Jérôme Darmont, Nouria Harbi, Sabine Loudcher)
- De nombreuses autres responsabilités au niveau de conférences internationales et nationales, de comités éditoriaux etc.

5 ANNEXE : DESCRIPTIONS DES AXES THEORIQUES ET DOMAINES TRANSVERSAUX DE RECHERCHE

5.1 ENtrepôts et Analyse en ligne de Données Complexes (ENA-DC)

Notre objectif de mener des recherches sur l'entreposage et l'analyse en ligne de données complexes. En effet, dans le processus d'entreposage de données, lorsque les données sont complexes (multiformats, multistructures multisources, multimodales, multiversions…), les problématiques d'intégration, de modélisation et d'analyse de données nécessitent des méthodes adaptées.

Face aux différents problèmes soulevés par l'entreposage de données complexes, notre recherche s'articulera d'une part autour de trois axes majeurs qui sont l'intégration de données, la représentation d'objets complexes et l'analyse en ligne et d'autre part autour de deux axes transversaux qui sont la sécurité et la personnalisation dans les entrepôts de données.

1. Intégration de données

Il s'agit ici de repenser le problème de l'intégration classique de données dans le contexte de données complexes. En effet, il est non seulement nécessaire d'intégrer les données brutes issues de différentes sources, mais aussi de sélectionner, de générer et d'intégrer des descripteurs de plus haut niveau (par exemple, des histogrammes de couleur ou de texture pour une image) ainsi que des descripteurs sémantiques (comme des mots clés ou des concepts) qui permettront plus tard d'améliorer le stockage et l'exploitation des données complexes. Par exemple, les sources de données complexes (comme le Web) ne sont pas systématiquement structurées. Elles le sont souvent même peu ou pas du tout. La description de ces sources est nécessaire pour un dispositif d'intégration. Outre une description structurelle usuelle, celle de leur sémantique est de surcroît indispensable, ce point constituant actuellement un verrou scientifique.

L'évolution du processus d'ETL (Extract-Transform-Load) devrait assurer une centralisation des données intégrées ou la possibilité d'intégrer celles-ci à la demande. Le rafraîchissement des données reste un problème ouvert. Les aspects temps réel et de flux de données sont des problématiques à reconsidérer dans le cadre des données complexes. L'historisation des données est un problème à reposer également. Car l'élargissement des capacités de l'analyse en ligne devrait permettre de relâcher cette contrainte.

2. Représentation d'objets complexes

Un objet complexe peut être considéré comme un agrégat hétérogène de données qui, une fois réunies, forment une unité sémantique. Ainsi, pour représenter un objet complexe, il est nécessaire de représenter, en plus des descripteurs de bas niveau, des connaissances et des métadonnées associées. Le challenge est alors de pouvoir trouver de nouveaux modèles, pour représenter les objets complexes, orientés analyse. Les concepts multidimensionnels (fait, dimension, hiérarchie, niveau, attribut…) sont à redéfinir à l'aide des objets complexes. Cela passera par la définition de nouvelles métriques afin d'agréger ou de comparer les objets complexes entre eux. En effet, les métriques quantitatives sont insuffisantes d'où la nécessité de définir des métriques qualitatives ou sémantiques.

Un autre problème à traiter est la représentation des connaissances associées à ces objets, mais également à l'ensemble du processus d'entreposage. En première analyse, des langages dérivés d'XML comme RDF et OWL devraient permettre de représenter aux niveaux logique et physique à la fois les objets complexes et leurs métadonnées et connaissances associées. Toutefois, ces modes de représentation impliquent un stockage de données XML dont la performance, que ce soit dans des systèmes compatibles XML ou natifs, est actuellement encore limitée. Nous mènerons donc des travaux de recherche afin de l'optimiser.

3. Analyse en ligne

L'objectif de cet axe de recherche réside dans l'élargissement des capacités de l'analyse en ligne (OLAP : On-Line Analytical Processing). Celle-ci doit permettre de faire plus que de naviguer dans les objets complexes stockés dans l'entrepôt. Trois volets doivent être approfondis : l'enrichissement de la navigation, le besoin de résumer (agréger) les objets complexes et le besoin de les expliquer. L'un des enjeux réside alors dans le fait de pouvoir garder l'esprit de la navigation OLAP dans le cadre de l'analyse des objets complexes. Et au-delà, nous souhaitons enrichir l'OLAP avec des opérateurs de prédiction et d'explication grâce au couplage avec les techniques de fouille de données et en particulier dans les documents semi-structurés. De même coupler OLAP avec des techniques de recherche d'information (RI) devra permettre d'enrichir la navigation par des analyses de types top-k mots clés ou autres descripteurs sémantiques, attribution d'auteurs à des documents, analyse lexicale, analogies entre objets, etc.

