3ème Atelier " Systèmes d'Information et de Décision pour l'Environnement " Inforsid - 25 mai 2010

Sandro Bimonte, André Miralles, François Pinet

Congrès Inforsid

25 mai 2010

3ème Atelier « Systèmes d'Information et de Décision pour l'Environnement »

Actes de l'atelier

Edités par :

Sandro Bimonte, UR TSCF, Cemagref, Clermont-Ferrand
André Miralles, UMR TETIS, Cemagref, Montpellier
François Pinet, UR TSCF, Cemagref, Clermont-Ferrand

3ème Atelier « Systèmes d'Information et de Décision pour l'Environnement »
Inforsid - 25 mai 2010

La recherche en informatique et en systèmes d'information offre depuis des années des solutions de plus en plus performantes pour relever les récents challenges environnementaux. Les données environnementales acquises sont de plus en plus nombreuses et sont aujourd'hui structurées et analysées au sein de Systèmes d'Information et/ou de Systèmes d'Aide à la Décision.

L'objectif de l'atelier est de présenter comment les toutes dernières avancées de la recherche en systèmes d'information ou de système de décision s'appliquent au domaine environnemental. L'atelier est ouvert aussi bien à la présentation de travaux de recherche déjà appliqués au contexte de l'environnement, qu'à des réflexions plus prospectives sur les possibilités d'utilisation d'un produit de la recherche en informatique pour une application environnementale.

La journée d'atelier a été découpée en trois sessions. La première porte sur les SI dédiés à l'eau, la seconde s'intéresse aux SI pour les feux de forêt, et la troisième concerne les SI destinés à la qualité environnementale.

La qualité des travaux laisse présager une journée d'atelier particulièrement enrichissante. Nous remercions les auteurs pour leurs soumissions et tous les membres du comité de programme pour leur excellent travail de relecture.

Les organisateurs de l'atelier
Sandro Bimonte, UR TSCF, Cemagref, Clermont-Ferrand
André Miralles, UMR TETIS, Cemagref, Montpellier
François Pinet, UR TSCF, Cemagref, Clermont-Ferrand

3ème Atelier « Systèmes d'Information et de Décision pour l'Environnement » Inforsid - 25 mai 2010

Programme

8h30-9h00 : Accueil des participants

9h00-10h30 : Session "Du cours d'eau à la mer"

- **Mise en place d'un Système d'Information Géographique Participatif pour cartographier la connaissance et la gestion des risques côtiers –** *Chancerel R., Lopistéguy P., Dagorret P.* (p.1-11)
- **Un système d'information pour le suivi de la qualité des cours d'eau –** *Grac C., Braud A., Le Ber F., Trémolières M.* (p.12-21)

10h30-11h00 : Pause

11h00-11h30 : Session "Du cours d'eau à la mer" (suite)

- **ObServe : Un système d'acquisition et de gestion de données d'observations –** *Cauquil P., Libourel T., Pierkot C., Tissot A., Tornare J.* (p. 22-34)

11h30-12h00 : Session "Incendies de fôrets"

- **Integration of image processing methods for fuel mapping –** *Maillé E., Borgniet L., Lampin-Maillet C., Jappiot M., Bouillon C., Long-Fournel M., Morge D., Amine El Gacemi M., Sorin D.* (p. 35-57)

Repas

14h00-14h45 : Session "Incendies de fôrets" (suite)

- **Une approche innovante de modélisation du risque d'incendie de forêt fondée sur la cartographie des interfaces habitat-forêt, nouvelle clé de lecture du territoire** - *Lampin-Maillet C., Jappiot M., Ferrier J.P.* (p. 58-70)

14h45-15h30 : Session "Qualité Environnementale"

- **Structure informatique pour la réponse aux plaintes liées à l'air au sein des logements** - *Bellia Heddadji Z., Vincent N., Kirchneret S., Stamon G.* (p. 71-89)

15h30-16h00 : Pause

16h00-16h30 : Session "Qualité Environnementale" (suite)

- **Gestion intelligente d'entrepôts de données énergétiques : quels défis? –** *Copin L., Laurent A., Rey H., Teisseire M., Vasques X.* (p. 90-103)

16h30-17h00 : Clôture et bilan de la journée

3ème Atelier « Systèmes d'Information et de Décision pour l'Environnement »
Inforsid - 25 mai 2010

Organisateurs :

Les organisateurs de l'atelier
Sandro Bimonte, UR TSCF, Cemagref, Clermont-Ferrand
André Miralles, UMR TETIS, Cemagref, Montpellier
François Pinet, UR TSCF, Cemagref, Clermont-Ferrand

Comité de programme :

Ahmed Lbath, LIG – Grenoble
Beniamino Murgante, Université de Basilicate – Italie
Catherine Roussey, LIRIS – Lyon
Florence Le Ber, Ecole Nationale du Génie de l'Eau et de l'Environnement de Strasbourg
Fréderic Flouvat, Université de Nouvelle Calédonie
Gil De Sousa, TSCF - Cemagref, Clermont-Ferrand
Jean-Christophe Desconnets, IRD – Montpellier
Jean-Paul Donnay, Université de Lièges –Belgique
Jean-Pierre Chanet, TSCF - Cemagref - Clermont-Ferrand
Jérôme Jensel, IMAG - Grenoble
Karine Zeitouni, Université de Versailles
Maguelonne Teisseire, UMR TETIS - Cemagref Montpellier
Mathieu Roche, LIRMM - Montpellier
Michel Passouant, Cirad - Montpellier
Michel Schneider, LIMOS - Clermont-Ferrand
Myoung-Ah Kang, LIMOS - Clermont-Ferrand
Sylvie Servigne, LIRIS - Lyon
Thérèse Libourel, LIRMM - Montpellier
Thierry Badard, CRG - Université Laval, Québec
Vincent Abt, TSCF - Cemagref - Clermont-Ferrand
Yvan Bédard, Université de Laval - Québec

Mise en place d'un Système d'Information Géographique Participatif pour cartographier la connaissance et la gestion des risques côtiers

Romain Chancerel*, Philippe Lopistéguy, Pantxika Dagorret****

** CERCO - Centre Européen sur les Risques Côtiers - Centre de la Merde Biarritz Plateau de l'Atalaye, 64200 Biarritz*

cerco@museedelamer.com

*** LIUPPA - Laboratoire d'Informatique de l'Université de Pau et Pays de l'Adour IUT de Bayonne, 2 allée du Parc Montaury, 64600 Anglet*

Philippe.Lopisteguy@iutbayonne.univ-pau.fr, Pantxika.Dagorret@iutbayonne.univ-pau.fr

RÉSUMÉ. La connaissance scientifique des littoraux est importante mais peu diffusée ; elle est variée, fragmentée, en constante évolution et les gestionnaires de même que le monde de la science ont vocation à participer à la réflexion concernant la gestion du littoral. Des inventaires de la connaissance ont été mis en place par différents organismes. Ils constituent une photographie du paysage institutionnel à un moment donné mais beaucoup reste à faire pour 1) pérenniser ces initiatives, 2) diffuser efficacement cette information et 3) rendre compte des relations entre les éléments constitutifs de ces inventaires. Cet article présente un outil de collecte, de pérennisation et de diffusion de cette information. Notre proposition comporte deux volets complémentaires qui s'articulent autour d'un Système d'Information Géographique Participatif hébergé sur une plateforme WEB. Le premier volet concerne le référencement des travaux en rendant compte des relations institutionnelles entre les éléments constitutifs de l'inventaire. Le second volet concerne le rôle joué par les acteurs référencés au sein du processus de gestion pour un aléa bien défini. Les informations référencées pourront concerner la connaissance de l'aléa mais aussi les stratégies de gestion mises en place pour s'adapter à cet aléa.

ABSTRACT. Scientific knowledge is important but little publicized; it is diverse, fragmented, constantly changing and managers as well as the scientists are expected to participate in the discussion on coastal management. Inventories of knowledge have been implemented by different agencies. They are photographs of the institutional landscape at one point but much remains to be done to 1) sustain these initiatives, 2) disseminate this information effectively and 3) report relationships between the components of these inventories. This paper presents a tool for collection, perpetuation and dissemination of this information.

MOTS-CLÉS : Gestion des Risques côtiers, Système d'information géographique participatif.

KEY WORDS: Coastal risk management, Participative Geographic Information System

1. Introduction

L'actualité des problèmes liés aux perturbations des zones côtières témoigne de la complexité des risques pour l'environnement côtier. Une gestion efficace de ces risques ne pourra se faire que sur la base d'une information pratique basée sur une expertise scientifique rigoureuse. Il faut pouvoir offrir aux gestionnaires une information organisée et synthétique des phénomènes à l'origine de ces risques (*i.e.* aléas) et également leur proposer des stratégies de prévention, de gestion de crise et de résilience. L'interaction entre scientifiques et décideurs est généralement faible. La connaissance disponible n'est pas efficacement utilisée car mal reçue par des messages inadaptés et les mécanismes institutionnels de transmission ne sont pas bien développés. Ceci est illustré par les incohérences entre les planifications de la terre et de la mer, par des incohérences parmi ses règlements environnementaux existants et par le manque d'instruments pour leur mise en œuvre.

Il faut établir un bilan des connaissances concernant les zones côtières afin de disposer d'un éventail de données plus conséquent et plus performant, pour une meilleure gestion intégrée et durable des perturbations. De nombreuses études de grande qualité ont été réalisées sur les zones côtières, mais rares sont les documents de synthèse faisant état des acquis scientifiques et techniques mais aussi des lacunes en termes de connaissance et de gestion. La création d'un centre de ressources, d'éducation et de formation à travers lequel scientifiques, technologues et gestionnaires pourront échanger connaissances et expériences est apparu comme l'une des mesures les plus urgentes à mettre en place.

Le Centre de la Mer de Biarritz va donc assurer un rôle d'interface dans la connaissance des océans et des risques côtiers dans le but de mettre à disposition des gestionnaires des éléments pour prévenir les risques et les gérer si nécessaire. Il deviendra à terme une interface privilégiée à travers laquelle les scientifiques et les usagers de la côte pourront partager informations et expériences. Il est devenu, en tant que Centre Européen des Risques Côtiers (CerCo, 2008), l'un des centres de ressources de l'Accord européen et méditerranéen sur les risques majeurs (EUR-OPA) dont le domaine d'action englobe la connaissance des aléas, la prévention des risques, la gestion des crises ainsi que l'analyse post-crise et la réhabilitation.

Dans le cadre de cette mission, le Centre de la Mer a proposé un certain nombre d'actions spécifiques: conférences, séminaires, formations, expositions. L'engouement suscité auprès des gestionnaires pour la formation « Connaissance et gestion des risques côtiers » organisée en octobre 2009 et adressée aux agents des collectivités territoriales témoigne d'un manque réel en formation et en information concrète sur les risques côtiers.

La stratégie préconisée par EUR-OPA comprend la réalisation d'un inventaire des organismes responsables de la gestion des risques côtiers en France et en Europe, la mise en place de formations adressées aux agents territoriaux et le développement d'un système d'information permettant un meilleur échange entre les gestionnaires et le monde de la science. Les connaissances scientifiques ainsi que les

stratégies de défense contre les risques côtiers sont importantes mais peu diffusées ; elles sont variées, fragmentées, en constante évolution et de ce fait difficiles à suivre.

Au-delà des besoins en formation et en information, les gestionnaires et le monde de la science ont vocation à participer à la réflexion concernant la gestion du littoral. Il faut donc donner une dimension 'participative' à l'outil qui leur sera proposé. Le Centre de la Mer de Biarritz a prévu de mettre en place un centre de ressources qui soit capable d'intégrer des informations provenant des organismes scientifiques (équipes, projets, programmes de recherche) et gestionnaires (leurs pratiques, stratégies, techniques, politiques de défense) et ce à différentes échelles territoriales tout en tenant compte de leur caractère évolutif.

Ces informations doivent être diffusées dans un souci de clarté; c'est pourquoi les outils qui seront mis en place permettront de cartographier efficacement l'état des connaissances ainsi que les stratégies d'adaptation qui ont été mises en place sur le littoral. Cette initiative couvre au moins deux objectifs à court terme:

- L'identification d'un cadre conceptuel permettant de catégoriser la connaissance (au sens large) du risque côtier et de rendre compte des relations entre les organismes de connaissance et de gestion
- La mise en ligne d'une plateforme WEB permettant une mise en commun de la connaissance scientifique des littoraux et des stratégies de défense avec interface participative et information géo-référençable.

2. Étude préliminaire

2.1. *Développement d'un cadre conceptuel pour le développement de la base de données*

Les risques côtiers peuvent être d'origine naturelle (érosion, inondation,…), anthropique (pollution, etc.) mais sont le plus souvent la conséquence d'une combinaison de forçages naturels et anthropiques (Klein, 1999). Le CerCo développe en partenariat avec le laboratoire ADES Aménagement, Développement, Environnement, Santé et Sociétés - UMR 5185 – et l'université de Southampton (GB) un cadre conceptuel qui permet de catégoriser d'une part, les connaissances scientifiques et d'autre part, les adaptations socio-économiques (prévention, gestion de crises, résilience) qui ont été mises en œuvre pour répondre aux risques étudiés par la communauté scientifique. Cette étude permettra notamment de déterminer une terminologie rigoureuse pour les différents concepts du risque (susceptibilité, adaptation, vulnérabilité, etc.) et de comprendre comment ces concepts sont liés entre eux.

- Connaissance du risque côtier :
  En fonction de l'aléa que l'on souhaite traiter, un bilan des connaissances doit être établi. Il mettra en avant la connaissance scientifique sur les

forçages naturels à l'origine du risque et la réponse des milieux naturels à ces forçages. Par exemple, pour le risque de submersion/d'inondation, un bilan des connaissances scientifiques doit être fait pour connaître quelle sera l'ampleur du changement climatique dans les décennies avenir et quel sera son effet potentiel sur la vulnérabilité des habitats et des hommes.

- Stratégies/politiques de défense (adaptation) :
  Elles seront classées selon l'aléa qu'elles doivent traiter (submersion, pollution, etc.), leur emprise géographique (locale, nationale, internationale) et le type d'action (actions de prévention du risque, de gestion de crise ou de réhabilitation). A partir de cette classification, un inventaire des techniques et des stratégies de défense (ou d'adaptation) pourra être réalisé au niveau des différentes régions de la façade atlantique.

2.2. *Objectifs opérationnels du SIG-Participatif*

Le CerCo développe en collaboration avec le Laboratoire d'Informatique de l'Université de Pau et des Pays de l'Adour (LIUPPA) une plateforme WEB interactive qui permet un peuplement décentralisé de la base de données. Les informations qui sont saisies peuvent-être géoréférencées et représentées sur une carte. Cet outil permettra de représenter à la fois l'état de la connaissance scientifique des risques, et les stratégies d'adaptation qui ont été mises en place pour lutter contre ces risques.

2.2.1. *Représentation géographique*

Le développement des technologies de l'information géographique permet de développer de nouveaux outils capables d'organiser et de présenter des données alphanumériques géoréférencées, ainsi que de synthétiser ces données grâce à des plans et des cartes. Concrètement, cet outil permettra de visualiser sur une carte l'ensemble des informations relatives à la connaissance et à la gestion d'un aléa déterminé et, à travers le référencement de l'information, de situer la place et la portée de ces informations au sein des processus de gestion et de connaissance de l'aléa.

2.2.2. *Mise en place d'une veille scientifique*

En pratique, la veille scientifique se met en place naturellement par les centres de recherche eux-mêmes afin de promouvoir leurs travaux, dans l'optique de transferts de technologie, de partenariats financiers, etc. Cette diffusion prend la forme de communications lors de congrès, de conférence et de publications dans divers médias et revus spécialisées ou toute autre manifestation intra- ou extracommunautaire.

Ce SIG se pose comme un nouvel outil d'échanges au service des centres de recherche et des gestionnaires. Les fonctionnalités qu'il proposera sont étudiées pour

que l'effort de coopération soit entretenu sachant le propriétaire de l'information est seul juge du bien fondé du transfert.

En effet, comme le souligne le Réseau de Recherche du Littoral Aquitain (RRLA), il est souvent difficile de demander aux organismes de rester en veille en particulier pour le compte d'organismes (dans ce cas, le CerCo) qui leur sont étrangers (aucune convention n'est signée) et avec lesquels les contacts ne sont qu'épisodiques.

Aussi, une réelle stratégie d'industrialisation doit être mise en place auprès des organismes gestionnaires (collectivités territoriales, associations d'élus, groupements d'intérêts publiques, etc.) et des organismes de recherche et prestataires de service. Des partenariats stratégiques devront être mis en place avec les réseaux de coopérations existants (réseau des Centres spécialisés de l'accord EUR-OPA, CNRS) et nous pourrons nous appuyer sur le réseau de gestionnaires que nous sommes en train de créer avec la formation « connaissance et gestion des risques côtiers ».

3. Mise en place de la plateforme WEB participative

Dans sa première phase, le développement du prototype WEB permettant le peuplement décentralisé de connaissances scientifiques ne s'est attaché qu'au référencement de l'information dite *scientifique* ; c'est-à-dire aux travaux publiés par les laboratoires de recherche spécialisés. Les aspects stratégies et techniques de gestion n'ont pas été traités lors de cette première phase de développement.

Les inventaires des études scientifiques, des laboratoires, des centres de recherche, etc. (ex : projet ANCORIM), sont traditionnellement réalisés manuellement par des agents qui ont pour mission de contacter les centres de recherches afin d'obtenir les informations. Ce travail est long et fastidieux et les informations récoltées peuvent devenir obsolètes avant même la fin du référencement à cause du caractère évolutif de ce tissu institutionnel.

L'objectif actuel du SIG développé est de permettre à tout centre de référencer ses propres études grâce à un login obtenu auprès d'un modérateur. Le SIG couvre les tâches essentielles d'ajout, de visualisation et de modification décentralisée d'études. Le modèle de données correspondant aux écrans ici présentés, est le suivant :

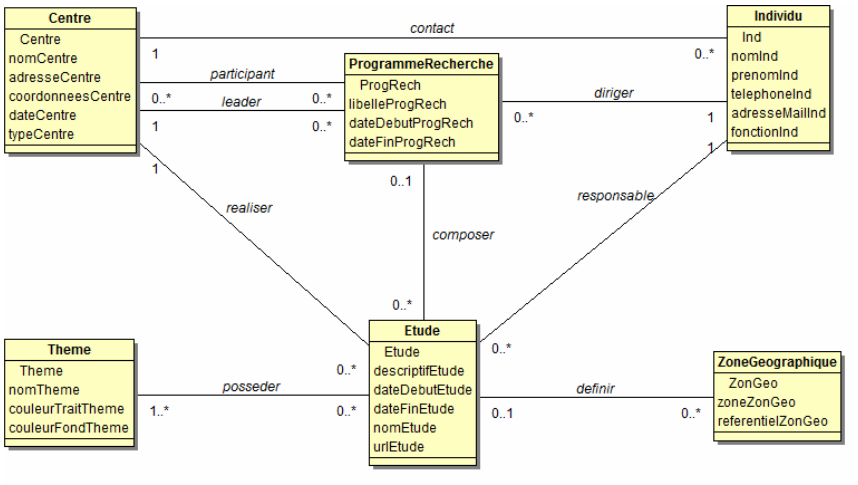


Figure 1. *Schéma de base de données*

Une étude possède entre autres, un nom, un descriptif, une URL et des dates de début et de fin. Ces informations facultatives, hormis le nom, sont renseignées par l'auteur de la fiche. Une étude peut relever d'un programme de recherche, est pilotée par un centre de recherche et est menée par un responsable. Une étude peut concerner plusieurs zones géographiques définies par défaut dans le référentiel WGS84 (utilisé par Google Maps[1]). Une couleur spécifique à chaque thème est utilisée pour délimiter les zones géographiques des études correspondantes. Si plusieurs thèmes sont traités dans une étude la zone de l'étude est colorée en gris.

Actuellement les thèmes éligibles pour une étude sont au nombre de quatre <Biodiversité, Pollution, Submersion, Aménagement>. Ces quatre thèmes correspondent à une synthèse issue des premières études répertoriées et d'une vue cooptée par divers centres de recherche contactés. L'étude conceptuelle permettra par la suite d'affiner ces quatre thématiques et de définir une classification articulée sur les aléas (ex : pollutions accidentelles, coastal squeeze, etc.) afin de délivrer une information pratique et concrète basée sur une description précise de l'aléa (forçages naturels et anthropiques impliqués, cadre légal, risques résiduels, etc.).

L'interface permet deux types de scénarios d'interaction pour 1) la consultation et 2) le peuplement d'études scientifiques.