4. Sécurité et qualité des données

La sécurité et la qualité constituent un axe transversal et nécessaire aux phases d'intégration, de représentation et d'analyse des objets complexes. Les enjeux se situent à la fois au niveau de l'intégrité des données, de leur qualité, de leur confidentialité et du contrôle d'accès. La stratégie développée pourrait se fonder sur une forme de veille (monitoring) avec un système d'alertes renforçant ainsi les détections d'intrusion et la prévention d'actes malveillants ou accidentels. Ce point est particulièrement pertinent dans le contexte des données sous forme de flux (data streams). Il faudra formaliser ces

aspects de sécurité dans le modèle générique de l'entrepôt et proposer des mécanismes de contrôle d'accès adéquats aux représentations multidimensionnelles pour assurer la confidentialité des données sensibles.

5. Personnalisation

La personnalisation constitue une problématique nouvelle dans les entrepôts de données qui pose plusieurs enjeux peu ou pas étudiés. Une plus grande interaction de l'usager avec le système décisionnel permettrait d'envisager des bénéfices à deux niveaux :

- du point de vue système, la connaissance accrue de l'utilisateur ou du groupe d'utilisateurs doit pouvoir servir à mieux paramétrer celui-ci, et par conséquent, doit permettre un fonctionnement plus proche des utilisateurs. Cela passe par la construction de profils utilisateurs qui sont principalement constitués par des préférences utilisateurs sur à la fois les données et les structures;
- du point de vue utilisateur, un système mieux adapté doit permettre une réduction des efforts nécessaires pour accéder, manipuler et structurer une information pertinente afin de faciliter davantage le processus décisionnel qui en découle. Cela inclut à la fois les problèmes de recommandation de nouvelles analyses en utilisant par exemple des techniques de fouille ou de RI et les problématiques d'autoadministration des entrepôts de données, qui entrent dans les recherches actuelles sur les systèmes autoadaptables.

5.2 FOuille de Données et Apprentissage (FODA)

Les données complexes (volumineuses, distribuées, hétérogènes, évolutives, non structurées) sont aujourd'hui le principal véhicule d'information. Nous disposons de volumineuses bases de données complexes dont la croissance est exponentielle mais dont la valorisation reste encore très faible.

Le défi qu'il faut relever est de tirer profit, dans tous les sens du terme, de ces données : recherche d'information, extraction de connaissances, création de valeurs économiques, … . La fouille de données complexes tente de répondre à ce besoin. Dans le cadre de ce projet scientifique, l'équipe propose de concevoir de nouveaux cadres méthodologiques et de développer des outils mathématiques et informatiques innovants pour aider à la structuration de ces données complexes, les analyser en vue d'extraire des connaissances ou des informations non accessibles par des moyens classiques. Une des manières d'aborder ces problèmes est l'apprentissage qui est un des domaines de recherche privilégié de l'équipe.

1. Les défis scientifiques

Les défis scientifiques que soulèvent ces particularités des données complexes sont multiples. Parmi ceux que l'on peut assez facilement identifier, on peut lister :

- La grande dimensionnalité

Les attributs issus des images, des textes et des autres modalités peuvent atteindre plusieurs milliers de variables, voir par exemple dans le cadre de l'étude du génome, plusieurs millions de variables. Comment évaluer la pertinence de cet espace de représentation par rapport aux tâches que l'on souhaite effectuer comme l'apprentissage supervisé ou la classification ? comment réduire l'espace si tant est que cela soit possible ?

- L'absence de structure mathématique

Généralement, les codages effectués sur les données complexes font en sorte que les tableaux qui en résultent sont assimilés à des points plongés dans des espaces multidimensionnels et, dans le meilleur des cas, dans des espaces vectoriels. Dans ce cadre, l'outillage mathématique, issu notamment de l'algèbre linéaire et de la programmation mathématique, permet de traiter ces données. Or ce codage n'est pas toujours possible notamment en présence de données hétérogènes (qualitatives, quantitatives, symboliques, …). Comment alors analyser ces ensembles de données ? Quel codage faut-il adopter ? Comment construire des indices de proximités qui sont des outils indispensables pour les taches d'apprentissage notamment non supervisé ? Quelles propriétés mathématiques résultent de ces choix pour savoir si oui ou non des algorithmes classiques de fouille peuvent être utilisés ?

- La différence de niveau sémantique

Les données qui se rapportent à un objet complexe ne se situent pas toujours toutes sur le même niveau d'abstraction. Ce phénomène bien connu dans le domaine de la représentation des connaissances et notamment dans les ontologies prend une nouvelle dimension encore plus difficile à maîtriser. Par exemple, un compte-rendu médical écrit par un médecin sur un patient peut être le résultat d'une interprétation de clichés et d'examens biologiques. Par conséquent, le niveau sémantique du texte est différent de celui des images. Dans ce cas, les attributs issus des comptes-rendus textuels auront du mal à être alignés sur ceux issus des images radiologiques ou des examens biologiques. Comment alors intégrer ces niveaux sémantiques pour ensuite pouvoir décrire des patients ou les comparer ? Le texte devrait il être vu comme un subsumant des images et des données biologiques ? Il est difficile d'y répondre promptement.

- La fusion des données et l'intégration des connaissances du domaine

Souvent, dans les processus d'interprétation des situations qui l'entourent, l'être humain peut mieux inférer grâce à une contextualisation des données qu'il reçoit par rapport à d'autres qui leurs sont liées indirectement. Comment intégrer ce processus d'interprétation complexe dans les modèles d'analyse de données ? Comment améliorer les méthodes d'apprentissage ?

2. Les axes de recherche

Les axes de recherche de l'équipe sur la problématique de la « fouille de données et apprentissage » couvrent une partie des différents défis scientifiques évoqués ci-dessus. Ils portent sur différents points du processus de fouille de données :

2.1. Recherche d'un bon espace de représentation : Pour toutes les taches de fouille, qu'elle relève de l'apprentissage, de la recherche d'information ou de la visualisation, la recherche d'un bon espace de représentation est une tâche cruciale. On lui associe généralement un ensemble de méthodes et de techniques dites de sélection et de construction d'attributs, d'échantillonnage d'instance, de recodage etc. La recherche d'un bon espace suppose d'une part, de disposer d'une fonction d'évaluation de la qualité de l'espace de représentation, et, d'autre part, d'un algorithme de recherche d'une solution à partir d'un état initial. De nombreuses méthodes, statistiques ou heuristiques, existent pour les situations où les données sont de même nature, par exemple toutes quantitatives. Dans le cas de données hétérogènes (quantitatives, qualitatives, temporelles, spatiales etc.) des difficultés diverses apparaissent. Quand à ces difficultés s'ajoutent la dimensionnalités, cela devient parfois une barrière pour la fouille de données. Dans ce cadre, nous comptons explorer différents pistes que nous citons en vrac :

- l'extraction des relations « causales » entre les variables d'étude et la variable à prédire ; en particulier par la construction d'une « Markov Boundary » de la variable à prédire, qui correspond à un sous ensemble minimal de variables rendant le reste des variables indépendant de la variable à prédire. L'équipe se propose de résoudre certains problèmes associés à ces techniques de sélection comme, d'un point de vue algorithmique, le passage à l'échelle ou l'étude, plus générale, de la causalité du point de vue théorique et empirique.
- La sélection prétopologique de variables, qui consiste à construire itérativement un ensemble de variables à sélectionner dans des bases en se dotant à la fois de critères d'acceptation et de critères de rejet.
- La recherche d'un codage à variance minimale dans le cadre de la discrétisation d'attributs continus et sa comparaison aux méthodes de codages flous,
- Passage de d'une représentation topologique à une représentation vectorielle et réciproquement. Il s'agit ici de trouver un moyen de contourner les limitations des méthodes basées sur le Mutidimensional scaling (MDS).
- …

2.2. Questions liées au algorithmes de fouilles (Description, Classification et Prédiction) : En Fouille de Données, il existe généralement trois grandes familles de méthodes :

- celles dont le but est de visualiser ou décrire les ressemblances entre individus et/ou variables comme les méthodes factorielles,
- celles dont le but est de synthétiser des grands tableaux (individus et/ou variables) comme les méthodes de classification,
- enfin, celles dont le but est d'expliquer et/ou de prédire une variable par la connaissance d'autres variables comme les méthodes de classement.