[1] Nous avons choisi l'outil GoogleMaps à cause de son niveau de qualité acceptable sur les zones côtières, sa grande popularité et sa facilité d'usage. Nous n'avons pas retenu Google Earth car il requiert une installation préalable et son usage est moins intuitif. Nous travaillons actuellement au changement de fonds cartographiques en cours de session (IGN, GoogleMaps, BingMaps, Yahoo, etc.) tout en conservant la visualisation des mêmes informations géolocalisées.

Le scénario de consultation débute par une recherche en termes de Thème(s), de Centre, de Période et de Géolocalisation. Les critères, lorsqu'ils sont renseignés, sont reliés par une relation « ET ». Par exemple la recherche d'études de <Thèmes=Submersion> ET <Centre=?> ET <Période=(2008,?)> ET dont la géolocalisation a une intersection non vide avec l'embouchure de la Garonne est une requête exprimée dans la **Figure 1**.

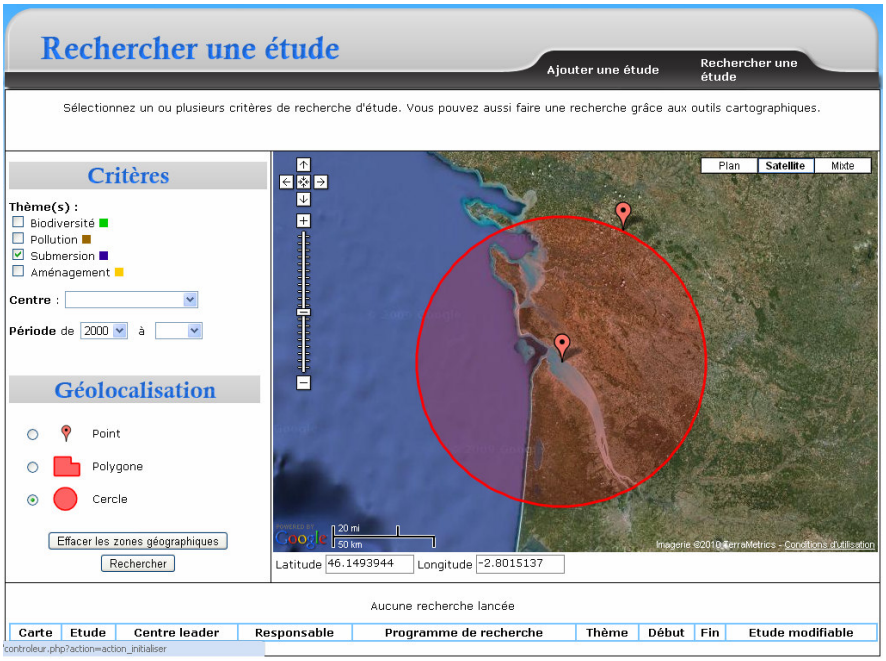


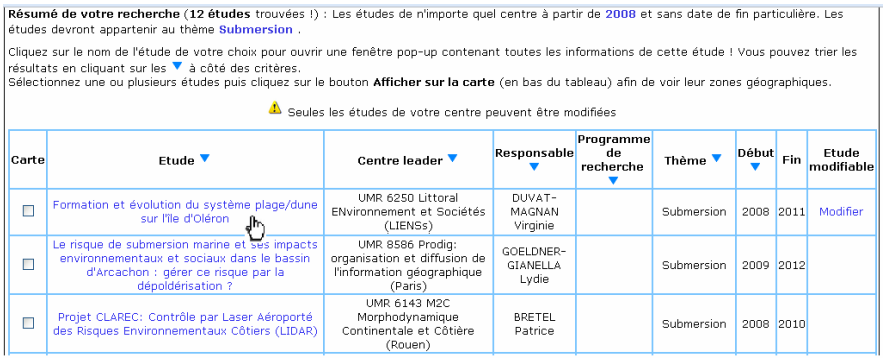
Figure 2. *Exemple de requête*

La zone sur laquelle les études sont recherchées est délimitée par les outils cercle, polygone et point. Lorsque seule la zone de recherche est définie, le résultat retourne la liste les études géoréférencées ayant une intersection non vide avec la zone définie par l'utilisateur.

Le résultat de la requête est présenté par une liste ordonnable selon chacun des critères Étude, Centre, Responsable, Programme, Thèmes, Début et Fin (cf. **Figure 2**). Lorsqu'aucun critère n'est renseigné dans la requête, toutes les études sont listées dans le résultat.

Pour les études de son choix, le lecteur peut activer ou désactiver la visualisation de sa géolocalisation (case à cocher sur la gauche). Sur la carte, le passage de la souris sur une zone géographique informe sur l'étude concernée. En cliquant sur l'étude qui l'intéresse (cf. **Figure 2**) ou sur la zone géographique correspondante, le lecteur peut accéder à la fiche descriptive de l'étude archivée

dans le SIG (cf. **Figure 3**). Lorsque l'utilisateur est auteur de la fiche il peut accéder à son contenu pour la modifier.



Résumé de votre recherche (**12 études** trouvées !) : Les études de n'importe quel centre à partir de **2008** et sans date de fin particulière. Les études devront appartenir au thème **Submersion** .

Cliquez sur le nom de l'étude de votre choix pour ouvrir une fenêtre pop-up contenant toutes les informations de cette étude ! Vous pouvez trier les résultats en cliquant sur les ▼ à côté des critères.
Sélectionnez une ou plusieurs études puis cliquez sur le bouton **Afficher sur la carte** (en bas du tableau) afin de voir leur zones géographiques.

⚠ Seules les études de votre centre peuvent être modifiées

Carte	Etude ▼	Centre leader ▼	Responsable ▼	Programme de recherche ▼	Thème ▼	Début ▼	Fin	Etude modifiable
☐	Formation et évolution du système plage/dune sur l'île d'Oléron	UMR 6250 Littoral ENvironnement et Sociétés (LIENSs)	DUVAT-MAGNAN Virginie		Submersion	2008	2011	Modifier
☐	Le risque de submersion marine et ses impacts environnementaux et sociaux dans le bassin d'Arcachon : gérer ce risque par la déspoldérisation ?	UMR 8586 Prodig: organisation et diffusion de l'information géographique (Paris)	GOELDNER-GIANELLA Lydie		Submersion	2009	2012	
☐	Projet CLAREC: Contrôle par Laser Aéroporté des Risques Environnementaux Côtiers (LIDAR)	UMR 6143 M2C Morphodynamique Continentale et Côtière (Rouen)	BRETEL Patrice		Submersion	2008	2010	

Figure 3. *Exemple de résultat d'une requête*

La spécialisation de ces capacités d'édition pour informations géo-référençables dans des domaines diagnostiqués d'intérêt par les acteurs du paysage institutionnel (labos de recherche, organismes prestataires de services, organismes gestionnaires, monde associatif) contribue au peuplement de la base de données.

Système d'Information Géographique - Fiche d'identité Etude - Google Chrome

http://projetmer.free.fr/modules/afficherPopUpEtude.php?idEtude=8

Fiche d'identité

Formation et évolution du système plage/dune sur l'île d'Oléron

Contact

NOM et prénom du responsable : DUVAT-MAGNAN Virginie
Fonction : Professeur de géographie
Adresse mail : virginie.duvat@univ-lr.fr
Téléphone : 0546458274

Plan | Satellite | Mixte

10 mi
20 km

Imagerie ©2010 TerraMetrics - Conditions d'Utilisation

Informations

Programme de recherche : aucun
Date de début de l'étude : 2008
Date de fin : 2011
Thème : Submersion
Description de l'étude : Vise deux types d'objectifs, développer la connaissance et proposer des outils opérationnels aux gestionnaires. Il comporte plusieurs axes : 1) Un travail de reconstitution de l'évolution du littoral qui repose (a) sur l'analyse de documents anciens, intégrés dans la base de données HISTOLITTO et (b) sur la réalisation de datations ; 2) L'étude de l'évolution de ce système au cours des dernières décennies à partir de l'analyse diachronique de photographies aériennes et d'images satellites ; 3) L'évaluation de la vulnérabilité de ce système aux facteurs de pression existants ; 4) La mise au point d'un protocole de gestion rationnelle de l'érosion

Figure 4. *Exemple de consultation de fiche d'une étude*

Par exemple, la spécialisation de l'outil pour décrire des programmes et stratégies d'adaptation contre les risques relève de cette tâche.

L'intégration dans le SIG, d'outils et de techniques complémentaires pour faciliter la gestion des risques côtiers, est conduite en accord avec le CerCo et ses partenaires, en fonction des objectifs et stratégies qu'ils se fixent. Les choix méthodologiques et conceptuels médiatisés seront conduits selon des approches pratiquées en Systèmes d'Information Décisionnels (Roy, 1985), en Systèmes d'Information Géographiques (Heywood, 2006), en Réseaux Sociaux et en Ergonomie (Bastien, 1993).

4. Perspectives

4.1. *Référencement des travaux*

Nous utilisons ici le terme « travaux » au sens large du terme. Il comprend à la fois les études scientifiques produites par les laboratoires de recherche et les stratégies de défense, pratiques, techniques ou aménagements qui ont été mises en place par les services techniques des collectivités ou autres organismes prestataires pour lutter contre un risque bien défini.

L'intérêt de cet outil résidera dans la façon dont il répondra concrètement aux besoins des collectivités. Des initiatives similaires existent aux niveaux national (Institut Français pour l'environnement) ou régional (Syscolab, Observatoire de la Côte Aquitaine, etc.) mais il s'agit surtout de veilles assez généralistes sur les programmes de recherche ou de mise à disposition de données cartographiques. L'originalité de notre approche est qu'elle se base sur une caractérisation rigoureuse du risque avec un référencement de l'information axé sur l'aléa afin de pouvoir scénariser le processus de gestion (prévention, crise, résilience) et de permettre aux différents acteurs scientifiques et techniques de situer leur action dans ce processus de connaissance et de gestion du risque.

4.2. *Interopérabilité des bases de données de la connaissance et de la gestion*

Le travail sur le cadre conceptuel devra faire émerger des passerelles entre les informations, issues d'une part de la connaissance scientifique et d'autre part de stratégies de gestion.

En effet, la gestion des risques côtiers et la recherche scientifique sur le littoral sont des domaines qui peuvent différer en termes organisationnels et en termes d'objectifs. La recherche scientifique va, par exemple, plutôt s'intéresser à décrire

les forçages[2], c'est-à-dire les modifications environnementales qui sont à l'origine du risque, alors que la gestion traite plutôt de l'adaptation à ces forçages. Les adaptations qui peuvent être du domaine naturelle (ex : migration des espèces en cas de changement climatique) ou socio-économique (prévention, gestion de crise ou résilience) sont les réponses possibles à ces forçages. Nous nous intéresserons ici aux adaptations socio-économiques, c'est-à-dire aux techniques, stratégies, politiques ou encore pratiques mises en place par les gestionnaires pour s'adapter à ces forçages.

Plus globalement, les thématiques qui ont été retenues pour le développement du système actuel sont la biodiversité, les aménagements, la pollution et la submersion. Des modifications devront être envisagées pour considérer des indicateurs métiers permettant d'aborder les risques selon des prismes plus variés tels que par exemple le tourisme, l'industrie ou le commerce.

4.2. *Animation et suivi*

Comme toute interface participative, cet outil devra évoluer en fonction de l'utilisation que les acteurs suggèreront. De nombreuses fonctionnalités pourront ainsi être greffées tout au long de la vie de cet outil (exemples : forums de discussions sur des projets d'aménagement, publication d'appels d'offre, messagerie entre les utilisateurs, etc.).

Outre l'utilisation de cet outil, la plateforme WEB prévoit la mise en place d'un espace personnalisé pour les gestionnaires avec des fonctionnalités telles que l'accès aux ressources/bases de données, l'inscription à des formations spécialisées, la participation à des forums de discussion, la saisie de documents ou encore une messagerie entre les membres du réseau.

5. Conclusion

Les risques côtiers sont un chapitre indissociable de la gestion intégrée des espaces littoraux. Indépendants des frontières des États, les risques côtiers nécessitent une prise en considération nouvelle au niveau européen, en relation étroite avec les caractéristiques relevant en première ligne essentiellement des sciences et techniques de la Terre.

---

[2] On appelle *forçages* les changements environnementaux qui peuvent être d'origines climatique (ex: élévation du niveau de la mer) ou non-climatique (ex: urbanisation de la bande littorale) mais aussi socioéconomique (ex : la croissance démographique des zones littorales) qui représentent une menace pour les écosystèmes et/ou pour des systèmes socioéconomiques. C'est souvent une conjonction de forçages qui cause le risque c'est pourquoi on parle de scenarii de forçages ou encore d'*aléas*.

Actuellement de nombreux pays voient avec impuissance leurs zones côtières se dégrader et subir des modifications désastreuses car ils ne possèdent pas les moyens nécessaires en vue d'une action de réhabilitation sur la base de connaissances bien étayées. Il en résulte souvent des interventions reposant sur des initiatives peuvent se révéler malencontreuses. Il est donc important de développer l'enseignement d'une gestion intégrée et raisonnée des milieux côtiers.

Les programmes européens de recherche scientifique ont permis de grandes avancées dans la connaissance des risques côtiers, mais il reste encore à assurer la diffusion de ces connaissances vers les gestionnaires. Cet outil de diffusion ainsi que les formations mises en place par le Centre de la Mer de Biarritz ont pour objectifs de mettre à disposition et de confronter ces avancées aux besoins rencontrés par les professionnels européens des zones littorales.

Les zones côtières doivent représenter un élément de la coopération internationale entre les pays disposant d'un savoir faire reconnu et ceux ne possédant pas les potentialités d'intervention adéquates. Ces différentes initiatives permettront de diffuser les connaissances et les innovations issues de recherche menées par les partenaires du projet dans les domaines de la connaissance et de la gestion des risques côtiers.

Ainsi, en s'appuyant sur la Fédération Européenne des Réseaux de Coopération Scientifique et Technique (FER) et l'Accord partiel-ouvert (EUR-OPA) sur les risques naturels et technologiques majeurs, un réseau de partenaires européens spécialisés dans la gestion des risques permettra d'offrir aux participants une vision globale des meilleures pratiques européennes dans ce domaine.

Bibliographie

CerCo, 2008, « Projet de Centre Européen Spécialisé sur les Risques Côtiers », *55ème réunion du Comité des correspondants permanents, Accord européen et méditerranéen sur les risque majeurs (EUR-OPA)*, Istanbul.

Bastien J.M.C., Scapin D.L., 1993, Ergonomic criteria for the evaluation of human-computer interfaces, Rapport Technique n° 156, Institut National de Recherche en Informatique et en Automatique, Rocquencourt, France.
< http://www.ergoweb.ca/criteres.html >

Heywood, I.; Cornelius, S.; Carver, S.J., 2006, An Introduction to Geographical Information Systems. 3rdth ed. Prentice Hall/Pearson Education Hill.

Klein, R.J.T., Nicholls, R.J., 1999, Assessment of Coastal Vulnerability to Climate Change », Ambio 28, 182–187.

Roy, 1985. « Méthodologie multicritère d'aide à la décision ». Economica, Paris.

Un système d'information pour le suivi et l'évaluation de la qualité des cours d'eau

Corinne Grac* — **Agnès Braud**** — **Florence Le Ber*,***** — **Michèle Trémolières***

* *LHyGeS UMR 7517 - ENGEES, UdS, CNRS, F 67000 Strasbourg*
{corinne.grac, florence.leber}@engees.unistra.fr, tremolie@unistra.fr
** *LSIIT UMR 7005, UdS, CNRS, F 67400 Illkirch*
agnes.braud@unistra.fr
*** *LORIA UMR 7503, F 54500 Vandœuvre-lès-Nancy*

RÉSUMÉ. *La directive cadre européenne sur l'eau (2000) impose la mise au point de nouveaux outils pour l'évaluation et le suivi de la qualité des masses d'eau. Dans ce but, nous avons réalisé depuis 2005 divers prélèvements sur un ensemble de stations (en rivières) de la plaine d'Alsace. Une base de données a été conçue pour organiser et partager les informations collectées : informations relatives aux espèces présentes dans les cours d'eau alsaciens et résultats des prélèvements (physiques, chimiques et biologiques) effectués sur les stations. Pour simplifier et enrichir l'analyse de ces informations nous avons développé un ensemble d'outils comprenant : l'interrogation de la base, la visualisation sur carte des stations respectant des critères donnés, un classifieur traitant les caractéristiques des stations. Notre objectif à terme est de constituer un système d'information utilisable à l'échelle du bassin Rhin-Meuse.*

ABSTRACT. *The European Water Framework Directive (2000) requires the development of new tools for monitoring and assessing the quality of waterbodies. Following this aim, we collected since 2005 various data from selected sites (of streams) in the Alsace Plain. A database was built to organise and share the whole information: information about the species living in alsacian streams and (physical, chemical and biological) data collected on the sites. Besides we developed tools to facilitate and enrich the analysis of this information: a query tool, a map-based visualisation tool, and a classification tool based on site characteristics. Our further aim is to develop an information system that could be used in the Rhin-Meuse watershed.*

MOTS-CLÉS : *état biologique des masses d'eau, rivières, indices biologiques, système d'information*

KEYWORDS: *biological quality of waterbodies, streams, biological indices, information system*

Introduction

La qualité des eaux de surface est un problème majeur en Europe, comme l'a souligné la Directive Cadre Européenne sur l'Eau (DCE), datant de l'année 2000. L'évaluation de la qualité de l'eau sur les seuls critères physico-chimiques est apparue comme insuffisante depuis les années 70 et l'usage complémentaire d'outils biologiques, tel que le premier indice français, l'IBGN (Indice Biologique Global Normalisé), basé sur sur les invertébrés, s'est généralisé depuis 1992 (AFNOR, 1992). Depuis 2000, quatre autres indices ont été normalisés en France, mais leur utilisation conjointe afin d'évaluer l'état d'un écosystème dans son ensemble n'a pas été immédiate (Bazerques, 2004) et elle n'a débuté que depuis peu.

Notre projet a pour but de proposer un tel outil global d'évaluation, au moyen d'un système d'information permettant de gérer et d'analyser les différentes données concernant les systèmes aquatiques de la plaine d'Alsace. Pour cela, nous avons d'abord développé une base de données regroupant les données existantes, collectées sur environ 700 stations de l'hydroéco-région de la plaine d'Alsace (Wasson *et al.*, 2002) depuis 20 ans pour les plus anciennes : il s'agit de données physiques, physico-chimiques, floristiques et faunistiques. La base inclut également les informations utiles à l'analyse et la synthèse de ces données, en particulier les caractéristiques des taxons (traits biologiques et écologiques) ainsi que les seuils de qualité (physique, chimique et biologique) et les valeurs des différents indices biologiques français calculés sur les stations. Par la suite, la base a été dotée d'une interface permettant d'accompagner les biologistes dans leurs analyses, au travers d'outils de recherche et de recoupement des informations stockées dans la base. Cette interface autorise différentes vues et en particulier un accès cartographique de l'information. Finalement nous développons un classifieur permettant de comparer les caractéristiques des stations et d'en donner une évaluation globale. La base de données originelle est ainsi complétée d'un ensemble d'outils cohérents qui en font un véritable système d'information pour l'évaluation et le suivi de la qualité des cours d'eau.

Cet article présente les différents aspects du système d'information ainsi développé. La première partie s'attache à la description de la base de données, la seconde à la description des interfaces avancées et la dernière à la description du classifieur en cours de développement. Nous discutons de notre approche puis concluons sur l'intérêt et les perspectives de ce projet.

1. Structure et contenu de la base de données

La base contient différents types de données sur les stations de cours d'eau : des données environnementales telles que le débit et le temps (climat) au moment du prélèvement ; des données physiques, concernant l'état hydromorphologique du cours d'eau ; des données chimiques, telles que les taux de nitrates, phosphates, matières organiques présents dans l'eau ; des données floristiques ; des données faunistiques. Les données floristiques recouvrent les diatomées (algues) et les macrophytes (ou

hydrophytes). Les données faunistiques concernent les invertébrés, les oligochètes benthiques et les poissons. Une partie des données a été collectée par le laboratoire LHyGeS (et anciennement le CEVH), une autre partie provient d'organismes publics tels que l'Office National de l'Eau et des Milieux Aquatiques (ONEMA) et l'Agence de l'Eau Rhin Meuse (AERM). Les plus anciennes données concernent essentiellement les aspects physico-chimiques et les macrophytes (Trémolières *et al.,* 1994, Trémolières, 2004). Les données récentes, collectées dans le cadre du projet INDICES (2005-2009) (Grac *et al.,* 2009) concernent tous les compartiments biologiques pour une quarantaine de stations, choisies parmi les sept types de cours d'eau présents en plaine d'Alsace. Les méthodes d'échantillonnage utilisées pour les relevés floristiques et faunistiques sont les méthodes normalisées des indices biologiques, modifiées selon les recommandations du programme européen de recherches AQEM[1] ; pour les invertébrés, nous avons suivi le protocole établi dans (Usseglio-Polatera *et al.,* 2004).