Dans ce contexte, de nombreux problèmes restent encore posés comme :

- Le mélange d'attributs hétérogènes dans des méthodes de type analyses factorielles (ACP, AFC, AFD, etc.)
- La prise en compte dans les méthodes d'apprentissage supervisé des problèmes liés à l'asymétrie des coûts et des populations dans les échantillons d'apprentissage,
- La mesure de la séparabilité des classes en apprentissage supervisé
- La réduction de la dimensionnalité de l'espace de représentation que cela soit en apprentissage non supervisé ou supervisé
- La prise en compte des connaissances du domaine et/ou de l'expert dans les problèmes d'apprentissage notamment semi-supervisé
- La construction de mesures de similarités sur des objets complexes comme les graphes, les images, les textes et/ou toutes combinaisons de ces derniers avec ou sans caractère spatio-temporel.

Dans les futurs travaux d'ERIC, nous allons aborder ces différentes questions selon différents points de vue en exploitant notamment la dualité entre l'espace métrique et l'espace topologique. Cette réflexion est en plein essor dans le contexte de l'apprentissage topologique (Topological learning). Ces recherches devraient déboucher sur de nouvelles méthodes de fouille de données.

2.3. Évaluation de la pertinence des résultats en fouille de données : Les sorties des algorithmes de fouille de données doivent nécessairement être évaluées avant d'être soumises à l'usager. De nombreuses question restent encore à approfondir et notamment :

- Unifier les mesures de qualité proposées dans les différents domaines,
- l'étude des propriétés algorithmiques des mesures d'intérêt des règles pour accélérer certains processus de fouille comme la recherche de règles d'association.

D'autres questions transversales en fouille de données comme la réduction de l'erreur d'apprentissage par des stratégies de renforcement ou de pondération comme Adaboost seront également abordées.

5.3 DECision et COmplexité (DECCO)

Dans cet axe, nous nous plaçons principalement en aval des recherches développées dans les autres axes. En effet, après l'extraction d'information issue de la fouille de données, des problématiques, à la fois d'agrégation de ces informations, mais aussi de décision multicritère ou collective, sont posées : comment agréger des données souvent hétérogènes ou prendre une décision à partir de différents critères ou différentes opinions de décideurs. La première thématique de recherche de cet axe se propose d'étudier le **processus de décision** dans son ensemble, afin de développer de nouveaux modèles mathématiques pour formaliser la **prise de décision en environnement complexe**. Une deuxième thématique de recherche de l'axe DECCO est centrée sur l'étude des objets dits « complexes » et plus particulièrement sur l'étude des **réseaux complexes**. L'objectif est de développer de nouveaux modèles d'étude de ces réseaux en intégrant la complexité de ces objets (structure dynamique, hétérogénéité, …).

La principale thématique de recherche de cet axe s'intéresse donc à la prise de décision et au processus de décision. La décision doit être étudiée comme un processus délivrant un résultat, fondé sur l'appréciation et l'évaluation de différentes informations, de différents critères ou de différentes stratégies par un ou plusieurs acteurs. Dans le cas d'acteurs, des hypothèses sont généralement placées à la fois sur le comportement des différents acteurs et sur l'environnement dans lequel ces acteurs sont amenés à se prononcer quant à la décision finale. Il s'agit d'un objet scientifique qui a été très étudié, pour lequel on sait déjà qu'il n'existe pas de solution « optimale », excepté dans quelques situations spécifiques assez rares. Les solutions proposées relèvent donc d'un processus d'arbitrage entre des avantages et des inconvénients qu'elles sont susceptibles de posséder plutôt que d'un processus d'optimisation. Les hypothèses placées sur les comportements des acteurs, sur le niveau d'information qu'ils possèdent jouent un grand rôle dans l'obtention des solutions et sur les propriétés dont elles sont dotées. Différentes approches méthodologiques s'intéressent à produire de telles solutions. Il existe, par exemple, les approches fondées sur le concept d'utilité très commode pour traduire simplement des préférences que pourraient avoir les acteurs sur les différentes possibilités qui s'offrent à eux. Ces approches ont donné lieu à la publication de très nombreux travaux relevant de l'économie, des mathématiques, et plus récemment, de l'informatique. Ces approches souffrent cependant souvent de différents inconvénients : des hypothèses fortes placées sur les préférences des acteurs qui induisent très rapidement le résultat, des hypothèses parfois invérifiables en pratique, des faiblesses de formalisation pour des cas réels, une relative pauvreté dans la modélisation des comportements, … . Clairement, on peut en attribuer, en partie, la cause au fait que la plupart des modèles proposés relèvent d'une approche cartésienne trop réductrice pour étudier un objet complexe. Nous proposons donc de

passer à une approche complexe de la décision en nous fondant sur des théories et des modèles intégrant la complexité de l'objet et utilisant des méthodologies nouvelles tant du point de vue mathématique (prétopologie, ensembles aléatoires, floue, …) que du point de vue informatique (simulation à base d'agents). L'objectif est aussi de développer des modèles intégrant davantage les comportements humains dans la prise de décision ; il est donc clair que ces modèles doivent mettre en jeu diverses disciplines : mathématiques, informatique, mais aussi les sciences humaines et sociales qui doivent être placées au centre de la réflexion.

Nous avons publié des modèles de débats entre décideurs et développé des protocoles de négociation entre ces décideurs. Nos travaux utilisent des modèles flous (basés sur l'intégrale de Sipos et de Choquet pour modéliser les changements de convictions et de préférences entre décideurs) et des modèles issus de la théorie des jeux coopératifs. Des simulations à base d'agents nous ont permis de définir des propriétés liées au processus de décision et de valider nos protocoles de négociation entre agents.

De manière plus générale, les thématiques scientifiques abordées dans cette thématique de l'axe sont :

- Modélisation, analyse et simulation de la négociation et de la décision collective (intégrant l'argumentation dans les processus de négociation)
- Caractérisation et apprentissage du comportement d'agents (lors de la négociation)
- Intégration de l'a priori du décideur dans les modèles de décision (approche bayésienne)
- Modélisation de la causalité

L'étude des objets dits « complexes » est focalisée, dans un premier temps, sur les réseaux complexes. Un réseau est une manière naturelle de représenter formellement les relations entre les individus (réseaux sociaux) ou les objets de différentes natures. Dans le cas des réseaux mettant en jeu des individus, la construction, l'analyse, l'utilisation de ces réseaux constituent un domaine de recherche fécond aux frontières de plusieurs domaines comme l'informatique bien sûr, mais également la sociologie et les représentations sociales. Ils constituent un outil tout à fait stratégique dans les systèmes d'aide à la décision moderne : détection des coalitions/communautés, des individus charnières pour ces communautés, des individus atypiques, isolés, apparition de nouveaux acteurs. Or, les modèles développés actuellement pour modéliser ces réseaux sont essentiellement centrés sur les individus et ne prennent pas toujours en compte l'hétérogénéité des informations permettant d'analyser ces relations de façon complète : thématiques (de quoi parlent les individus), opinions (de quelle manière), … . Nos premiers travaux s'inscrivent justement dans cette perspective de recherche et proposent de nouveaux modèles afin de mieux représenter les réseaux de relation dans toute leur diversité.

D'un point de vue plus général, les réseaux complexes se caractérisent par une structure non-triviale, dynamique et des relations multicritère. Il existe, dans ces réseaux, une intrication forte entre structure et dynamique : la structure évolue et s'auto-organise, ce qui permet l'émergence de certaines propriétés non trivialement prévisibles : des phénomènes de coopération entre les nœuds (acteurs, objets, …) du réseau apparaissent. L'ensemble de ces caractéristiques pose encore beaucoup de difficultés aussi bien théoriques que pratiques. Tout ceci est également accentué par la considération de réseaux « de masse », où un grand nombre d'informations sont représentées. Le volume de données aujourd'hui accessible induit sans conteste la prise en compte du problème clé qu'est la visualisation, non seulement des réseaux eux-mêmes, mais des multiples points de vue sur ces derniers, de leur évolution, et également des connaissances que l'on peut extraire de ces multiples aspects.

De manière plus générale, les thématiques scientifiques abordées dans cette thématique de l'axe sont :

- Modélisation des réseaux complexes intégrant leur dynamique et leur hétérogénéité (graphes d'opinion, prétopologie stochastique, …)
- Caractérisation mathématique de certains phénomènes, comme la contraction, la fusion, l'explosion, …
- Visualisation des réseaux, de leur structure dynamique, des connaissances extraites, …

5.4 Santé et Environnement

1 – Introduction

Les systèmes de santé et d'environnement constituent une illustration exemplaire de systèmes complexes (interconnectés d'ailleurs) au sein desquels une grande diversité d'agents poursuivent chacun des objectifs bien précis tout en étant au service d'une cause commune : maintenir le niveau de santé d'une population ainsi qu'un grand nombre de facteurs incontrôlables qui en influencnent le comportement et les dynamiques.