La base de données développée (Ehrhard, 2005) obéit au format national SANDRE[2] pour les données aquatiques. Elle contient 38 tables. Les principales tables concernent la description des stations et des données physico-chimiques et hydrobiologiques : stations échantillonnées, dates d'échantillonnage, conditions environnementales, méthodes d'échantillonnage, résultats chimiques, résultats biologiques (voir figure 1).

Les autres tables contiennent des informations sur les différents paramètres, en particulier les paramètres chimiques et biologiques. Par exemple, chaque taxon floristique ou faunistique est représenté par sa nomenclature taxonomique dans une table « Taxon ». Cette table provient du SANDRE (environ 2000 enregistrements à l'époque de la création de la base) et a été complétée (environ 4500 enregistrements). Trois tables reliées à la table « Taxon » détaillent les caractéristiques de chaque taxon. Les informations concernant les traits biologiques et écologiques des taxons présents en Alsace ont également été enregistrées dans des tables spécifiques. Ces informations ont été collectées dans la littérature (Willby *et al.,* 2000, Usseglio-Polatera *et al.,* 2002, Tachet *et al.,* 2000, Van Dam *et al.,* 1994) et adaptées ou complétées pour le contexte local.

La base a été implantée sous MySQL. Elle a été alimentée régulièrement au cours du projet INDICES avec des données existantes et suite aux prélèvements effectués et analyses faites en laboratoire. À l'heure actuelle, les données concernant quasiment tous les prélèvements effectués au cours du projet sont saisies et tous les paramètres utiles à l'évaluation de la qualité de l'eau sont renseignés, permettant ainsi une analyse complète sur un ensemble de stations représentatif de la plaine d'Alsace.

---

1. http ://www.aqem.de/
2. http ://sandre.eaufrance.fr/

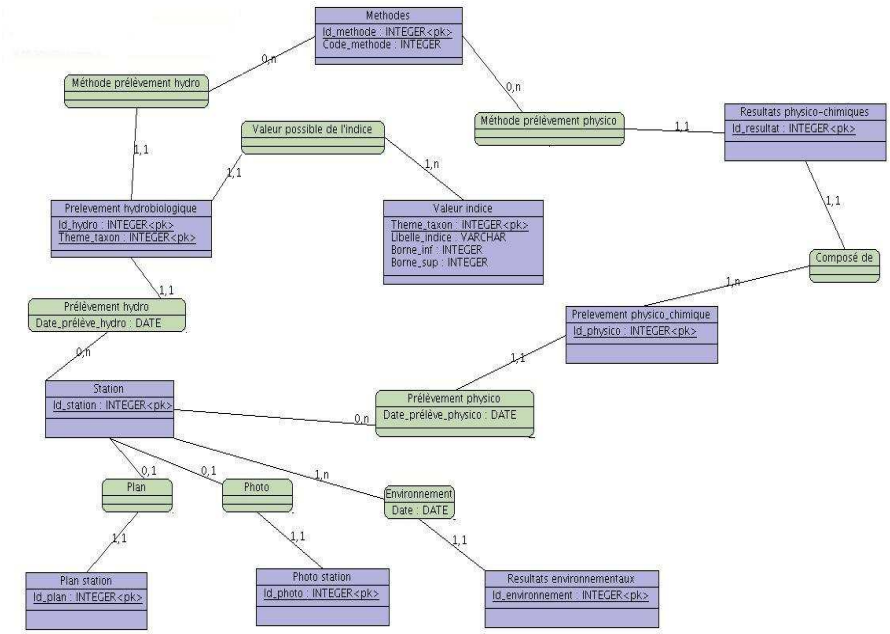


Figure 1. *MCD partiel de la base de données : les tables décrivant les stations, les méthodes d'échantillonnage et les résultats, avec leurs liens*

2. Interfaces avancées

Les données enregistrées dans la base sont utilisées par différentes personnes, enseignants-chercheurs et étudiants pour le moment, mais ingénieurs ou techniciens ultérieurement. L'entrée naturelle, pour ces utilisateurs, est une entrée cartographique, car elle permet de relier le contenu de la base au terrain qu'ils ont enquêté ou qu'ils veulent diagnostiquer. Les données enregistrées dans la base sont également utilisées dans différentes perspectives : recherche de stations ayant les mêmes caractéristiques (mêmes valeurs d'indices biologiques, par exemple), possédant tels ou tels taxons (macrophytes, poissons, …), échantillonnées à telle ou telle période ; ou bien recherche des stations où se trouvent tels ou tels taxons à différentes périodes, etc. Enfin, les utilisateurs veulent pouvoir extraire les résultats de leurs requêtes sous forme de tableaux sur lesquels ils pourront ensuite faire des analyses.

Ces besoins ont été établis progressivement et nous ont conduites à développer une interface cartographique doublée d'une interface de requête avancée, permettant de sélectionner simultanément un ensemble de stations *via* la carte, puis de rechercher les taxons s'y trouvant. Outre ces informations, le tableau extrait peut contenir des informations synthétiques, calculées à la volée (consultation des stations, figure 2).

Symétriquement, nous développons une vue permettant de sélectionner des taxons et de visualiser puis extraire les stations où ils se trouvent (consultation des taxons). L'ensemble est accessible via un site web, rendu nécessaire par la dispersion physique des utilisateurs. Le choix d'une interface *ad hoc* plutôt que d'un couplage de la base avec un système d'information géographique s'est fondé sur différentes raisons dont nous discutons plus loin.

Figure 2. *Vue de l'interface : consultation des stations*

3. Principes du classifieur

Un des objectifs du projet INDICES, et donc de la constitution du système d'information, est de comparer et combiner les valeurs d'indices biologiques. Pour un ensemble de 40 stations, nous possédons en effet les informations nécessaires pour calculer les cinq indices biologiques normalisés français, à savoir l'IBD (établi sur les diatomées), l'IBGN (invertébrés), l'IBMR (macrophytes), l'IOBS (oligochètes) et l'IPR (poissons). Ces valeurs d'indices sont ensuite transformées en cinq classes de qualité (cf. figure 2, en bas de l'interface).

L'idée du classifieur, issue de réflexions sur le système d'harmonisation proposé dans (Lafont *et al.*, 2001), est d'évaluer la qualité d'une station en l'associant automatiquement à un "profil caractérisé". Ce profil correspond à un ensemble de stations

connues, stockées dans le système d'informations, et partageant des caractéristiques communes, à partir desquelles ont peut définir un profil. Par exemple, les stations possédant les classes de qualité suivantes (IBGN=[1,2], IOBS=3, IBMR=3, IBD=[4,5]) sont associées au profil suivant : *"début de dégradation des sédiments, forte dégradation physico-chimique au moins liée à un niveau trophique moyen, mais hors matière organique, bon potentiel de résilience général et possibilité de résilience sur les sédiments"*. L'intérêt ici pour les biologistes est de disposer d'une évaluation globale des stations, y compris dans les compartiments dont ils ne sont pas spécialistes. De plus, ce classifieur permet d'observer l'évolution temporelle d'une station pour laquelle on dispose d'un suivi pluriannuel.

Pour déterminer de tels ensembles de stations, nous utilisons la technique des treillis de Galois (Barbut *et al.,* 1970). Un algorithme de construction de treillis de Galois prend en entrée un ensemble d'objets, un ensemble de propriétés, et une relation d'incidence précisant pour chaque couple (objet, propriété) si l'objet possède la propriété. Il fournit en sortie un ensemble hiérarchisé de concepts, c'est-à-dire des groupes d'objets définis par leurs propriétés communes. Il permet également de générer les règles d'association entre ces propriétés.

La démarche de construction du treillis est présentée dans (Braud *et al.,* 2009). Les propriétés prises en compte pour définir les groupes ont été dans un premier temps les cinq indices biologiques normalisés, et nous travaillons à l'heure actuelle sur l'intégration des paramètres physico-chimiques. La structure hiérarchique obtenue permettra alors de parcourir efficacement l'ensemble des concepts afin de répondre à des requêtes, comme cela a été fait dans les domaines biologique ou géographique (Messai *et al.,* 2008, Bedel *et al.,* 2007). Nous comptons utiliser une approche similaire pour évaluer une nouvelle station, en déterminant le concept auquel elle appartient ou ceux dont elle se rapproche le plus.

4. Discussion

Les fonctionnalités proposées dans ce système d'information viennent combler un manque d'outils exprimé par les chercheurs en biologie du LHyGeS pour ce qui concerne certaines de leurs tâches.

Un premier problème concerne le stockage et la manipulation des données collectées sur le terrain. Elles sont stockées classiquement dans des feuilles de tableur de grande taille et les recoupements sont effectués à la main, engendrant un travail fastidieux et des risques d'erreurs. À l'inverse, le système de gestion de base de données apporte des garanties en termes de cohérence et de sécurité des données. Le système d'information offre également des fonctionnalités de requêtes avancées qui apportent un mécanisme puissant de recherche et de recoupement de données issues de sources multiples.

Un second problème concerne la création de cartes permettant de visualiser la répartition géographique des valeurs des paramètres mesurés sur les stations. Jusqu'à

présent ces cartes sont créées à la main : un tableau contenant les informations à visualiser doit être constitué (étape manuelle) puis enregistré dans un système d'information géographique (SIG) qui affiche alors les données thématiques sur la carte. Grâce à l'interface cartographique du système d'information, les cartes sont construites par une simple requête et, de plus, l'ensemble des informations concernant une station ou plusieurs stations sélectionnées est accessible par simple clic sur ces stations.

Pour répondre à ces différents besoins, et en particulier aux besoins cartographiques, il a été tout d'abord envisagé d'utiliser un SIG, couplé à la base de données. Nous avons mené une étude comparative confrontant différents SIG et un développement *ad hoc* (Buleandra, 2007). Nous avons considéré uniquement des SIG Open Source, tels que GRASS[3], GeOxygene[4], ou GvSIG[5], afin de pouvoir adapter le SIG choisi à nos besoins. À l'issue de cette étude, nous avons opté pour un développement *ad hoc*, pour les raisons suivantes :

– les données considérées ont une spatialité faible, seules les stations doivent être localisées et ceci par un point, leur extension spatiale n'étant pas précisée ;

– les besoins en fonctionnalités cartographiques se limitent à l'heure actuelle au positionnement de stations sur une carte, et éventuellement plus tard à la détermination d'un voisinage ;

– en revanche de nombreuses fonctionnalités sont nécessaires pour aider les biologistes dans le recoupement d'informations issues de plusieurs sources ;

– la visualisation sur carte est une fonctionnalité supplémentaire offerte aux biologistes pour faciliter la recherche des informations liées aux stations, sachant qu'il est plus facile de les identifier visuellement, par zooms successifs, que par un code.

Il apparaît ainsi que le recouvrement entre les nombreuses fonctionnalités offertes par les SIG et nos besoins spécifiques est assez faible. De plus, utiliser un SIG impose certaines contraintes alors qu'un développement *ad hoc* permet une plus grande souplesse pour obtenir les résultats attendus. En l'occurrence, GvSIG était l'option la plus appropriée, mais il n'y avait pas de version disponible stable à l'époque. Une étude menée aujourd'hui nous conduirait peut-être à des conclusions différentes.

Finalement, et ceci justifie largement notre choix, faire un développement *ad hoc* nous a permis d'adapter en continu l'interface aux biologistes qui en sont destinataires. L'outil est, à leurs yeux, simple, intuitif, et ne nécessite pas de formation. Ceci est dû au fait qu'il est conçu en collaboration directe avec les biologistes (chercheurs et étudiants) qui font leurs propres propositions, testent et valident les différents affichages et les fonctionnalités.

---

3. http ://grass.itc.it/
4. http ://oxygene-project.sourceforge.net/
5. http ://www.gvsig.gva.es/

5. Conclusions et perspectives

Le système d'information présenté dans cet article a été conçu dans le cadre du projet INDICES et dans l'objectif d'intégrer tous les éléments nécessaires à l'étude d'outils d'évaluation globale de la qualité des cours d'eau tels que demandés par la DCE. Il rassemble pour cela une grande variété d'informations ayant trait à l'évaluation des cours d'eau de la plaine d'Alsace. Les données stockées sont issues de bases de données nationales faisant référence (pour les taxons par exemple), issues d'une synthèse bibliographique réalisées par les biologistes du LHyGeS (traits biologiques), ou résultats d'analyses faites sur des prélèvements. Ces prélèvements ont en particulier été effectués sur un échantillon de stations sélectionnées en plaine d'Alsace pour représenter les différents types de stations identifiés par la DCE.

Pour permettre une exploitation efficace de ce système d'information par les biologistes et les aider à valoriser et recouper l'ensemble des informations, nous avons conçu différents outils d'interrogation simple ou avancée, de visualisation sur carte, de classification automatique de stations. Ces outils sont intégrés dans un site web. L'ensemble est simple d'utilisation et accessible depuis n'importe quel poste sans nécessité d'installer un logiciel autre qu'un navigateur.

Les agences de l'eau ont développé des bases de données où sont recensées les informations sur les nombreuses stations qu'elles surveillent. Certaines offrent une visualisation cartographique, comme l'Agence de l'Eau Loire-Bretagne, avec l'interface Osur Web[6]. Toutefois, même si elles recouvrent des zones géographiques étendues, les informations disponibles dans ces bases sont très limitées, par exemple on ne dispose généralement pas des relevés taxonomiques établis sur les stations.

Sur cette même thématique, l'ONEMA a commandé le projet SEEE-cours d'eau (Système d'Évaluation de l'État de l'Eau). Il s'agit de rassembler des banques de données afin de permettre la mise au point puis l'intégration d'outils d'évaluation. Nos objectifs sont similaires, les principales différences viennent des dimensions des deux projets : le SEEE a un cadre national tandis que nous travaillons sur la plaine d'Alsace ; de plus, le SEEE a pour vocation de réaliser l'évaluation DCE des cours d'eau, autrement dit, en se fondant sur la plus mauvaise valeur, alors que nous considérons tous les indices ensemble.

Par ailleurs, le projet INDICES ne se limite pas aux seules données nécessaires au calcul des indices biologiques ou physico-chimiques, mais vise à intégrer des informations supplémentaires, tels que les traits biologiques et écologiques des taxons, pour construire un outil d'évaluation de la qualité biologique globale des cours d'eau (Bertaux *et al.*, 2009). Finalement nous mettons en œuvre des outils innovants, comme les treillis de Galois, afin d'explorer au mieux les données collectées. L'approche fondée sur le classifieur sera prochainement finalisée et testée sur des données équivalentes recueillies en Bretagne. Ceci pourrait conduire à confronter des expertises différen-

6. http ://carto.eau-loire-bretagne.fr/osur/

ciées, les espèces considérées pour construire les indices pouvant être plus ou moins présentes dans les différentes régions.

Remerciements

Les auteurs remercient l'Agence de l'Eau Rhin-Meuse et l'ONEMA pour leur soutien à ce projet.

6. Bibliographie

AFNOR, « Qualité de l'eau : détermination de l'Indice Biologique Global Normalisé (IBGN) », 1992. NF T90-350.

Barbut M., Monjardet B., *Ordre et classification – Algèbre et combinatoire*, Hachette, 1970.

Bazerques M.-F., « Directive-cadre sur l'eau : le bon état écologique des eaux douces de surface : sa définition, son évaluation », 2004, Communication au Ministère de l'Écologie et du Développement Durable, Paris.

Bedel O., Ferré S., Ridoux O., Quesseveur E., « GEOLIS : a logical information system for geographical data », *Revue Internationale de Géomatique*, vol. 17, p. 371-390, 2007.

Bertaux A., Le Ber F., Braud A., Trémolières M., « Identifying ecological traits : a concrete FCA-based approach », *7th International Conference on Formal Concept Analysis, ICFCA 2009, Darmstadt*, LNAI 5548, Springer-Verlag, p. 224-236, 2009.

Braud A., Grac C., Pristavu S., Dor E., Le Ber F., « Une démarche fondée sur les treillis de Galois pour l'aide à la qualification de l'état des milieux aquatiques », *Actes du 2ème Atelier « Systèmes d'Information et de Décision pour l'Environnement » - SIDE 2009, Toulouse*, p. 94-105, 2009.

Buleandra M., « Visualisation de données sur la qualité des cours d'eau en Alsace », Mémoire de stage Erasmus, LSIIT, Illkirch, France et Université Dunarea de Jos, Galati, Roumanie, 2007.

Ehrhard J.-L., « Mise en œuvre d'un système de comparaison des réponses des indices biologiques sur les cours d'eau de la plaine d'Alsace », Mémoire de diplôme d'ingénieur CNAM en informatique, Strasbourg, 2005. CEVH.

Grac C., Le Ber F., Herrmann A., Trémolières M., Programme de recherche-développement *Indices* – Rapport d'avancement scientifique de la deuxième année (2008), Contrat pluriannuel 1463 de l'Agence de l'Eau Rhin-Meuse, CEVH, 2009.

Lafont M., Vigneron S., Fournier A., Evaluation de l'effet des rejets polluants sur les milieux aquatiques situés dans des environnements imperméabilisés : Proposition d'une approche intégrée, Rapport 01-0784, Cemagref, 2001.

Messai N., Devignes M.-D., Napoli A., Smaïl Tabbone M., « Correction et complétude d'un algorithme de recherche d'information par treillis de concepts », *Classification : points de vue croisées*, Revue des Nouvelles Technologies de l'Information (RNTI), Cépaduès Éditions, p. 147-158, 2008.

Tachet H., Richoux P., Bournaud M., Usseglio-Polatera P., *Invertébrés d'eau douce : Systématique, biologie, écologie*, CNRS Éditions, 2000. 588 pages.

Trémolières M., Carbierner R., Orstcheit A., Klein J.-P., « Changes in aquatic vegetation in Rhine floodplain streams in Alsace in relation to disturbance », *Journal of Vegetation Science*, vol. 5, p. 169-178, 1994.

Trémolières M., « Fiches descriptives des habitats aquatiques », *Référentiel des habitats reconnus d'intérêt communautaire de la bande rhénane*, Conservatoire des Sites Alsaciens et Office National de Forêts, p. 73-110, 2004. Programme LIFE Rhin vivant.

Usseglio-Polatera P., Beisel J.-N., « Longitudinal changes in macroinvertebrate assemblages in the Meuse river : anthropogenic effects versus natural change », *River Res. Applic.*, vol. 18, p. 197-211, 2002.

Usseglio-Polatera P., Wasson J.-G., « Protocole de prélèvement et de traitement des échantillons des macro-invertébrés benthiques sur les sites de référence "cours d'eau" », Université de Metz et Cemagref de Lyon, 2004. 7 pages.

Van Dam H., Mertens A., Sinkeldam J., « A coded checklist and ecological indicator values of freshwater diatoms from The Netherlands », *Aquatic Ecology*, vol. 28, n° 1, p. 117-133, 1994.

Wasson J., Chandesris A., Pella H., Blanc L., « Les hydroécorégions de France métropolitaine - approche régionale de la typologie des eaux courantes et éléments pour la définition des peuplements de référence d'invertébrés », Rapport 02-0413, Cemagref, 2002. 190 pages.

Willby N., Abernethy V., Demars B., « Attribute-based classification of European hydrophytes and its relationship to habitat utilization », *Freshwater Biology*, vol. 43, p. 43-74, 2000.

ObServe : Un système d'acquisition et de gestion de données d'observations

Application à la pêche thonière

Pascal Cauquil* — **Thérèse Libourel**** — **Christelle Pierkot**** — **Anthony Tissot***** — **Julien Tornare*****

** IRD, CRHMT (Centre de Recherche Halieutique Méditerranéenne et Tropicale)*
UMR 212 EME (Ecosystèmes Marins Exploités)
Avenue Jean Monnet
34200 SETE
prenom.nom@ird.fr

*** LIRMM*
161 rue ADA
34095 Montpellier Cedex 5
nom@lirmm.fr

****Université de Montpellier 2*
Prenom.nom@etud.univ-montp2.fr

RÉSUMÉ. *L'Observatoire Thonier (OT), un dispositif de l'Institut de Recherche pour le Développement (IRD), a en charge le suivi de la pêche thonière tropicale française et conduit en particulier un programme d'observateurs embarqués à bord des navires. Plusieurs types de données (captures, efforts, mensurations, etc.) sont collectées dans le cadre de l'OT (certifié ISO 9001) et l'un des objectif majeurs est l'amélioration continue de la qualité des données statistiques produites. Pour y parvenir, un nouveau système d'information, ObServe, a été mis au point pour gérer la collecte des données d'observation. L'article présente les diverses réflexions menées dans le contexte de ce projet. Celles-ci portent sur les divers points suivants : création d'une base conforme à un schéma "générique" des données relatives aux observations nécessaires, acquisition et optimisation de ces observations (potentialité d'automatisation par usage de capteurs), intégration des données et traitements effectués.*

ABSTRACT. *The Tropical Tuna Observatory (OT), a team from the French Research Institute for Development (IRD), is in charge for monitoring the French tropical tuna fishery and leads in particular a scientific observer program on board of vessels. Several types of data (catches,*

effort, length measurements, etc.) are collected. One major objective of the OT is the continuous improvement of the quality of statistical data produced. To achieve this, has been set up a new information system, ObServe, intended to manage observation data. This paper presents the diverses reflexions conducted in the context of this project. These address various issues: creation of a database conforming to a generical schema, acquisition and optimization of these observations (potential automation by the use of sensors), data integration and treatments.

MOTS-CLÉS : *Acquisition de données, Capteurs, SI ObServe, Intégration*

KEYWORDS: *Data acquisition, Sensors, ObServe IS, Integration*

---

1. Introduction

Dans le cadre du programme européen de suivi des pêcheries Data Collection Framework (DCF), l'IRD (Institut de Recherche pour le Développement) est maître d'oeuvre pour l'Union Européenne et la Direction des Pêches Maritimes et de l'Aquaculture (DPMA) française d'un système de suivi des pêches.

La pêche thonière tropicale française de surface se déroule dans les océans Atlantique et Indien. Elle concerne une trentaine de bateaux qui opèrent à la canne et à la senne[1] tournante et capturent chaque année en moyenne 150 000 tonnes de thons tropicaux.

Plusieurs programmes d'observations ont été mis en place pour effectuer ce suivi. Chacun des programmes a eu ses spécificités (e.g. protocoles, espèces d'intérêt, granularité de la donnée), mais on retrouve toujours un noyau de données communes. Pour chacun d'entre eux, des observateurs embarqués se sont succédés afin de collecter manuellement ce noyau de données.
Cependant, malgré le soin apporté à la collecte et à la gestion de ces observations, on constate fréquemment en fin de chaîne que des erreurs importantes et grossières ponctuent les jeux de données obtenus. Une première hypothèse avancée est que la majorité de ces erreurs pourraient être identifiées et écartées dès la phase d'acquisition à l'aide d'outils informatiques adaptés.

Dans ce but, la mise en place d'un système d'information permettant de gérer la collecte et le stockage des données d'observation a donc été décidée. Ce projet dénommé ObServe, doit atteindre les objectifs suivants :

1) le développement d'un modèle de données robuste destiné aux données d'observation de la pêche et suffisamment générique pour pouvoir recevoir les données des programmes d'observation passés, présents et à venir,

2) la création d'une base de données conforme au modèle précédent et la récupération des données d'observation historiques au sein de cette base,

3) le développement d'un logiciel d'acquisition capable de transférer directement les observations au sein de cette base et ceci afin de limiter les étapes intermédiaires de ressaisie et de migration qui pouvaient générer des erreurs dans les données,

4) Le nombre d'observations diverses et le caractère souvent fastidieux ou contraignant de celles-ci, nous a aussi amené à réfléchir sur l'automatisation de l'observation elle-même via l'usage de capteurs.

L'article présente donc les grandes étapes de la mise en place du projet ObServe et des ses extensions.

Dans ce papier, nous présentons en section 2 en quoi consiste les observations : données à collecter et erreurs les plus fréquemment rencontrées. Dans la section 3,

1. La senne est une technique de pêche qui consiste à capturer les poissons à la surface en pleine eau en l'encerclant à l'aide d'un filet.

nous discutons de la nécessité d'intégrer les données dans un système d'information et nous présentons le modèle de données mis en place au sein du projet. La suite de cet article concerne les perspectives de ce travail et notamment l'automatisation de l'acquisition des observations grâce à la potentialité nouvelle offerte par les capteurs. Nous décrivons dans la partie 4.1, les capteurs SunSpot qui servent de base à notre étude. Puis, nous montrons dans la partie 4.2, comment restituer ces nouvelles informations dans le projet ObServe. Enfin, nous concluons dans la partie 5.

2. Données d'observation

L'acquisition de données d'observation est indispensable pour construire l'information statistique utilisée pour la gestion des ressources halieutiques et des pêcheries. Chaque année, de telles données sont collectées sur le terrain par des membres de l'Observatoire Thonier Tropical (OT), une équipe IRD de l'UMR Ecosystèmes Marins Exploités (EME).

Ces données constituent les statistiques officielles de la France et sont communiquées aux Commissions Internationales thonières : la Commission Internationale pour la Conservation des Thonidés de l'Atlantique (CICTA) et la Commission Thonière de l'Océan Indien (CTOI).

2.1. *Méthodes d'acquisition*

L'OT entretient trois filières de fourniture de données :

– **La collecte des journaux de bord**. Il s'agit des relevés rédigés par l'équipage et qui cataloguent les activités menées à bord pendant les marées. Ces données sont disponibles et collectées pour presque la totalité des marées. Elles donnent des informations sur les poids totaux pêchés et les efforts de pêche mais sont approximatives.

– **Les enquêtes aux débarquements**. Ces enquêtes sont menées aux deux ports de débarquements des canneurs et senneurs objets du suivi (Abidjan en Côte d'Ivoire et Victoria aux Seychelles), par des techniciens de l'IRD. Ces enquêtes parviennent à couvrir l'ensemble des débarquements. Elles consistent en des échantillonnages de taille et de composition spécifique réalisés à l'intérieur des cuves des thoniers au cours du débarquement. Elles fournissent donc des informations précises mais parcellaires.

– **L'embarquement d'observateurs scientifiques à bord**. Ces observateurs embarquent à bord des bateaux et réalisent l'intégralité de la marée sur laquelle ils se sont engagés. Cela leur permet de collecter des informations complètes et très précises sur cette marée, en particulier sur les prises accessoires et les rejets. Par contre, seules environ 10% des marée sont couvertes par ce programme.

Chaque filière permet de conduire par la suite divers types d'analyses statistiques. En effet, l'objectif majeur de la collecte de données sur les activités des navires de pêche est de parvenir à estimer, pour chaque année, le poids et le nombre total de

poissons pêchés. Cette estimation quantitative est rendue possible par l'utilisation simultanée des données d'enquêtes et des livres de bord, en leur appliquant une série de corrections et d'extrapolations parfaitement définies. Les données d'observation quant à elles, sont essentiellement utilisées pour analyser de façon qualitative les espèces qui font l'objet de prises accessoires et de rejets d'une part, et l'évolution des stratégies de pêche d'autre part.

2.2. *Données collectées et erreurs liées aux collectes*

Plusieurs programmes d'observateurs embarqués se sont succédés depuis 1995 afin d'obtenir des données sur les rejets (essentiellement des thons de petites tailles) et les prises accessoires qui concernent une cinquantaine d'espèces (Stretta *et al.,* 1997), (Gaertner *et al.,* 1998), (Goujon, 2004), Moratoire (1997-2006) (Gonzalez *et al.,* 2007), (Romanov, 2002), (Amandè *et al.,* 2008), (Romanov, 2008) et actuellement le programme DCF (2003-2013).

Un des axes importants de recherche est l'étude de l'effort et des stratégies de pêche appliquées par les équipages pour localiser les bancs de thons. Dans le protocole d'observation, cet intérêt est traduit par la collecte de plusieurs paramètres :

– Le comportement général du bateau pendant une marée : la position géographique, la vitesse, des paramètres environnementaux (température de surface et vitesse du vent), ou encore l'activité éventuelle d'autres bateaux dans la même zone.

– La présence d'éléments susceptibles de révéler la présence d'un banc de thons et d'influencer les choix de l'équipage (les systèmes observés) : les oiseaux, les objets flottants, ...

– Les conditions de capture lors des coups de pêches réussis : banc libre ou dispositif de concentration de poissons, durée de l'opération, ...

– La quantification et l'échantillonnage des captures accessoires (les espèces non visées qui seront conservées pour un marché parallèle ou rejetées en mer).

– La ou les raisons du rejet d'espèces recherchées : cuves pleines, taille trop petite ou poisson trop abîmé, ...

– Les paramètres de capture des espèces ciblées[2] : tonnage capturé, discrétisation par espèces et par catégories de poids, ...

Les données des programmes d'observation historiques, de même que les données du programme DCF actuel collectées avant le lancement du projet ObServe, ont

2. Ces informations peuvent être aussi récupérées via le livre de bord du capitaine et l'enquête du port, mais le relevé effectué à bord par l'observateur scientifique est nettement plus précis. Ces observations permettent alors d'apprécier le niveau de qualité des données issues des deux autres filières de collecte, qui sont les seules disponibles pour les 90% de marées non observées, et d'instaurer en conséquence les règles de correction et d'extrapolation qui amèneront les estimations à un niveau de qualité convenable en termes de poids total capturé et de composition spécifique.

toujours été récoltées au travers de formulaires papier, puis informatisées localement sous forme de fichiers MS Excel ou de tables MS Access non structurées. Une analyse de ces jeux de données a été effectuée dans le but d'identifier les erreurs les plus fréquemment rencontrées.

Cette étude a montré que les erreurs sont rarement dues à de mauvaises observations mais qu'elles se produisent plutôt au moment de leur transcription sur le papier. Beaucoup d'erreurs viennent d'une mauvaise utilisation des formulaires : unités non respectées, indication d'un type de mesure incohérent avec la mesure, etc. Certains champs normalement obligatoires sont non renseignés comme les positions géographiques ou les références entre formulaires. D'autres relèvent de l'inattention (erreurs sur les horaires). D'autres encore sont dues à des erreurs de transcription entre brouillon et formulaire, ou entre formulaire et saisie informatique. Toutes ces erreurs témoignent de la nécessité d'améliorer à la fois la structuration des observations et la méthode de saisie.

3. Le système d'information Observe

Les méthodes d'acquisition informatique et les systèmes de stockage des données collectées au cours des programmes passés, sont différents et ne permettent pas aisément une extraction de données dans un format commun afin de conduire des analyses sur de longues périodes. La mise en commun de données historiques et récentes est cependant indispensable pour analyser et mettre en évidence des changements dans les pratiques de pêche et/ou dans les communautés exploitées.

La principale idée mise en œuvre pour pallier les imperfections et répondre aux exigences précédentes, est la mise en place d'un système d'information conçu à partir d'un modèle de données adapté, correctement contraint et d'un logiciel de saisie adapté au contexte. Ce système doit contribuer de manière significative à l'étude des effets de la pêche sur les écosystèmes et à la compréhension de leur structure et de leur fonctionnement (Zeller *et al.,* 2005), il doit donc structurer l'information pertinente afin de la rendre accessible et utilisable dans le temps et par des équipes distinctes.

Les principaux objectifs du SI ont ainsi été fixés afin d'assurer une collecte de données de qualité :

1) Concevoir un modèle de données adapté au contexte de l'étude et stocker l'information durablement,

2) Proposer une interface de saisie ergonomique pour l'observateur,

3) Contraindre et vérifier les saisies de l'observateur,

4) Limiter le trajet de l'information entre son observation et sa consolidation en base centrale,

5) Utiliser des technologies informatiques pérennes.

3.1. *Le modèle de données*

En concevant le modèle de données, nous avons dû de répondre à deux objectifs : Le modèle se doit tout d'abord d'être suffisamment générique pour pouvoir accueillir, en plus des données des programmes de collecte actuel et à venir, celles des programmes passés. Ensuite, le modèle doit être facilement compréhensible et manipulable par des biologistes, premiers utilisateurs de l'information collectée. Les entités modélisées sont donc celles du métier de la pêche.

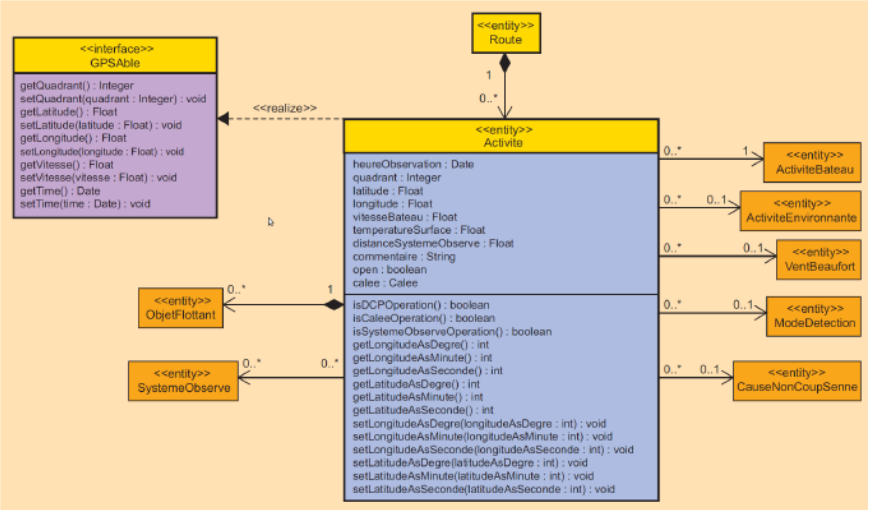


Figure 1. *Modèle de données relatif à l'entité "Activité" dans ObServe*

La figure 1 présente une vue partielle du modèle de données. Une activité correspond à une opération de collecte d'observations effectuée à un instant t et en un lieu précis. Elle est identifiée par une heure et qualifiée par une position géographique. Une activité d'observation est associée à d'autres activités comme celle du bateau et permet de définir certains paramètres environnementaux tels que la température, la vitesse du vent, les paramètres de la pêche le cas échéant, etc...

3.2. *Trajet et stockage de l'information*

Plus une information subit d'opérations, de changements de formats, de médias, plus le risque d'altération augmente. Pour améliorer la qualité des données collectées, il faut donc simplifier autant que possible la chaîne de traitements qui les conduit à leur forme finale consolidée. A ce titre, l'objectif poursuivi est d'informatiser les données aussi près que possible de leur source et de les insérer au plus vite dans la base de données.

Dans ObServe, le stockage durable des données d'observation est effectué dans une base de données centrale située à Montpellier (Cf. figure 2). Idéalement, il aurait été souhaitable que les données issues de la saisie à bord soient insérées en temps réel dans la base de données centrale, au travers du réseau. Mais le contexte de la pêche en haute mer et l'absence de liaison internet descente et exploitable a nécessité la mise en place de solutions techniques adaptées, telles que la mémorisation temporaire des données en local et la mise en oeuvre d'un système de synchronisation entre base centrale et applications de saisie[3].

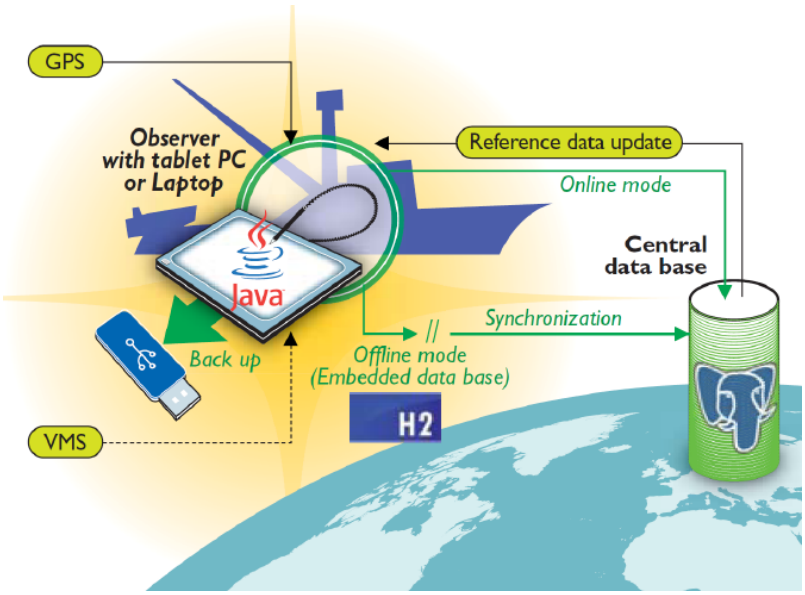


Figure 2. *Système d'information ObServe*

Afin de limiter les risques lors de la synchronisation de données, les choix de conception suivants ont été mis en place :

– Le même modèle de données relationnel a été implémenté sur les bases de données locales et sur la base de données centrale.

– La manipulation des données passe par une librairie de *mapping* qui prend en charge les particularités d'implémentations relatives aux moteurs de bases de données[4].

3. Le système dispose néanmoins des deux modes de fonctionnement à savoir le mode déconnecté, que nous venons de décrire et le mode connecté qui permet de saisir les données directement dans la base centrale.
4. PostGreSQL est utilisé pour la base de données centrale et H2 sur pour les bases de données locales

– Les enregistrements de toutes les tables possèdent des identifiants uniques permettant de résoudre aisément les conflits de mise à jour lors de la synchronisation entre la base de données locale et la base de données centrale.

4. Automatisation de l'acquisition

Afin d'améliorer les performances du système ObServe, de faciliter le travail des observateurs et de limiter encore un peu plus les erreurs de saisie, il nous semble judicieux d'automatiser l'acquisition de certaines observations. L'idée sous-jacente serait d'acquérir certaines données grâce à la technologie des capteurs et et de les intégrer directement dans la base de données locale sans intervention de l'observateur.
Dans le contexte d'ObServe, plusieurs informations telles que la géolocalisation ou encore la collecte de données environnementales (température, luminosité,...) ou contextuelles (vitesse du bateau, taille des thons, ...) pourraient ainsi être automatisées.

Afin d'assurer la faisabilité, plusieurs critères doivent être pris en considération :

– Les capteurs doivent être peu encombrants et suffisamment robustes pour être utilisés sur les ponts des bâteaux, milieu par définition hostile.

– Les capteurs doivent avoir des capacités de communication et d'autonomie suffisamment importantes pour pouvoir être utilisés sur des bâteaux de pêche lors des campagnes de relevés d'observations.

– Les données issues des capteurs doivent pouvoir être facilement intégrées dans le modèle de données existant.

Pour vérifier notre hypothèse, une première étude basée sur l'utilisation de capteurs SunSPOT est menée en collaboration avec l'université de Montpellier 2.

4.1. *Les capteurs Sunspot*

SunSPOT est une technique de réseaux de capteurs créée par l'entreprise américaine Sun Microsystems (SunMicroSystems, n.d.). Le kit de distribution comprend deux capteurs et une base. SPOT (Small Programmable Object Technology) se démarque par sa plateforme logicielle (compatible J2ME) et matérielle homogène. Les composants matériels sont tous basés sur des standards ou des architectures très répandues. Les capteurs SunSPOT proposent donc une solution simple et surtout extrêmement modulaire car reposant sur des technologies récentes, connues et pour la plupart standardisées.

Outre le fait que les SunSPOT sont très facilement maniables, l'un des principaux avantages est qu'ils possèdent trois capteurs par défaut, à savoir, un thermomètre, un luminomètre et un accéléromètre 3D. Dans le cadre du projet ObServe, l'accélérateur pourrait être mis à profit pour détecter un mouvement de giration du bateau et ainsi déduire la durée du coup de senne (l'action de pêche). Le capteur de luminosité pour-

rait quant à lui servir à enregistrer les conditions d'ensoleillement dans lesquelles les analyses du plan d'échantillonnage ont été faites.

Les SunSPOT disposent également d'une interface radio leur permettant de communiquer entre eux et avec la base. L'interface radio, bien que de portée plutôt faible (40 mètre annoncés, 20 mètres constatés) reste un outil très interressant pour la mise en réseau de plusieurs capteurs distribués. Sur un bateau de pêche, il est souvent impossible d'utiliser un ordinateur sur le pont du fait des conditions difficiles (humidité, saleté, vitesse,...), impliquant la saisie des données sur une ardoise PVC hydrophobe et le report en temps différé dans l'ordinateur ; l'interface radio des capteurs permettraient de lever cette difficulté et d'envoyer directement les données sur l'ordinateur.

Par ailleurs, les SunSPOT disposent de broches additionnelles permettant d'ajouter des capteurs supplémentaires, augmentant ainsi leurs possibilités d'exploitation. On pourrait imaginer utiliser le connecteur d'extension pour interfacer une balance marinisée et un ichthyomètre (pied à coulisse) afin de peser et de mesurer automatiquement les espèces péchées.. Mais à notre avis, l'utilisation la plus intéressante et plus qu'abordable est l'ajout d'une puce GPS pour le géoréférencement des données. Dans le cadre du projet ObServe, une puce GPS permettrait de récupérer la trace des bateaux de manière automatique[5].

Le schéma 3 propose une modélisation des entités du SunSPOT, définissant ainsi les différentes possibilités de mesures offertes par les capteurs.