En tant que systèmes complexes, systèmes de santé et systèmes environnementaux doivent donc être bien compris, à divers niveaux et selon différents points de vue, de manière à ce que leur fonctionnement et leur impact sur la société soient bien compris et/ou optimisés. C'est précisément le rôle du chercheur de se pencher sur les problèmes de ces systèmes afin de proposer des solutions aux décideurs.

2 – Défis scientifiques

Les défis scientifiques que nous posent les systèmes de santé et l'environnement sont nombreux et variés. Ils relèvent de différentes problématiques et mettent directement en jeu la société en la plaçant au cœur de ces systèmes. Le pollué est aussi le pollueur, ses comportements économiques et culturels participent à l'émission d'effluents de toutes sortes. Le patient est aussi un agent économique dont les

objectifs divergent de ceux du médeinc, autre agent économique d'un système où la collectivité joue un rôle important, en tant que payeur, etc. La méthodologie permettant d'aborder les problèmes ne peut que relever d'une approche « système complexe » interdisciplinaire. Nous nous intéresserons ici exclusivement à l'approche mathématico-informatique, en précisant que celle-ci est une approche parmi d'autres : économqiue, sociologique, anthropologique, etc.

3 – Approche scientifique

La recherche scientifique a pour objectif de proposer des concepts, développer des méthodes et construire des outils. L'objet de la recherche menée dans cet axe sur le plan des concepts relève des systèmes complexes, de la complexité et de l'aide à la décision. Il porte sur l'analyse des politiques de santé, des stratégies thérapeutiques, sur le fincnement des systèmes de santé et sur la gestion des risques sanitaires.

Pour ce qui est des méthodes développées, le travail s'effectue selon quatre directions : la théorie de la complexité, l'ananlyse systémique, la prétopologie et la théorie des ensembles aléatoires. Ces méthodes permettent la modélisation des comportements des acteurs, l'analyse de l'agrégation des préférences, l'évaluation et la négociation entre acteurs.

Les outils mis au point sont dans le champ de l'informatique distribuée, des algorithmes statistiques de décision, des outils de requétage et de fouille des données ainsi que du méta-apprentissage. Cela permet de travailler sur les problèmes de décision statistique, de simulation des comportements d'extraction de connaissances dans les grandes bases de données médico-économiques, du e-learning pour les professionnels de santé, etc.

Cette approche scientifique s'appuie donc sur les axes méthodologiques du laboratoire et contribue à alimenter la réflexion au sein de ces axes. L'ensemble des chercheurs de ERIC sont donc concernés par l'axe santé-environnement sur le plan de la conforntation de leurs modèles avec une réalité complexe. Les bases de données dont on dispose permettent une réelle évaluation des résultats de recherche fondamentale des axes.

Par ailleurs, sur le plan structurel, l'axe santé-environnement bénéficie de la présence d'une chaire d'excellence (Chaire Market Access) gérée par l'ex équipe MA2D au sein de Lyon 1 dont l'objet principal est l'analyse des systèmes de santé. Cette chaire a été construite pour une durée de cinq ans et a été financée par les Laboratoires Lundbeck SAS.

Enfin, par l'intermédiare de la spécialité « Aide à la décision médico et pharmaco-économique » du master « Santé et Populations » de Lyon 1, un grand nombrede thèses sont en cours sur ces problématiques.

4 – Projets support

La recherche de l'axe santé-environnement s'appuie et est validée par différents projets :

- le projet ECHOUTCOME, projet européen (DG SANCO, FP7) débutant le 1er février 2010, financé à hauteur de 1M€ environ, responsable : M. Lamure. Ce projet vise à analyser les guidelines existantes en matière de stratégie de financement du médicament au niveau européen, de comparer ces stratégies, d'en proposer de nouvelles de manière à fournir à l'Europe des guidelines intégrées tenant compte des spécificits des systèmes de santé de chaque pays. Sa durée est de trois ans, le consortium est constitué de 8 membres de 4 pays différents.