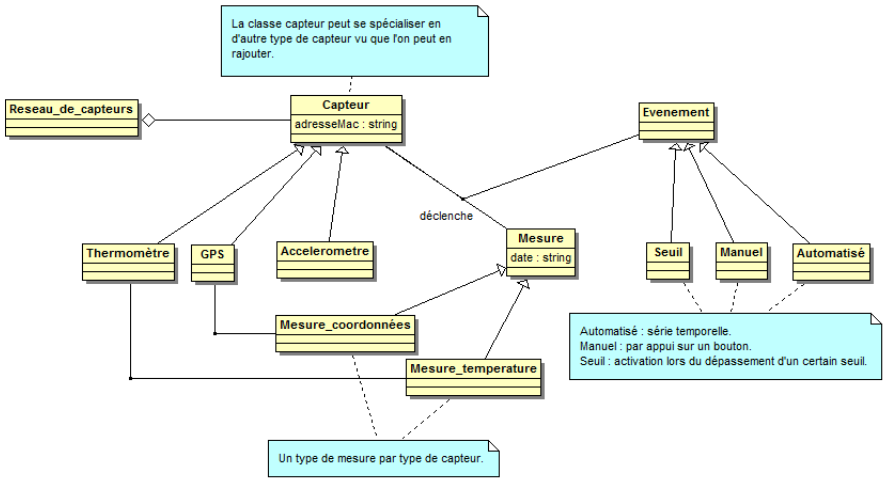


Figure 3. *Modélisation des entités SunSPOT*

5. A ce jour, l'observateur retranscrit manuellement dans le système, les informations issues d'un GPS

Enfin, les capteurs possèdent tous les éléments nécessaires à l'exécution de programmes embarqués (écrits en Java) permettant ainsi de spécifier les paramètres d'acquisition et de fonctionnement en fonction du contexte. De ce point de vue, ils sont donc des entités autonomes qui peuvent fournir une solution pour automatiser l'acquisition de données d'observation. Sachant qu'une puce GPS consomme beaucoup d'énergie, on peut imaginer conditionner l'activation de celle ci au type de mouvement du bateau capté par l'accéléromètre.

4.2. *Restitution des données issues des capteurs*

Trois possibilités sont envisageables pour restituer le résultat de la mesure :

– Le capteur est relié directement à un ordinateur et les données d'acquisition sont transmises en temps réel pour être intégrées dans la base de données locale (liaison 1 de la figure 4).

– Le capteur effectue une communication radio avec sa base qui est reliée filiairement à un ordinateur et lui communique les données en temps réel qui sont ensuite intégrées dans la base de données locale (liaisons 2 de la figure 4).

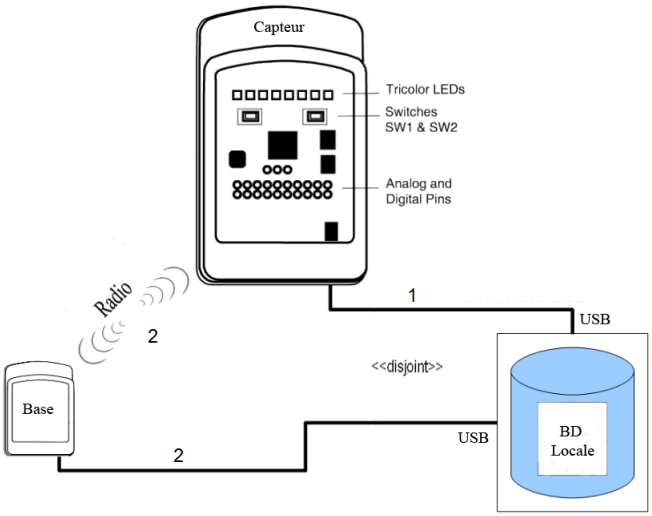


Figure 4. *Liaisons entre les SunSpot et la base de données locale*

– Le capteur est autonome est stocke les données acquises dans un journal afin de les restituer ultérieurement. Une synchronisation avec la base de données centrale est

alors nécessaire pour intégrer les observations.

Quelque soit le mode de fonctionnement choisi pour restituer le résultat des mesures des capteurs, nous sommes toujours limité par le fait qu'il n'y a pas de connection internet satisfaisante entre le bateau et la base centrale située à Montpellier. Nous sommes donc toujours contraints à recueillir les données dans une base locale ou dans un journal et à les intégrer ensuite dans la base centrale.

5. Conclusions

A ce jour, le système d'information ObServe tel qu'il a été décrit dans la section 3 est opérationnel et utilisé pour collecter les données d'observation de la pêche thonière.
L'étude se poursuit concernant l'utilisation de capteurs pour automatiser la collecte de certaines données. La liaison entre capteurs et une base de données a été établie, l'ajout d'une puce GPS est en cours de réalisation et la prochaine étape concerne l'intégration des données dans la base de données du système d'information ObServe. Cela suppose de définir des formats de données compatibles avec la modélisation définie au préalable dans le cadre du projet et de définir des formats de fichiers permettant si besoin est de collecter les données dans des journaux. Une étude de plusieurs formats (GML, KML, GPX) est en cours de réalisation afin de trouver la meilleure solution.

Les résultats issus de cette étude vont déterminer la faisabilité quant à l'utilisation de capteurs pour effectuer la collecte et l'intégration automatique des données d'observation et vont permettre de déterminer les spécificités que doivent avoir les capteurs. Cependant, nous pensons que les SunSPOT ne seront pas suffisants pour réaliser une intégration automatique en situation reelle. En effet, les capteurs SunSPOT sont intéressants pour amorcer une discussion sur le sujet mais restent malgré tout limités de part leur faible portée radio et leur faible mémoire de stockage. Il faudrait donc certainement envisager de transposer les résultats obtenus à d'autres technologies de capteurs (telles que celles mises au point au laboratoire de recherche Loemi de l'IGN (Martin, 2009)).

Cependant, les perspectives de cette étude sont multiples car l'utilisation de capteurs pour automatiser l'acquisition de données d'observation ne se limite pas au domaine de la pêche thonière et pourrait être transposée à d'autres domaines d'application.

6. Bibliographie

Amandè M., Ariz J., Chassot E., de Molina A. D., Gaertner D., Murua H., Pianet R., Ruiz J., Chavance P., Bycatch and discards of the European purse seine tuna fishery in the Atlantic ocean, Technical report, 2008. Estimation and characteristics for the 2003-2007 period.

Gaertner D., Pallares P., Efficiency of Tuna Purse-Seiners and Effective Effort, Technical report, 1998. (ESTHER). Sci. Rep. EU Programme N.98/061.

Gonzalez I., Ruiz J., Moreno G., Murua H., Artetxe I., AZTI discards sampling programme in the Spanish Purse seiner fleet in the western Indian Ocean (2003-2006), Technical report, 2007. IOCT-2007-WPTT-31.

Goujon M., Informations sur les captures accessoires des thoniers senneurs gérés par les armements français d'après les observations faites par les observateurs embarqués pendant les plans de protection des thonidés de l'Atlantique de 1997 à 2002., Technical report, 2004. International Commission for the Conservation of Atlantic Tunas, 56(2), 414-431.

Martin O., *SANY - an open service architecture for sensor networks*, vol. Chap 7.13 : IGN Geocubes, 2009. ISBN : 978-3-00-028571-4, Laboratoire de recherche Loemi de l'IGN : http ://recherche.ign.fr/LOEMI.

Romanov E., « Bycatch in the tuna purse-seine fisheries of the western Indian Ocean. », *Fish. Bull.*, vol. 100, n° 1, p. 90-105, 2002.

Romanov E., Bycatch and discards in the Soviet purse seine tuna fisheries on FAD associated schools in the north equatorial area of the Western Indian Ocean., Technical report, 2008. West. Ind. Oce. J. Mar. Sci. 7, 163-174.

Stretta J., de Molina A. D., Ariz J., Domalain G., Santana J., Les espèces associées aux pêches thonières tropicales., Technical report, 1997. ICCAT, 46(4), 250-254.

SunMicroSystems, « The SunSPOT project. http ://www.sunspotworld.com », n.d.

Zeller D., Pauly D., « Good news, bad news : global fisheries discards are declining, but so are total catches. », *Fish and Fisheries*, vol. 6, p. 156-159, 2005.

Integration of image processing methods for fuel mapping

Eric Maillé, Laurent Borgniet, Corine Lampin-Maillet, Marielle Jappiot, Christophe Bouillon, Marlène Long-Fournel, Denis Morge, Mohamed Amine El Gacemi, Dorian Sorin

Cemagref, Unité de Recherche Ecosystèmes Méditerranéens et Risques
Département Gestion des Territoires,
CS40061, Le Tholonet
13182 Aix en Provence cedex 5

ABSTRACT. Fuel mapping is a key activity for forest fire risk management. It is based on remote sensing images processing *methods*. These *methods* are versatile and validated in some particular contexts. They are usually implemented in one specific software environment. We propose a distributed solution for sharing and integration of image processing methods developed in their own computer environment and validated in some particular contexts, using specific data, to respond to specific needs. Its architecture includes a knowledge database of methods and resources, and an expert system for methods selection in relation to the user needs specification. Selected methods can then be organised into *demarches*. An executive engine is designed to execute the different methods of the *demarche* in their respective computer environment, through mediating wrappers. A research prototype called “Fuel Mapping Methods Integration Platform” (FMMIP) was developed.

RESUME. La cartographie du combustible est une activité clé pour la gestion du risque d’incendie de forêt. Elle se base sur des méthodes de traitement d’images télé-acquises. Ces méthodes sont variées, validées dans leur contexte de mise en œuvre, et généralement implémentées dans des environnements logiciels spécifiques. Nous proposons une solution d'intégration distribuée permettant le partage et la réutilisation de *méthodes* de traitement d'images pour la cartographie du combustible, développées dans des environnements informatiques en utilisant des données hétérogènes afin de répondre à des besoins spécifiques dans des contextes différents. Son architecture s'articule autour d'une base de connaissance et d'un système expert permettant d'évaluer la capacité de chacune des méthodes à répondre aux besoins spécifiés par l'utilisateur. Les méthodes sélectionnées sont alors organisées dans des *démarches* de traitements d'images, exécutables. Un moteur d'exécution permet l'exécution séquentielle de chacune des méthodes, dans leur environnement informatique respectif, au travers d'adaptateurs logiciels de médiation. Un prototype de recherche, appelé "Fuel Mapping Methods Integration Platform" (FMMIP) a été développé.

KEYWORDS: fuel mapping, remote sensing image processing, image processing integration, decision support systems, forest fire risks.

MOTS-CLES : cartographie du combustible, traitement d’images téléacquises, intégration de traitement d’image, aide à la décision, risque d’incendie de forêt.

1. Introduction

Forest fire risk management is one of the major concerns of Mediterranean local territories land planning (Moulignier, 2007). Land management decision makers require risk maps and risk models based on fuels maps. Fuels are vegetative covers, classed in different *types* in relation to their combustibility (Jounet, 2008).

In order to produce risk maps, fuel types have to be mapped using remote sensing images. At the European scale, both fuel typologies and the image processing methods used to map them are very different depending on the context, in particular the ecosystem type, as well as the available data and available computer resources to process the images.

In the context of the FIREPARADOX European research project, aimed at proposing a generic forest fire risk mapping method valid all over Europe, different fuel mapping methods were proposed by the different partners, adapted to particular contexts and using specific images available for their zone of interest. Moreover, most of the methods don't lead to a final fuel type map, but to some spatial variables useful to assess the combustibility of the vegetative cover: cover ratio, vegetation height, biomass, etc. As a result, it was not possible to propose a unique method valid all over Europe to map the whole diversity of fuels.

So we proposed an integration solution that aims to articulate different fuel mapping methods in a global processing *demarche*, as well adapted as possible to the user working context (Borgniet, 2009). It is a distributed solution, where methods are assessed in relation to the user specified context, and then can be associated and sequentially executed in their respective computer environment. The solution was developed as a research prototype called "Fuel Mapping Methods Integration Platform" (FMMIP).

In section 2 of this paper, we briefly describe different images processing issues and methods for fuel mapping. In section 3, we present the conceptual basis of systems integration on which we will specify a tool for image processing methods integration. Section 3 describes a general architecture of the specified fuel mapping method integration framework and its different components. Finally, we present the implementation of the FMMIP, and an example of the use of the developed tool.

2. Images processing methods for fuel mapping

Complexity of fuel mapping by using remote sensing images is related to the complexity of objects to be detected. So image processing methods are designed to try to solve the different levels of complexity.

2.1. Vegetation combustibility and fuel types

Spatial patches of fuel types are complex and highly heterogeneous spatial entities. We can classify the complexity of fuel types, in relation to their remote sensing-based mapping problem, in four sorts (Borgniet 2009):

- ***Purely spectral complexity*** of fuel types. Two different fuel types might have very close spectral signatures. This is in particular due to biomass obscured under the canopy and whose structure determines different fuel types. On the other hand, two different spectral signatures might correspond to two very close fuel types. This is due to the fact that combustibility of fuel types is mainly determined by vegetation structure, while spectral response depends on many other factors, like the soil type for discontinuous fuel types.
- ***Spatial heterogeneity*** of the spectral signal, for one given elementary object of interest characterising a fuel type. One elementary "object" of interest (for example one tree or even one homogeneous and continuous cover of trees) is represented by a set of pixels with very different spectral signal. Most of the pixels are then mixed, their spectral values being the average of several spectral responses. This heterogeneity is known as the **textural characters** of the patches.
- ***Spatial complexity of fuel types themselves***. Fuel types have spatial horizontal structures that determine their fuel characteristics. This structure represents the spatial organisation of the smallest elements of interest (trees, shrubs, etc.) of the fuel types. This heterogeneity is known as the **structural characters** of the patches.
- ***The vertical complexity of the fuel types.*** Fuel characteristics of a fuel type are highly determined by the vegetation stand "structure", i.e. the description of the grass, bush and trees strata. Simple remote sensing methods can only "view" the vegetation cover, i.e. highest stratum. Advanced remote sensing methods and tools have to be used in order to map stands vegetation structure.

The different methods studied by the different partners of the project permit to solve or to place elements in order to contribute to solve one or more components of the complexity of fuel mapping.

2.2. Notion of "methods"

Methods are defined as ***series* of several image *atomic processing***. The series is not necessarily strictly sequential, but might also be parallel in some cases.

Methods are usually developed in one given computer environment, in particular one given image (or geo-data) processing software[1]. However, methods are not necessarily computer implemented (automated). They often require the user to follow an interactive *demarche* of successive processing (in that case, we will talk about *literal* methods). Some specific processing is implemented on specialised commercial software, particularly in the case of "object based" processing and LIDAR processing.

Automation of such sequences might be implemented in macro-languages of the software. However, as many parameters are required, a great interactivity with the user is necessary. For example, for supervised processing like supervised classifications, patterns (areas of interest) have to be provided to the procedure. Such interactivity might require complex capabilities that software simple macro-language might not provide. In particular, such interactivity requires elaborated graphical user interface that might be difficult to develop using the commercial software tools.

Methods were developed by partners using particular input data, generally satellite images, but also different geo-data. Such data might not be available everywhere, but other untested images or geo-data might also be used. Therefore methods are often *data dependent*, but some data might be substituted by other more available data.

Methods might not end in a complete fuel map, but allow the assessment of some important attribute of vegetative cover involved in fuel typology (percent of cover, vertical structure, etc.). Methods might have to be linked in order to reach a complete fuel map. Moreover, methods are often applied to fuel of a particular land cover (for example, continuous forest land, bushlands, etc.): so several methods might be required to map the fuel for the whole area.

Fuel mapping methods always depend on the geographical context. However, methods are also closely related to fuel typology and have been developed based on particular vegetation typologies. It is very difficult to propose only one method to map fuel, valid all over the world. Partners of the FIREPARADOX project have therefore proposed different methods adapted and validated to some particular areas.

We can distinguish four groups of methods: spectral methods, textural methods, object oriented methods, and 3D methods.

2.2.1 - Spectral methods (or "pixel based methods")

Spectral methods are convenient to solve purely spectral complexity. All image processing methods use the spectral values of the pixels, but "purely" spectral (or

[1] Some particular "auxiliary processing", like image format conversion, might require different particular software.

pixel based) methods use *only* this information. Some processing might also use spectral information of the neighbourhood of the pixel (filter, for example), but in that latest case, information about the spatial distribution of this neighbouring information is ignored.

Purely spectral methods are usually based on multi-spectral classification processing and/or on spectral indexes calculation, in particular vegetation indexes like the Normalized Difference Vegetation Index (NDVI), the Ratio Vegetation Index (RVI), Soil Adjusted Vegetation Index (SAVI), etc. Note that such processes are "sensor dependent", and might give very different results in relation to the spectral range of each bands of the sensor used to acquire the image.

These process capabilities are commonly implemented on commercial image processing software.

2.2.2. Textural methods

Textural methods are convenient to solve textural complexity of fuel types. This is a key element of *continuous* or *dense discontinuous* fuel type mapping that have a regular (not structured) heterogeneity. Such fuel types are very common in the Mediterranean area: Mediterranean forest, scrubland, dense shrub/bush lands, might be discriminated using textural methods.

Advanced textural processing are complex algorithms not always incorporated in standard commercial image processing software. Some "*add ons*" often exist including particular algorithms that are not always adapted to the specific problem of fuel type mapping.

In the FIREPARADOX project, a specific software, called the GLCM Tool (Grey Level Co-occurrence Matrix Tool), was develop and implemented. The algorithm is based on the co-occurrence analysis algorithm by multi-dimensional classification of Haralick indexes (Capel 08).

2.2.3. Object oriented methods

Objects oriented[2] methods are convenient to solve horizontal spatial structure complexity. They aim to detect geographical objects in relation to some of their spectral, textural, or geometrical attributes (shape, size, etc.). These methods are more particularly dedicated to discontinuous horizontally structured fuel types

[2] Note the term of "*object oriented method"* refers to a specific class of image processing and to *geographical* objects. It has no link with computer science *object oriented* conceptual modelling or development methods.

detection. In such fuel types, geographical fuel objects (for example shrubs, isolated trees or coppice, opened grass patches, etc.) are organised into a particular spatial structure characterizing the fuel type. Objects have to be first detected before analysing the structure using particular spatial analysis processes.

Geographical objects extraction is a quite empirical approach, based on the definition by the expert user of several parameter value intervals. A set of rules is defined that permit to identify particular classes of geographical objects. This learning phase is highly interactive.

Commercial object oriented image processing software allow the storage of rule sets in order to be able to reuse them in other contexts (other images, for example). This procedure permits to automate object oriented methods. However, applying the same set of rules to a different context has to be done very carefully.

A second phase of object oriented methods is the structure spatial analysis, in order to delimit iso-structure patches characterising fuel types. Common raster or vector spatial analysis processes might be used, like density calculation of each object class, buffer based envelop drawing, inter-object distance calculation, overlap rate calculation, etc.

2.2.4. 3D methods

3D methods are convenient to solve vegetation vertical complexity. At least three kinds of 3D methods were developed.

Some particular spectral methods aim to assess vegetation density under the top vegetation layer. A correlation between some vegetation indexes values (RVI) and some vertical structure characters of the vegetation was indentified and validated in some particular fuel types. Such methods are easy to automate, but should be calibrated for different fuel typologies.

Photogrammetric methods aim to assess the top layer vegetation height by calculating the difference between a surface numeric model and a terrain numeric model. Vegetation height is a key factor in order to calculate the fuel biomass. However, these methods require many user manipulations and are very difficult to automate.

LIDAR based methods are highly technological and costly methods that might provide a precise description of vertical fuel structure. Main limit of such methods is data availability, cost, and processing complexity. All processing requires specialised software, and automation can only be possible in specific computer software environments.

2.2.5. Mix methods

We put the emphasis on the fact that most of the methods studied by the different partners of the project are mixed. For example, the Cemagref RVI method articulates a purely index based spectral approach (RVI calculation) and an object oriented approach (ENVI Feature Extraction module), in order to assess the vertical cover rate (vertical complexity of fuel types). The Cemagref GLCM co-occurrence matrix-based method might also be considered as both a textural and a structural (horizontal) method.

2.3. Toward an integrative processing approach

Methods are defined as geo-data processing using one particular geo-data type, developed in one particular computer resource environment (one software type). Methods specified by the FIREPARADOX partners have their own application domain: particular geographical or ecological context, particular fuel typology to be detected, etc. Some methods are specialised in particular groups of ecosystems (scrublands, etc.). Finally, they usually contribute to map important factors of vegetation combustibility (horizontal density, vertical structure, etc.).

So it appears that it is not possible to propose one unique method able to produce a fuel map valid in any context with the same parameters. Proposed methods are context dependent and might be complementary in order to solve the global problem of fuel mapping in a given geographical and ecological context.

This lead us we choose an open knowledge based system, opposed to a closed processing solution. The system aims to help the user to build a global successive processing approach that we call a "*demarche*", in order to better respond to his needs. We present some aspects of processes integration in the next section 3, before presenting the architecture of an integration framework in section 4.

3. Integrative approaches of image processing

In the field of remote-sensed image processing, a process is a sequence of operations aimed at extracting semantic information from a raw multi-dimensional raster data set (image): so, it is considered as an *interpretation function* relating some pixel radiometric values of the image to a set of objects *classes*. As the function is often complex, composed of many elements, we will talk about processing *models.*

3.1. Coupling processes and models integration

Image processing methods are executive models implemented on specific image processing software systems. In order to apply sequences of different methods on a data set, these models have to be coupled or integrated.

3.1.1 Models coupling typology

Mandl (Mandl 96) proposes a typology of model coupling (it is specialised in coupling GIS and multi-agents based models, but it can be applied to any kind of model coupling), in relation to its architecture:

- *"weak coupling":* models are implemented in two independent systems exchanging only data. Massive dynamic data exchanges often limit the efficiency of the system. Because no mediating system exists, coupling requires an access to the internal structure of the model.

- *"tight coupling":* models are implemented in a same computer program. This kind of coupling solves the problem of massive data exchanges, but it does not permit re-use of existing software components. It requires a lot of development and implementation work.

- *« direct cooperative coupling »:* models are independent, but communicate through a client/server link. The user operates the resulting system through the user interface of the client system. This coupling requires a good compatibility between models and data of the coupled systems. It permits to overtake the limits of tight coupling, allows re-use of existing software, and keeps the whole functional potential of existing systems. It facilitates dynamic exchanges thanks to the client/server link. It also requires considerable development work for the client's user interface in order to be able to drive the server system.

- *Indirect cooperative coupling* is also based on a client/server link with a mediating system in charge of data interoperability that is endowed with a user interface. Indirect cooperative coupling is rather costly for development of the mediating system, but it solves the direct cooperative coupling limitation of data interoperability.

3.1.2 Integration levels

The concept of integration level might refer to the tightness of the link between models, (for example, the number of software and hardware layers to be crossed for information exchanges), or the number of software components that can be

considered as autonomous systems in the final solution. As tight coupling is supposed to be aimed at “melting” the different systems in a unique one, it is the only one that can be qualified as real “integration”. Direct cooperative coupling is more integrative than indirect cooperative coupling because of the lower number of software systems involved.

Data exchange efficiency is the second criterion to assess the integration level of a given software architecture. Massive data exchanges of weak coupling are not very efficient as opposed to message exchanges of client/server based cooperative architectures that can be considered more integrated than weak coupling.

3.1.3 Formal specification of integration

Models integration also refers to the usual modelling approach, in which a system is decomposed into different sub-systems, each of them representing a specific model different from the others in terms of semantic and formalism : Duboz (Duboz 04) proposes the coupled DEVS formalism (Ziegler 99), stemming from discrete events simulation, in order to produce a formal representation of connections between outputs of a models and inputs of the others one. The coupled DEVS formalism represents the scale transfer to a upper level (integrated level) by coupling atomics DEVS elements of a lower scale level (fig 1).

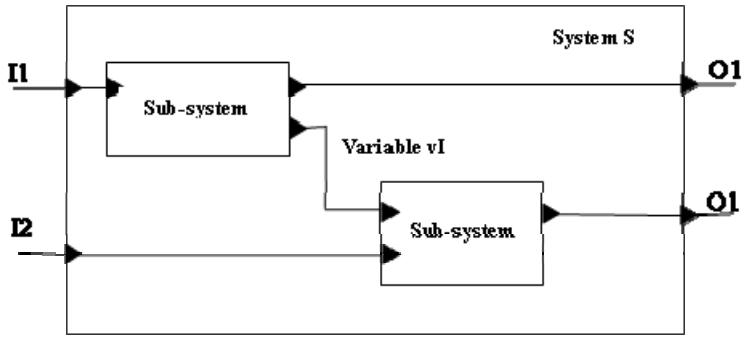


fig 1. *Graphic representation of a formal coupling of models with the DEVS formalism*

The formal system DEVS allows proving the confidence of the integration, but does not specify its architecture. In that perspective, we consider that the coupling is

the formal specification of the integration. Two other aspects of the integration have to be solved: the semantic aspects, and the syntax aspects.

3.2. Semantic and syntax integration

Semantic integration is aimed at solving semantic heterogeneity between models to be integrated at a conceptual level. It specifies the semantic relationships between *concepts* handled by the models to be integrated. If semantics (i.e. list of concepts) handled by the models are different, integration will require the specification of models for models integration (Maillé 08): such models specify the relationship between concepts of the initial models.

Syntax integration is aimed at solving heterogeneity of representation terms of information handled by the models to be integrated. It permits models interoperability which allows proper functioning of the resulting model, without referring to its semantic consistence (Müller 08).

Syntax integration might be specified at different abstraction levels: organisational level (architecture), logical level (data models, communication protocols, etc.), physical levels (networks), etc.

3.3. Specification of an integrative image processing system

As any decision support tool, specifications of an integrative image processing system depend on its intended use. Fuel mapping is an occasional activity decision-makers practice for forest fire risk management planning. Processing time is usually not a limiting factor, while resources availability (data, software and knowledge) might be strict constraints. As a consequence, we specified an integrative architecture based on a mixed solution:

– A weak coupling system is in charge of data exchange, while the processing sequence is managed by a direct cooperative coupling, between distributed “nodes”. This solution permits to use some distributed resources, without being constrained by their storage location.

– Specifications of processing integration are implemented into processing “demarches”. Demarches are specified by the end user of the produced tool, who can store and share them in a shared knowledge database.

– In the application field of image processing for fuel mapping, we assume that all processing models handled concepts belonging to the same unique ontological field (Grüber 93). So image processing model integration does not require any semantic specifications.

In the following section 4, we describe the integrative image processing tools developed for fuel mapping activity called the Fuel Mapping Methods Integration Platform (FMMIP).

4. The Fuel Mapping Method Integration Platform (FMMIP)

The fuel mapping methods integration platform is an open knowledge based system, aimed at helping the user to build and operate a global *demarche* by articulating different *methods* in order to produce a fuel map adapted to his context and responding to his needs. Context parameters might concern geographical variables related to the user's working zone (location, geology, climate, etc.) or the user's available data and available computer resources, in particular commercial image processing or spatial analysis software. Needs concern the targeted result to get (targeted fuel typology, scale of the fuel map) and/or the previewed use of this product (global risk calculation, operational planning, etc.).

The "*demarche*" is to organise different "*methods*" into a processing framework, allowing the user to take into account his different constraints and specifications. Then, the global *demarche* is not unique because it has to be adapted to the use of the fuel map. A global demarche articulates different methods with other standard geo-data processing in relation to the different available resources (figure 2).

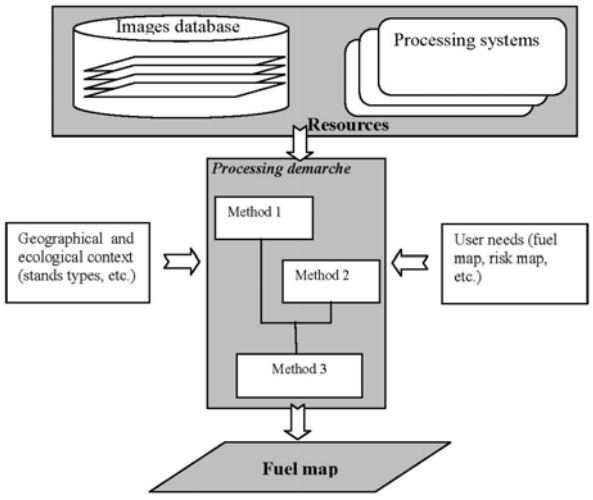


fig 2. *Processing "demarche" combining several processing "methods"*

In this purpose, the fuel mapping method integration platform manages and operates *resources*. Resources are either geo-data or geo-data processing systems. For example, implemented methods are considered as geo-data processing resources. Resources might be open access or limited access. Most of the geo-data used, in particular satellite images, are limited access resources because the user must have license rights to process them. Implemented methods depending upon commercial software are also limited access resources because of the required license to use the commercial software.

In order to access limited access resources, particular agreements will have to be passed between the platform user and the owner of the resource.

4.1. The FMMIP "nodes"

The fuel mapping methods integration framework is composed of a network of FMMIP "nodes". Nodes architecture is structured by a kernel surrounded by peripheral software modules, and linked to a knowledge database (figure 3). Software modules are image processing or GIS software and associated *methods* implemented in the macro-language of the given software.

The nodes kernel is composed of three main components: a driving Graphical User Interface, an *expert system* engine that helps the user to choose the best resources to use in relation to his needs, and an *executive engine* that can *operate* the resource, if possible. In particular, it can execute *methods* by operating their implementing software. To do so, the executive engine accesses the software modules through wrappers (figure 3).

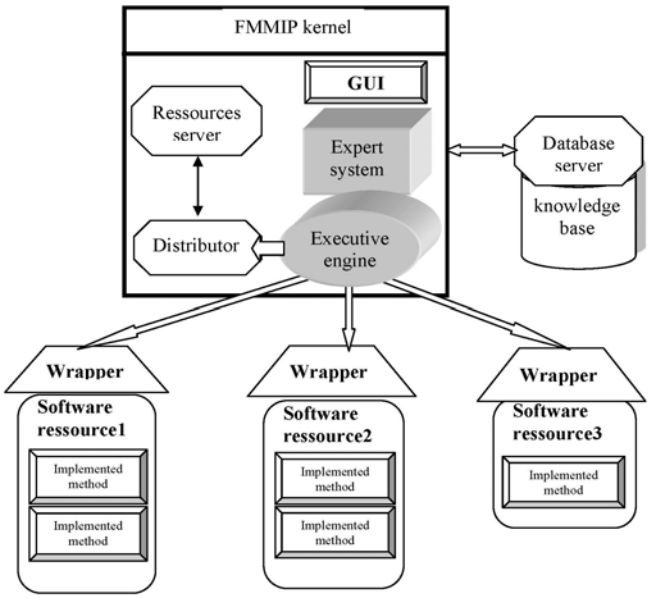


fig 3. *A FMMIP node architecture*

The *knowledge database* gathers information about available resources. It contains all information about the resources (location, accessibility, operability, etc.), but does not contain any resource i.e. data or data processing software. Most of the kernel components might get information from the database server.

Moreover, each node kernel might be endowed with three components dedicated to the system distribution: a database server, a process server and a distributor. We develop the role of these different components in the next part of this section.

4.2. Functioning scheme of a FMMIP node

Figure 4 presents the components used in a standard use case of the FMMIP.

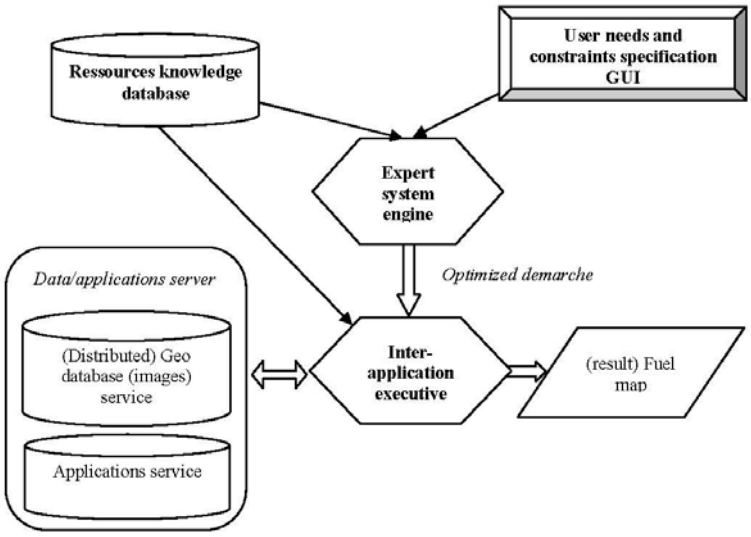


fig 4. *Functioning scheme of a FMMIP Node*

A standard use case is the following:

1) The user specifies his requirements, i.e. his needs (targeted typology, fuel mapping parameter estimator, etc.), his context (geographical and ecological), and his locally available resources (data and software), with the GUI.

2) From these data, the *expert system* queries the resource knowledge database in order to assess the available resources regarding the requirements specification, and then proposes a set of *demarches* and *methods* adapted to respond to the requirements.

3) The user chooses either a whole *demarche* or some proposed *methods* in order to build his particular *demarche*.

4) The demarche is executed, by invoking process

5) The output is displayed to the user.

6) If the user is not satisfied with the result he can either modify parameters of the process *or* specify à new demarche.

7) If the user is satisfied with the *demarche*, information provided by the user can be stored in the knowledge database in order to complete it and share it. In particular, context information is used to enlarge methods' usability in the knowledge database.

The fuel mapping tool accepts as input geo-data, mainly satellite images, and provides as output fuel maps. Data format depends on the processing software accepted formats. The tool provides either a classified image (raster) or a GIS vector layer endowed with semantic attributes related to fuel characters (fuel type, combustibility index, etc.). Produced data format depend on the processing software used.

All resources used by a node, data, software of knowledge, might be located on an other FMMIP node. In the next paragraph, we describe the distributed architecture of the system.

4.3. The system distribution

The fuel mapping methods integration framework is structured in a “cloud computing” service oriented architecture (SOA, Nickul 05), where nodes are resources of a “private cloud” (Catteddu & al., 09). Nodes might communicate through a wide area network (WAN) like the Internet network. However, it is not based on web’s standards and protocols (http, applets/servlets systems, etc.), only nodes being addressed through their URL.

Any node can be client, server or both. Moreover, a FMMIP node might be a *resource* server, so that it offers processing services or data providing services, and/or a *knowledge server* so that it can offer read access to its node database. Distribution is ensured by the "Distributor" component of each node, that takes charge of the "*client*" role, and the "Ressources server" that plays the role of the *server*.

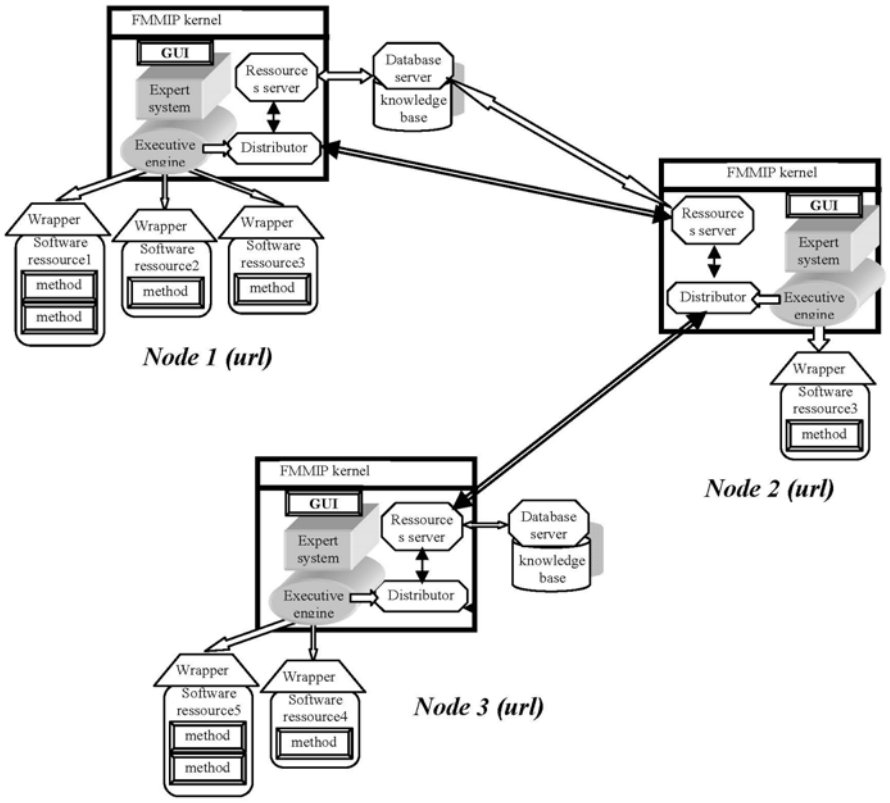


fig 5. *The FMMIP framework distribution*

When the executive engine of *node 1* has to execute a particular process on some given data, it invokes the "Distributor", which queries the local knowledge data base to check if all data and processing resources are available locally. If not, it finds the URL of a FMMIP *node 2* where data or processing resource might be found, so that the "Distributor" can invoke the remote "Ressource server" of *node 2*. The "Ressource server" checks into the database if the resource is available on *node 2*. Note that the database it consults is the shared database of *node 1*. If the resources are available locally on *node 2* and the required resources are data, the resources server temporarily uploads the data back to *node 1*. If the required resources are processing resources, the Resources server temporarily downloads data from *node 1* to *node 2*. Then it asks the Executive engine to process these data through the convenient wrapper. Finally, it uploads the result data back to node 1.

The required resource might also not be locally available on *node 2*. In that case, the *node 2*'s Resource server should find in the database the URL of a remote *node 3*. It will the invoke *node 2* Distributor, so that the process can be repeated recursively.

4.4. Agreements required between node administrators

The FMMIP network allows sharing processing resources as well as data. Although data can not be kept on the remote nodes, agreements between the different node administrators have to be passed so that resource access rights are respected. The FMMIP network cannot be an open access network, but operate within a limited community involved in a given topic, like the Fireparadox community in the fuel mapping topic.

5. The prototype implementation and validation

A prototype of the FMMIP platform was developed in the context of the Fireparadox project. It permits to build and share multi-environment processing *demarches* based on specific *methods* developed by different partners of the project.

5.1. Technical specifications

The platform is developed around a kernel and wrappers in the JAVA language, by using the respective software macro-languages (Gacemi, 2009). For example, in order to communicate with the image processing software ITT ENVI©, used to operate some "object oriented" methods, the IDL language is used. Many image processing software also use script-like macro-language (ESRI ArcInfo© AML, ERDAS Imagine© batch, etc.). The GIS software ESRI ArcGIS language is VBA© (Visual Basic for Application) or Pyton©. A specialised image processing tool kit, called the "fuel mapping resources tool kit" was also developed in C++ language, for open access standard image processing (Sorin, 2009). This software uses standard system script macro-language.

The node databases are managed by the shareware database server POSTGRES. So they might not be "*local*", but can also be remote. The node can use, for example, a centralised shared knowledge database. However, the database related to a node is unique for each node, during a FMMIP session.

5.2. Example of use

An example of processing *demarche* is proposed. A *demarche* is composed of "*meso-processes*", i.e. "linear" sequences of processing. The proposed *demarche* will try to differentiate some fuel types within the fuel zone extracted by a multi-spectral classification. To do so, it uses a particular “method”called the GCLM

textural analysis, specified by the FireParadox project (Capel, 2008), specifically on the extracted fuel zones.

This *demarche* has four meso-processes (fig.).

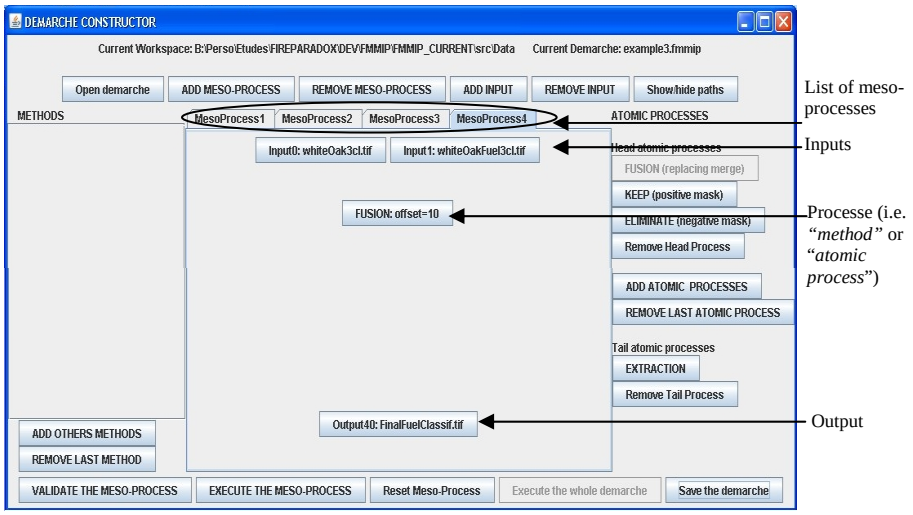


fig. 6. *A demarche with four mesoprocesses*

It is designed to process aerial images on a local French sub-Mediterranean context (fig.).