- le projet MOUSSON. Il s'agit d'un projet international, placé sous le sceau d'un accord cadre entre le CNRS (France) et le CNRST (Burkina Faso) initié en 1984. Le PIR-MOUSSON est lancé en octobre 2006 sous l'impulsion à la fois des Directions des Départements des Sciences Humaines et Sociales (SHS) du Centre National de la Recherche Scientifique (CNRS) en France et du Centre National de la Recherche Scientifique et Technique (CNRST) au Burkina Faso. Ancré dans les deux milieux scientifiques, ce partenariat est consolidé par la signature d'un protocole de coopération scientifique sur le PIR Mousson à Ouagadougou le 29 février 2008. L'objectif du projet Mousson vise à mettre en place à Ouagadougou (Burkina Faso) un système d'alarme pour prévenir les risques sanitaires ou les maladies liées à la pollution atmosphérique urbaine, tout en favorisant une meilleure qualité de vie et en améliorant le niveau de santé des populations locales.

A partir de 2009, des groupes interdisciplinaires Nord-Sud se sont constitués autour de 6 actions de recherche à traiter.

1- Action Analyse exploratoire des données hospitalières.

2- Action Caractérisation physico-chimique des polluants et analyses chimiques.

3- Action Enquêtes des pratiques quotidiennes et mesure de la pollution de l'air dans l'espace domestique.

4- Action Mise en relation de la mesure physique avec le ressenti des sujets

5- Action Questionnaires de qualité de vie (ciblée santé des patients)

6- Analyse de la construction interdisciplinaire (concerne tout le groupe)

D'autres projets en cours, viennent compléter cet esnsemble d'actions concrètes visant à tester les méthodes et outils développés dans la phase de recherhce fondamentale et approtant du fiancnement au laboratoire.

5.5 Sciences Humaines et Sociales (SHS)

Le laboratoire ERIC entretient depuis sa création des collaborations avec les secteurs de recherche des SHS. Nous comptons renforcer ce volet par des projets de collaboration comme celui qui est actuellement en cours avec les historiens et qui est décrit ci-dessous de manière plus détaillée. D'autres projets viendront par la suite étoffer cette collaboration.

Contexte :

La progression quasi exponentielle des capacités de stockage et d'échange de l'information, facilité par l'essor d'Internet, ouvre aujourd'hui des perspectives de mutations majeures dans les pratiques des historiens, et plus largement des chercheurs en Sciences Humaines et Sociales (SHS). Les données complexes qu'ils manipulent au quotidien (image, texte, dessins, etc.) constituent une véritable mine d'or pour déployer et évaluer les méthodes modernes de fouille de données, telles que développées au sein du laboratoire ERIC.

Dans ce contexte, l'objectif de ce projet scientifique consiste à définir un cadre méthodologique et des outils pour structurer efficacement les données complexes manipulées en SHS et pour ensuite les analyser en vue d'extraire de nouvelles connaissances ou, tout du moins, des informations encore non accessibles par les moyens disponibles avec les moyens actuels.

Partenaires : Institut des Sciences de l'Homme (ISH), Laboratoire d'Informatique Paris Descartes (LIPADE), Telecom ParisTech département TSI, Laboratoire Extraction et Exploitation de l'Information en Environnements Incertains (E3I2).

Positionnement :

Le projet vise à capitaliser sur plusieurs projets antérieurs ayant été menés entre des chercheurs du laboratoire ERIC et des historiens du laboratoire LARHRA à l'ISH de Lyon. Ces travaux sont ainsi à l'origine du projet SyMoGIH (Système Modulaire de Gestion de l'Information Historique) qui permet de modéliser l'information historique à l'aide d'un méta-modèle réalisé dans le formalisme entité-relation. D'un autre côté, un projet interne financé par l'Université de Lyon 2 pour une durée de un an est actuellement en cours. Il a notamment permis d'ouvrir plus largement le dialogue entre les chercheurs en histoire et en informatique. L'objectif de ce projet interne est d'aboutir à l'élaboration d'un document de réponse à un appel d'offre national de type ANR. Il fédère des chercheurs de plusieurs laboratoires, dont le LIPADE (Paris 5), Telecom ParisTech et le E3I2 (ENSIETA, Brest). Il est rattaché aux axes 3 (Numérisation, simulation, modélisation de la complexité) et à l'axe 8 (texte, discours et cultures) des axes de recherche fixés comme prioritaires par l'Université de Lyon 2.

Verrous scientifiques :

1) Modélisation des données historiques. Ces données sont, par nature, évolutives et demandent un traitement adapté. Les notions de qualité des sources et de traçabilité sont également très importante et

dépassent largement le « simple » cas des données historiques. On peut par exemple penser aux données médicales et au dossier patient.

2) Intégration de données issues de sources hétérogènes. Ces données sont complexes (image, texte, etc.), imparfaites, imprécises, voire manquantes dans de nombreux cas. Le verrou majeur est de parvenir à en extraire des informations de haut niveau sémantique pour instancier notre nouveau modèle. Par exemple, une grande partie des données est du texte rédigé en langage naturel et son utilisation pose les problèmes habituels liés à la langue : polysémie, ambiguïté, ellipses, etc. Ce problème de « gap sémantique » se retrouve bien entendu avec les données images.

3) Outils d'analyse pour le chercheur en SHS. Une fois le modèle construit et instancié, il convient de définir des outils d'analyse et des modes d'interactions avec l'utilisateur final de notre système. A ce titre, les besoins des chercheurs en SHS sont très spécifiques et demandent un traitement adapté.

Approches préconisées et résultats escomptés :

Concernant la modélisation de ce type particulier de données, nous envisageons de recourir à un entrepôt de données. Cet entrepôt pourra s'appuyer sur le langage XML qui présente des caractéristiques très intéressantes pour représenter et manipuler les données complexes. Ainsi, la flexibilité de ce modèle semi structuré nous semble tout à fait adaptée au traitement des données évolutives et aux questions de traçabilité. Il permet entre autre d'élaborer des analyses en ligne riches adaptées aux données complexes. Le traitement des images sera réalisé en collaboration avec le LIPADE dont l'une des spécialités concerne justement les documents anciens. Des techniques comme le wordspotting nous permettra d'extraire des caractéristiques de haut niveau qui alimenteront l'entrepôt. Le traitement des textes fera appel à des techniques de fouille de textes développées notamment au laboratoire ERIC comme l'extraction et la structuration de thématiques. Pour tous les problèmes de fusion, nous projetons d'utiliser la théorie de l'évidence de Dempster-Shafer avec l'aide du laboratoire E3I2 pour résoudre les problèmes de conflits et de fusion en les adaptant au cas des données complexes. Les outils d'analyse envisagés sont tout d'abord ceux naturellement associé aux entrepôts de données (analyse en ligne, requêtage), mais également la construction et la visualisation des réseaux sociaux impliquant les acteurs historiques. Parmi les résultats escomptés, nous souhaitons obtenir un prototype directement utilisable par les chercheurs en SHS. Cela signifie notamment que nous envisageons de tester ce prototype non seulement sur les données historiques du LARHRA, mais également sur d'autres types de données (BNF, INA, etc.).

5.6 Logiciels libres

L'équipe développe depuis de nombreuses années des outils logiciels afin de permettre une application réelle et rapide des résultats scientifiques. Par exemple, le logiciel libre Tanagra (http://eric.univ-lyon2.fr/~ricco/tanagra/) contribue fortement au développement des projets de recherche et de coopérations avec d'autres équipes. Plusieurs projets de coopérations doivent se mettre en place dans un avenir proche. Le premier concerne l'analyse d'images vidéo eu/ou statistiques avec l'équipe L3I de l'Université de la Rochelle (http://l3i.univ-larochelle.fr/). Le second concerne le classement automatique d'insectes dans le continent sud américain à l'aide de données météorologiques, en coopération avec le Pr Rabinovich de l'Universidad Nacional de la Plata (Argentine - http://www.cepave.edu.ar/).

La liste des logiciels développés au sein de l'équipe et diffusés gratuitement sur Internet sont listés ci-dessous.

Logiciels développés au sein de l'équipe et diffusés gratuitement sur Internet		
Nom	Auteur	Description
Amado	J.H. Chauchat – A. Risson	Visualisation de matrices. Il a été intégré à SPAD.
REGRESS32	Ricco Rakotomalala	Logiciel d'économétrie destiné à l'enseignement
SIPINA	Ricco Rakotomalala – D.A. Zighed	logiciel de DATA MINING implémentant les algorithmes de graphes et arbres d'induction
TANAGRA	Ricco Rakotomalala	Logiciel de Data Mining destiné à l'enseignement et la recherche. 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.
SMAIDoC	Omar Boussaid	Intégration de données complexes par des méthodes de système multiagent
DWEB	Jérôme Darmont	DWEB (Data Warehouse Engineering Benchmark) est un des rares bancs d'essais décisionnel actuellement opérationnel permettant d'évaluer les performances des entrepôts de données.
Pretopolib	Vincent Levorato	Librairie de calcul d'éléments pré-topologique, de simulation et de visualisation de la structure.
Donalysor	Ahmed Bounekkar	Logiciel d'analyse de donnée