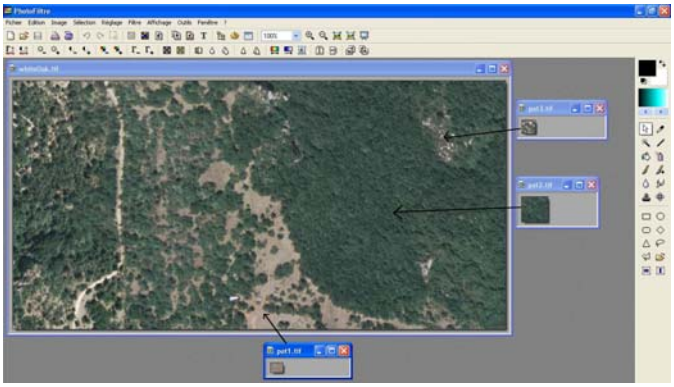


fig. 7. *« Patterns » extraction from the initial image*

Meso-process 1 operates a ERDAS Imagine© supervised multispectral classification in three classes, using learning patterns (method 1, fig. 7, ①). Then it extracts the class number 2 (fig. 7, ②), corresponding to fuel areas.

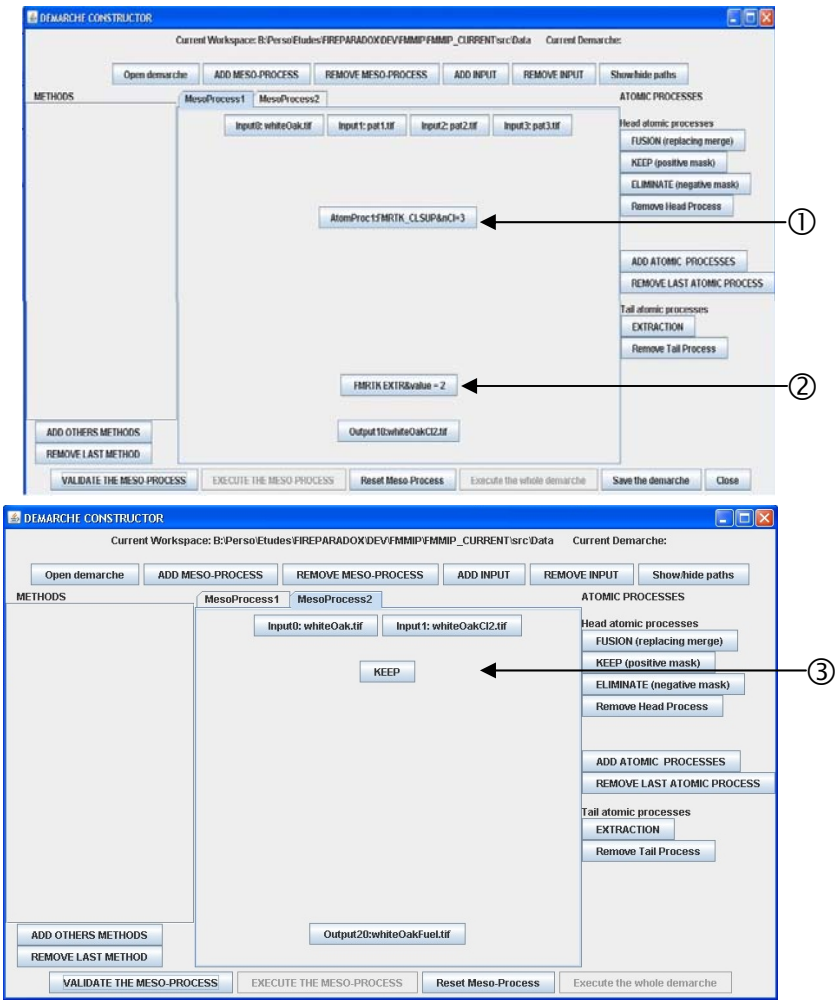


fig 8. *Meso-processes 1 and 2 of demarche*

Meso process 2 extracts from the initial image pixels corresponding to class 2. (fig. 7, ③). It is implemented in a different meso-process because it requires two inputs. Successive outputs of these processes are shown in fig. 8.

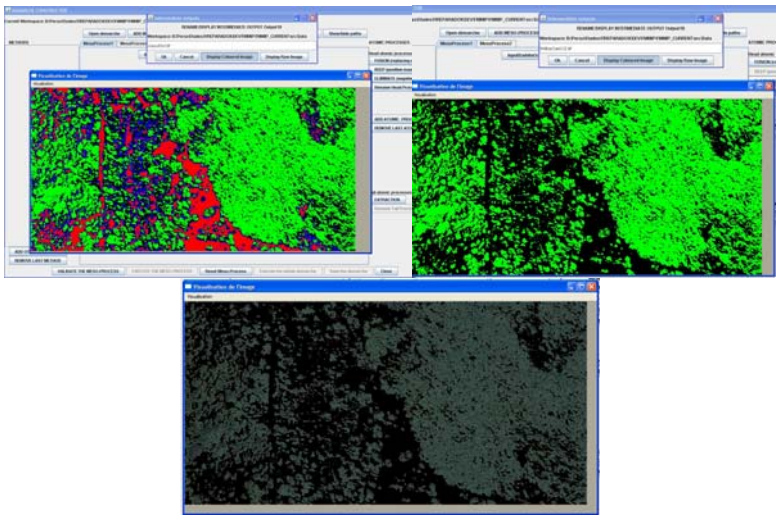


fig 9. *Successive outputs of the different process (form left to right, supervised classification, class 2 extraction, pixel extraction from initial image.*

The *meso-process 3* will only execute the GLCM method (fig. 9, ④), on the only extracted fuel zone (fig 10)

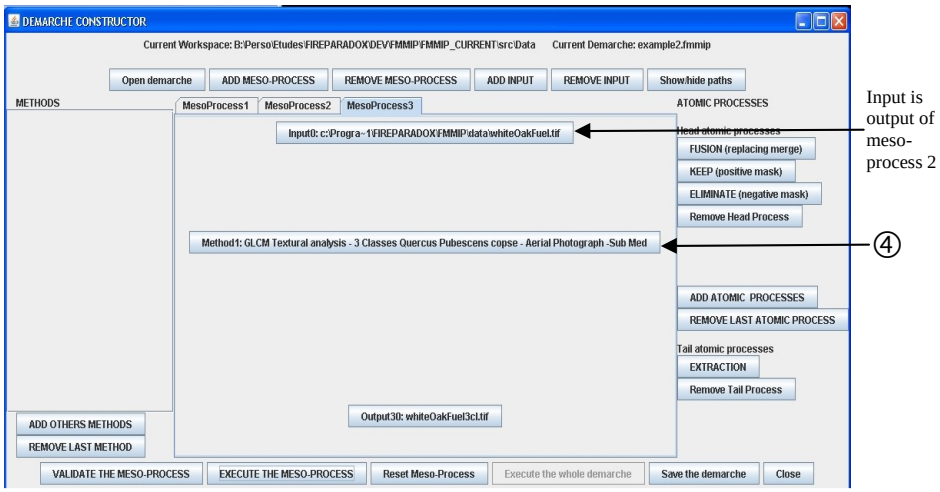


fig 10. *Meso-process 3 of the demarche*

The GLCM will class the fuel areas regarding the different patterns provided.

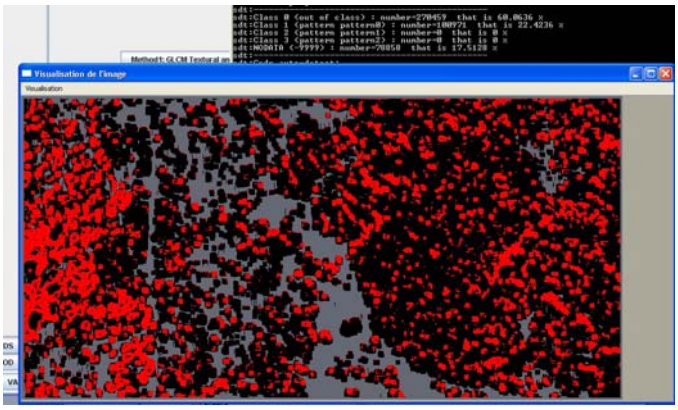


fig 11. *Output of the GLCM process*

Finally, the *mesoProcess 4* is designed to make a global fuel map, including fuel areas and non fuel areas, by merging the first classification executed by the *meso-process 1*, and the fuel type classification just obtained with the *meso-Process 3*. Result is displayed in figure 12.

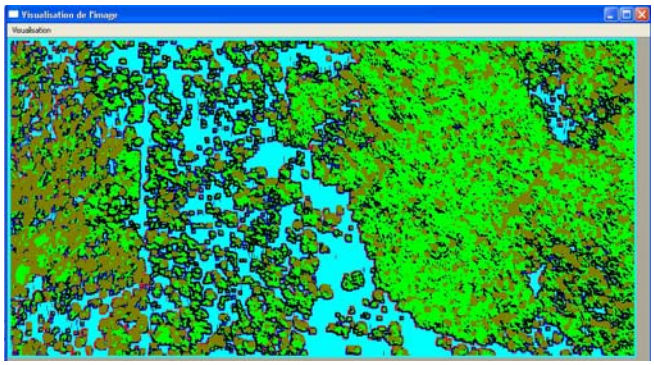


fig 12. *Final output of the demarche*

The classification has now 4 classes:

Class 1 (Red): these are rocky pixels

Class 3 (Light blue): neck soil (defined by the first classification

Class 10 (Light Green): fuel not classed by the GLCM analysis

Class 11 (dark green): fuel classed 1 by the GLCM analysis

6. Conclusion

We propose a fuel mapping system that aims to take into account the complexity of the fuel geographical object definition and typology, the wide diversity of geographical and ecological contexts in which fuel might be mapped , and finally the diversity of resources potential users can have access to. To do so, the system is organised around the concept of *method* and *demarche* that is processing sequence, based on specific data type, adapted to a particular context, and aiming to detect a specific fuel typology (or a particular sub-set of fuel types of a universal fuel typology).

The prototype being implemented is composed of several components: a system kernel, that includes a knowledge database and an expert system to choose the best adapted methods in relation to requirements, an open access resource toolkit that provide common processing algorithms, and different methods implemented on their particular software environment. All these elements have to be able to communicate, through an adapted distributed architecture. Potential distribution of this architecture makes possible a physically distributed system, based on a shared method knowledge database and also shared images database.

7. References

Borgniet, L., Maillé, E., Long, M., Capel, A-C., Bouillon, C., Morge, D., Ganteaume, A., Lampin-Maillet, C., Jappiot, M., Curt, T., Machrouh, A., Sesbou, A., Mantzavelas, A., Apostolopoulou,I., Partozis, T., Gitas, I., Marell A., Cassagne N., Pimont F., Rigolot E., Morsdorf F.,Koetz B., Allgower B., 2009, Development of an easy to use tool to recognize and mapfuel models: Deliverable 5.1-6 of the Integrated project “Fire Paradox”, Project no. FP6-018505, European Commission, 107 p.

Capel A.C., 2008, Gray Level Coocurence Matrix Tool (GLCMTool), Manuel d'utilisateur, *Cemagref*, UR Ecosystèmes Méditerranéens et Risques, Aix-en-Provence, F

Catteddu D., Hogben G. (eds.), 2009, “Cloud computing: Benefits, risks and recommendations for information security”, European Network and Information Security Agency (ENISA), http://www.enisa.europa.eu/act/rm/files/deliverables/cloud-computing-risk-assessment

Duboz R., 2004, « Intégration de modèles hétérogènes pour la modèlisation et la simulation de systèmes complexes, Application à la modélisation multi-échelle en écologie marine » Thèse de Doctorat, 2004, Laboratoire d’Informatique du Littoral, Université du Littoral Côte d’Opale

Gacemi M. A., 2009, "Spécification et développement d'une plateforme d'intégration de méthodes de traitement d'image pour la cartographie du combustible", rapport de stage, Ecole Nationale des Sciences Geographiques, *Cemagref*, UR Ecosystèmes Méditerranéens et Risques, Aix-en-Provence.

Grüber T. R., 1993, Towards Principles for the Design of Ontologies Used for Knowledge Sharing, In. Formal Ontology in Conceptual Analysis and Knowledge Representation, Kluwer Academic Publisher.

Journet S., 2008, Caractérisation et cartographie du combustible dans les interfaces habitat/forêt., Mémoire de Master II EGEPM Évaluation et Gestion de l'Environnement et des paysages de montagne, *Cemagref* Aix-en-Provence, 120p.

Lampin-Mailllet C. 2009. Caractérisation de la relation spatiale entre organisation spatiale d'un territoire et risque d'incendie : Le cas des interfaces habitat-forêts du sud de la France Thèse de doctorat en Géographie Analyse spatiale de l'université de Provence,. 321 p.

Maillé E., 2008, "Intégration conceptuelle et opérationnelle de modèles spatio-dynamiques, Application à la dynamique du risqué d'incendie de forêt", Thèse de doctorat, Laboratoire des Sciences de l'Information et des Systèmes, UMR CNRS 6168, Université Paul Cézanne Aix-Marseille III, F., *Cemagref*, UR Ecosystèmes Méditerranéens et Risques, Aix-en-Provence, F

Mandl P., 1996, Fuzzy-System-Umgebungen als regelgesteuerte Simulationsmaschinen für Geographische Informationssysteme, http://www.uni-klu.ac.at/groups/geo/gismosim/paper/mandl/mandl.htm

Moulignier A. (Dir), 2007, "Prévention des incendies de forêt", Dossier de presse, Direction Générale de la Forêt et des Affaires Rurales, Ministère de l'Agriculture et de la Pêche, Paris, F.

Müller J.P., 2008, MIMOSA user's manual, CIRAD-ES-GREEN, Montpellier, F

Nickul D., Reitman L., Ward J., Wilber J., 2005, "Service Oriented Architecture (SOA) and Specialized Messaging Patterns", Technical White Paper, Adobe.

Sorin D., 2009, "Spécification et développement.d'une solution de traitements d'images satellites pour la cartographie du combustible, dans le cadre du projet FIREPARADOX", rapport de stage, Institut Universitaire de Technologie d'Arles, *Cemagref*, UR Ecosystèmes Méditerranéens et Risques, Aix-en-Provence.

Zeigler B, 1999, "Theory of modelling and Simulation", Ed. John Wiley & Sons, 2nd Edition

Une approche innovante de modélisation du risque d'incendie de forêt

Fondée sur la cartographie des interfaces habitat-forêt, nouvelle clé de lecture du territoire

Lampin-Maillet C*, Jappiot M*, Ferrier J.P**

** Cemagref, UR EMAX,*
3275 route de Cézanne CS40061, 13182 Aix en Provence cedex 5, France
corinne.lampin@cemagref.fr

*** Université d'Aix-Marseille I, Professeur émérite,*
Aix-En-Provence, France

RÉSUMÉ. *Une méthode de cartographie des interfaces habitat-forêt est développée dans le contexte du risque d'incendie de forêt. La cartographie des interfaces habitat-forêt sur le territoire qui en résulte permet alors une nouvelle compartimentation du territoire : types de territoire interfacés et types situés en dehors des interfaces. En mettant en relation la distribution spatiale de ces types de territoire avec l'historique des incendies (départs de feu et surfaces brûlées), certains types d'espaces révèlent de haut niveau de risque d'incendie avec une forte densité de départs de feu, d'incendie et de taux de surfaces brûlées. Une modélisation du risque d'incendie est proposée de façon globale.*

ABSTRACT. *A method to characterize and to map wildland-urban interfaces (WUI) is proposed in the context of wildfire risk. The WUI mapping on the territory allows new spatial configurations: inside WUI and outside WUI. Establishing relationships between WUI distribution and forest fire history (departure of fires and burned areas) types of territory appear with high levels of risk: high fire ignition density values and high wildfire density and high burned area ratio. A model of a total fire risk index has also been developed.*

MOTS-CLÉS : *interface habitat-forêt, risque d'incendie, habitat, indice d'agrégation, densité d'éclosion, densité d'incendie, taux de surfaces brûlées, indice global de risque.*

KEYWORDS: *wildland-urban interface,wild fire risk, housing, aggregation index, fire ignition density, wildfire density, burned area ratio, total index of fire risk.*

1. Introduction

Les incendies de forêt affectent de grandes surfaces et causent d'importants dommages qui peuvent avoir de lourdes conséquences écologiques, sociales et économiques. Plus de 50 000 feux brûlent environ 500 000 hectares de végétation chaque année dans les pays du bassin méditerranéen européen (JRC, 2006; Lampin-Maillet, 2008). Les interfaces habitat-forêt sont directement concernées par ces incendies : 90% des départs de feux sont liés à l'activité humaine en Europe Méditerranéenne, et chaque année de nombreux morts sont à déplorer à cause de ces incendies de forêt, notamment parmi les habitants des interfaces habitat-forêt. Dans le contexte d'une forte pression d'urbanisation et d'une accumulation de biomasse combustible, les interfaces habitat-forêt représentent une véritable préoccupation pour la gestion du risque d'incendie (Davis, 1990; Velez, 1997; Cohen, 2000), particulièrement au regard des deux composantes du risque : l'aléa en termes de départs de feu causés par les activités humaines, et la vulnérabilité, en termes de surfaces brûlées menaçant les zones habitées et aussi de dégâts sur les habitations (Hardy, 2005; Jappiot *et al*, 2009).

Malgré les fortes préoccupations que causent les interfaces habitat-forêt, notamment en matière de gestion du territoire et de gestion de l'incendie, les données sur leur localisation sont imprécises et celles sur leur extension sont rares. Comme le soulignent Theobald et Romme (2007) ainsi que Dumas *et al.* (2008) des cartographies plus détaillées d'interfaces habitat-forêt permettraient d'utiliser les cartes produites à des fins d'activité de gestion et de prévention mais aussi de prospective en matière de développement futur. Le développement d'une méthode efficace pour cartographier précisément les interfaces habitat-forêt serait nécessaire pour la gestion du risque d'incendie.

Le risque d'incendie de forêt est une réalité en région méditerranéenne française, son intégration dans la gestion et l'aménagement du territoire est devenue incontournable. Cette intégration du risque doit s'appuyer sur des actions conjointes : (1) de gestion et de protection des massifs forestiers à travers leur aménagement ; (2) de planification et de réglementation pour maîtriser l'urbanisation avec la prise en compte du risque d'incendie dans l'aménagement des zones urbaines ; et (3) de maîtrise de l'utilisation du foncier localement pour la protection ou la mise en valeur d'espaces menacés par un risque d'incendie de forêt. Mais pour une intégration réussie, l'évaluation spatiale du risque d'incendie et sa cartographie sont une des composantes nécessaires. Et cette évaluation du risque doit s'orienter vers une approche globale.

L'objectif de l'article est donc de présenter une approche d'évaluation du risque d'incendie innovante. Celle-ci s'appuie sur une cartographie des interfaces habitat-forêt qu'il a fallu mettre au point. Puis à partir d'une analyse spatiale et statistique du territoire, des indicateurs de risque ont été définis pour construire un modèle de risque global.

2. Méthodologie

2.1. *Site d'étude et données*

La *zone d'étude* (Fig.1) est située dans le sud-est de la France entre les métropoles d'Aix-en-Provence et de Marseille dans le département des Bouches-du-Rhône (43°23'57" N, 5°22'00" E). Elle s'étend sur 167 736 ha couvrant 59 communes : 60 % de la zone est occupée par des espaces forestiers, 20 % par des espaces urbains et 20 % par des espaces agricoles (OccsolSPOT 5, 2003). Cette zone connait un haut niveau d'urbanisation et de pression urbaine (420 hab/km^2). Les interfaces habitat-forêt y sont très communes. L'extension urbaine occupe peu à peu les anciennes terres agricoles désormais en jachère mais elle est également marquée aux limites, voire au cœur, des massifs forestiers.

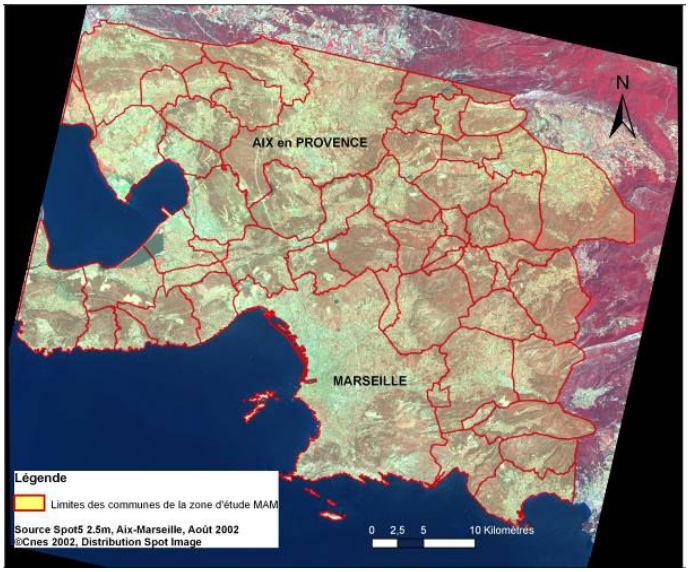


Figure 1. *Zone d'étude entre les métropoles Aix-en-Provence et Marseille*

Les *principales données* utilisées sont :

- La base de données géoréférencées de départs de feu, fournie par l'agence départementale de l'Office National des Forêts des Bouches-du-Rhône (ONF13), comptabilisant 565 points d'éclosion sur la zone d'étude située entre les métropoles d'Aix-en-Provence et de Marseille sur la période 1997 à 2007 ;

- Les surfaces incendiées dont les contours digitalisés sont issus d'une base de données géo référencées DDAF13/ONF13 concernent les feux de plus de 10 ha recensés de 1960 à 2007 et les sautes associées d'une surface généralement

inférieure à 1 ha. Dans l'étude, seuls les feux de 1990 à 2007 ont été pris en compte de façon à rendre acceptable l'hypothèse d'une évolution modérée de l'occupation du sol et de mettre en relation la surface brûlée avec le territoire existant ;

- La carte d'occupation du sol Occsol SPOT5 conçue pour décrire les paysages urbains et périurbains tout en intégrant les principales composantes naturelles extra-urbaines. La hiérarchisation des classes et le contenu thématique reprennent les principales nomenclatures Corine Land Cover. Elle est un dérivé du traitement et de la photo-interprétation des images SPOT5, 2.5 m couleur, (2002), assistée de l'utilisation de données exogènes nécessaires. Elle a été produite en 2004 à l'initiative du CNES, de l'ARPE PACA et de Spot Image ;

- D'autres bases de données relatives aux bâtis, au modèle numérique de terrain de l'IGN.

Le logiciel ArcGIS© Version 9.2 a été utilisé comme SIG et le logiciel STATGRAPHICS®Centurion comme outil des traitements statistiques.

2.2. Démarche de recherche

Une démarche de recherche a été construite en trois étapes. La *première étape* a consisté à définir précisément l'interface habitat-forêt dans le contexte du risque d'incendie et à développer une méthode de caractérisation de cette interface par une approche d'analyse spatiale. Cette analyse a conduit à définir des caractéristiques homogènes et des valeurs seuils identifiant des types d'interface habitat-forêt sur le territoire et à les cartographier. La *deuxième étape* a mis en relation l'organisation spatiale du territoire lue à travers la cartographie des interfaces habitat-forêt et l'historique des incendies de forêt. Fondé sur l'hypothèse que le risque d'incendie est lié à la structure spatiale du territoire selon des relations stables et reproductibles, l'objectif a été d'établir des relations entre les éléments du risque et les différents compartiments ou types de territoire (espaces dits « interfacés » et espaces dits « non interfacés »). Le risque d'incendie a été appréhendé en termes de distribution spatiale des points de départs de feu, des incendies et des surfaces brûlées correspondantes. La mise en relation a alors été recherchée entre les types de territoire identifiés à partir de la cartographie des interfaces habitat-forêt et les répartitions spatiales de départs de feu, d'incendie et de surfaces brûlées. Enfin la *dernière étape* a cherché à appréhender globalement et de façon synthétique les niveaux de risque d'incendie, à modéliser ce risque dans les interfaces habitat-forêt situées en région méditerranéenne française.

2. Caractériser et cartographier les interfaces habitat-forêt

Il y a de nombreuses manières de définir les interfaces habitat-forêt mais l'interface habitat-forêt est le plus communément définie comme une aire où les zones urbaines sont en contact et interagissent avec les zones rurales incluant les

bordures des grandes villes et petites agglomérations (Vince et *al.*, 2005), comme une zone où des dispositifs de développement humain se mélangent avec la végétation naturelle (Collins, 2005), comme une aire où les habitations ou autres activités humaines sont situées dans, ou au contact d'une végétation combustible (Summerfelt, 2001; Sanchez-Guisandez et *al.,* 2003). La définition développée dans l'article s'appuie sur l'existence de la loi d'orientation forestière du 11 juillet 2001 (Art. L.322.3) qui impose l'obligation de débroussailler dans un rayon de 50 m minimum autour des bâtis situés à moins de 200 m de forêts, garrigues ou maquis. L'interface habitat-forêt est délimitée par la surface dessinée par un rayon de 100 m autour des seuls bâtis de type résidentiel situés à moins de 200 m de tout massif forestier ou garrigues, maquis qu'ils soient occupés de façon permanente, temporaire ou saisonnière. Cette définition conduit à penser à élaborer une typologie d'interfaces habitat-forêt en fonction des parts relatives des systèmes « habitation » et systèmes « forêt ou autre espace naturel » et du niveau d'imbrication de leurs structures. La première hypothèse est de considérer que l'organisation de l'habitat résidentiel, sa structure spatiale a une influence sur le niveau de risque d'incendie. Selon sa nature - isolé, diffus ou groupé- la pression anthropique sera différente sur l'environnement, les enjeux seront plus ou moins importants en cas d'incendie, etc. La seconde hypothèse est de prendre en compte la structure horizontale de la végétation localisée en interface. L'analyse des retours d'expériences après incendie montre en effet que la structure de la végétation prime sur sa nature dans la propagation d'un incendie de forêt (Joliclercq, 2003). Une végétation éparse aura pour effet d'atténuer l'intensité d'un feu, de freiner sa propagation du fait d'une rupture dans la continuité de la végétation ; au contraire, une végétation compacte et continue alimentera le feu et lui maintiendra un niveau d'intensité critique jusqu'aux aux abords d'un bâti. Même si la structure verticale de la végétation joue un rôle également dans la propagation d'un feu en permettant le transfert du feu d'une végétation enflammée au sol jusqu'à la cime des arbres en cas de continuité verticale de la végétation, elle ne sera prise en compte dans l'approche spatiale.

La caractérisation et la cartographie les interfaces habitat-forêt a donc supposé de faire le choix de critères pertinents et quantifiables. Quatre types de structure d'habitat résidentiel (habitat isolé, diffus, groupé dense et groupé très dense) ont ainsi été définis, fondés sur des notions de distances entre bâtis et de regroupement de ces bâtis (Lampin-Maillet et *al.* 2009). Trois types de structure horizontale de végétation ont été identifiés (végétation continue et compacte, végétation éparse et discontinue, végétation absente) selon les valeurs fortes, faibles à nulles d'un indice d'agrégation calculé sur cette végétation (McGarigal, 2002). La combinaison de ces critères a conduit à construire une typologie d'interfaces habitat-forêt en 12 types (Lampin-Maillet et *al.* 2010). Le traitement des données spatiales sous SIG - pour l'essentiel, bâtis au format vecteur, végétation au format raster - a permis alors de cartographier ces interfaces habitat-forêt. Cette cartographie des interfaces habitat-forêt a produit une nouvelle carte du territoire. En effet le territoire s'est trouvé compartimenté selon une nouvelle clé de lecture : les espaces dits « interfacés » (Interfaces habitat-forêt avec une différenciation selon la typologie d'interfaces

élaborée en 12 types), et les espaces dits « non interfacés » (Espaces bâtis hors interfaces et le reste du territoire) (Fig.2).

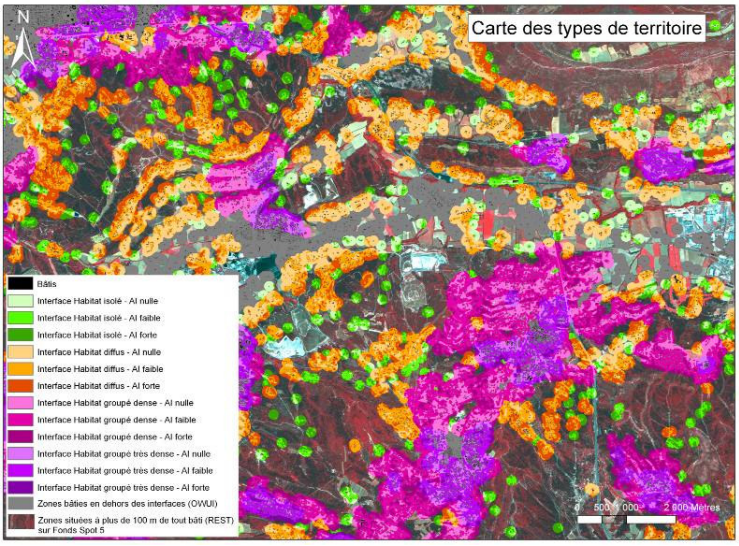


Figure 2. *Cartographie des types de territoire*

3. Mise en relation risque d'incendie et interfaces habitat-forêt

3.1 Approche globale à l'échelle du territoire

Le rapprochement de la distribution spatiale des départs de feu d'une part, et de la distribution spatiale des incendies d'autre part, avec les types de territoire, a mis en évidence une relation forte entre les types de territoire de nature interface habitat-forêt et l'importance des départs de feu et des taux de surfaces brûlées.

Les résultats ont montré que la densité des départs de feu est des plus élevées en interface habitat-forêt de type isolé et diffus au contact de la végétation (agrégation de la végétation non nulle), mais également en interface habitat-forêt de type groupé très dense au contact d'une végétation continue et compacte. Ces départs de feux sont essentiellement liés à la concentration d'activités humaines, mais aussi aux imprudences de la vie quotidienne (barbecue, activités de jardinage avec étincelles provenant d'outils…). Les résultats ont également montré que le taux de surfaces brûlées décroît des interfaces de type isolé aux interfaces de type groupé dense et très dense mais aussi avec une végétation de plus en plus éparse. Les interfaces de type isolé sont davantage menacées par les grands feux, du fait de la dispersion des

moyens de lutte, des temps de réponse plus longs liés à l'isolement et parfois à l'inaccessibilité des habitations correspondante (Sturtevand et Cleland, 2007). Si la diminution de la part de végétation dans les interfaces, souvent compensée par une augmentation de la part de surfaces urbanisées, conduit à un taux de couverture végétale inférieur à 30 %, la perméabilité de la végétation combustible est perdue et la propagation de l'incendie devient limitée, et une forte proportion d'espaces urbanisés interrompt la continuité de la végétation combustible (Syphard *et al.*, 2007). Ce qui est le cas des interfaces de type groupé dense et très dense (avec plus de 60 % de surfaces urbanisées). Enfin les valeurs du taux de surfaces brûlées ne sont pas nulles dans les interfaces avec une agrégation de la végétation nulle. Celles-ci peuvent en effet être parcourues par le feu qui se propage dans les champs, notamment les chaumes (Sturtevand et Cleland, 2007).

3.2 Approche analytique

L'approche globale a mis en lumière l'existence de relations entre types de territoire et risque d'incendie : certains types et, notamment certains types d'interface habitat-forêt sont davantage soumis au risque d'incendie en termes de densité de départs de feu et de taux de surfaces brûlées. Ces premiers résultats ont conduit à lancer de nouvelles investigations pour mieux connaître et comprendre l'environnement tant écologique, topographique que socio-économique qui conditionne les départs de feu et l'extension des surfaces incendiées. Une analyse spatiale et statistique approfondie a été entreprise en prenant en compte une large palette de variables d'occupation du sol disponibles, autres que les seuls types de territoire ou types d'interfaces habitat-forêt afin d'identifier les environnements les plus propices aux départs de feu et les plus affectés par les incendies. Trois indicateurs élémentaires de risque considérés comme pertinents ont été définis : densité de départ de feu, densité d'incendie et taux de surfaces brûlées.

Un espace plutôt naturel (forêts et garrigues), peu agricole mais avec une représentation urbaine plutôt forte (forte densité de bâtis de 178 bâtis/km^2 et de routes de 7 km/km^2) est propice à une densité de départ de feu non nulle. Et ce d'autant plus qu'il s'agira de zones d'expositions chaudes et très chaudes. En revanche la nature de la végétation ne joue pas de rôle prédéterminant. Cette densité de départ de feu augmente avec une plus forte représentation de l'espace naturel marqué par une densité de chemins élevée (7,3 km/km^2) et une moindre représentation de l'espace urbain (densité de bâtis de 59 bâtis/km^2 et de routes de 6 km/km^2). Ces caractéristiques s'apparentent davantage aux zones d'interfaces habitat-forêt qu'ailleurs.

Un espace naturel prédominant (forêts et garrigues) est propice à une densité d'incendie de forêt non nulle. Cette prédominance se confirme par une végétation très présente, continue ou éparse, constituée de peuplements mixtes et surtout de garrigues. La densité de chemins y est également plus forte avec 7,2 km/km^2 et les expositions plutôt chaudes et très chaudes sont davantage présentes. Cette densité

d'incendie augmente avec une moindre représentation de l'espace urbain (densité de bâtis de 42 bâtis/km^2 et de routes de 3 km/km^2. Là encore, ces caractéristiques s'apparentent davantage aux zones d'interfaces habitat-forêt qu'ailleurs.

Enfin l'environnement pour lequel le taux de surfaces brûlées est non nul est similaire à celui d'une densité d'incendie non nulle.

4. Calcul d'un indice global du risque d'incendie et cartographie

Les résultats de l'analyse réalisée dans l'approche analytique précédente ont permis de souligner quelques caractéristiques d'occupation du sol, d'environnement naturel et topographique, les plus propices aux départs de feu, à la présence d'incendies et à leur extension. La contribution de certaines variables à l'explication du risque d'incendie a alors été mise en évidence par la modélisation de chacun des trois indicateurs : densité de départ de feu, densité d'incendie et taux de surfaces brûlées. Par des régressions de type Moindres carrés partiels PLS Partial Least Squares, chaque indicateur a ainsi été modélisé sous la forme d'une combinaison linéaire de variables relatives à l'environnement naturel, physique et socio-économique. Ces variables se sont avérées comme les plus significatives avec des poids relatifs, contribuant de façon positive ou négative à l'explication de chacun des trois indicateurs sur le territoire étudié. Trois équations ont été produites.

La première équation [1] concerne l'indicateur de Densité de départ de feu ou d'éclosion DE. Sept variables contribuent de façon significative à l'explication de cet indicateur. Ces variables sont : (i) le type de territoire appelé interface habitat-forêt en habitat isolé I et la part occupée par les autres espaces naturels ESN qui contribuent positivement ; (ii) le type de territoire non bâti R, la densité de bâtis DB, la part occupée par l'espace urbain URB, l'interface habitat-forêt en habitat groupé dense GD et l'interface habitat-forêt en habitat groupé très dense GTD qui contribuent négativement.

Densité de départ de feu ou d'éclosion **DE** = exp (2,30258509*[***1,76489*** + ***0,00558842*** *I* **- *0,00240165*** *GD* **- *0,00105965*** *GTD* **- *0,00609774*** *R* **- *0,00065618*** *DB* **+ *0,00465397*** *ESN* **- *0,00512739*** *URB*]**)** [1]

La deuxième équation [2] concerne l'indicateur de Densité d'incendie DI. Neuf variables contribuent de façon significative à l'explication de l'indicateur. Ces variables sont : (i) le type de territoire appelé interface habitat-forêt en habitat isolé I et la part occupée par les autres espaces naturels ESN et les zones d'exposition très chaudes KR5 qui contribuent positivement ; (ii) le type de territoire non bâti R, la densité de bâtis DB, l'interface habitat-forêt en habitat groupé très dense GTD, la part occupée par l'espace urbain URB, la présence de végétation résineuse VG2 et l'interface habitat-forêt en habitat groupé dense GTD qui contribuent négativement.

Densité d'incendie **DI** = exp (2,30258509*[***2.09384*** + ***0,00247646*** *I* **-** ***0,0011186*** *GD* **-** ***0,00301069*** *GTD* **-** ***0,0117099*** *R* **-** ***0,000994732*** *DB* + ***0,00258941*** *KR5* – ***0,00420811*** *VG2* + ***0,00303519*** *ESN* **-** ***0,00301263*** *URB***]**) [2]

La troisième équation [3] concerne l'indicateur taux de surfaces brûlées SB. Onze variables contribuent de façon significative à l'explication de cet indicateur. Ces variables sont : (i) la part occupée par les autres espaces naturels ESN, et la végétation de garrigue VG4, le type de territoire, interface habitat-forêt en habitat isolé I, la densité de chemins DC, l'agrégation faible de la végétation AI2 et l'altitude ALT qui contribuent positivement ; (ii) la part occupée par l'espace agricole AGR, l'interface habitat-forêt en habitat groupé dense GD, la densité des routes DR, la densité de bâtis DB et la part occupée par l'espace urbain URB qui contribuent négativement.

Taux de surfaces brûlées **SB** = ***29,292*** + ***0,093933*** *I* **-** ***0,100626*** *GD* **-***0,0246026* *DB* **-***0,663865* *DR* + ***0,625354*** *DC* + ***0,408128*** *VG4* + ***0,149018*** *AI2* + ***0,0199612*** *ALT* **-** ***0,18559*** *AGR* + ***0,296497*** *ESN* **-** ***0,119762*** *URB***.** [3]

La modélisation de chacun des trois indicateurs de risque est obtenue avec des valeurs du coefficient de détermination R^2 qui permettent de mesurer la fiabilité des modèles. Ainsi les modélisations relatives à DE et DI présentent des valeurs de R^2 respectives de 51 % et 57 %. Les facteurs pris en compte peuvent expliquer de façon satisfaisante les densités d'éclosion et d'incendies. La modélisation SB présente une valeur plus faible de R^2 de 36 %. Le modèle est moins bon bien qu'il soit significatif. Syphard et *al.* (2007) ont trouvé également de meilleurs résultats sur la fréquence des incendies plutôt que sur les surfaces brûlées avec des coefficients de R^2 du même ordre de grandeur pour les surfaces brûlées.

Les trois indicateurs de risque définis DE, DI et SB, permettent d'appréhender la notion de risque d'incendie. Chacun de ces indicateurs est en effet porteur de tout ou partie d'informations relatives à l'aléa et à la vulnérabilité. Ainsi DE et DI contribuent à apprécier l'occurrence du phénomène incendie de forêt (Probabilité d'éclosion, Probabilité d'incendie) et SB permet quant à lui de mesurer le niveau d'intensité de l'incendie (à considérer en termes d'aléa et de vulnérabilité) et le taux de dommages notamment l'impact sur les espaces naturels. A partir de la combinaison de ces trois indicateurs élémentaires de risque Densité d'éclosion DE, Densité d'incendie DI, Taux de surfaces brûlées SB, un indice de risque IR, unique, global et synthétique a été produit. Sa formule est une combinaison linéaire des indicateurs DE, DI et SB pondérée de la performance du pouvoir explicatif de chaque indicateur. En considérant que le meilleur modèle obtenu pour les indicateurs de risque est celui de la densité d'incendie DI, DI a été pris en référence et on lui a affecté la valeur 1. A l'indicateur Densité d'éclosion DE dont le pouvoir explicatif est de 51 %, une valeur de 0,89 correspondant au ratio de 51/57 a été affectée, 57 étant la valeur du pouvoir explicatif de l'indicateur Densité d'incendie DI. De la même façon, une valeur 0,63 a été affectée à l'indicateur Taux de surfaces brûlées SB.

Ainsi, l'indice global de risque d'incendie IR (Lampin-Maillet, 2009) est défini par la formule suivante :

IR = 0,89 DE + DI + 0,63 SB

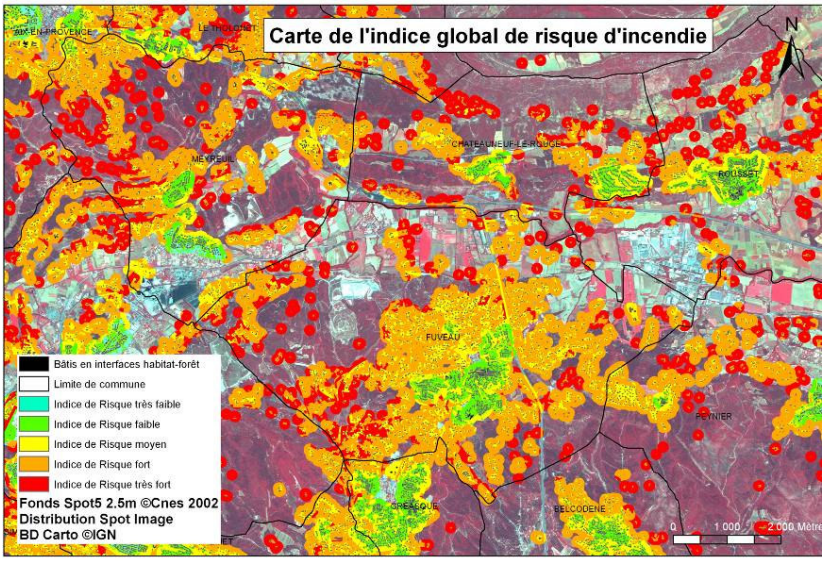


Figure 3. *Carte de l'indice global de risque d'incendie*

Ce processus d'évaluation du risque d'incendie évite un diagnostic selon la démarche analytique traditionnelle (aléa + vulnérabilité). Il est construit à partir de trois indicateurs élémentaires de risque, chacun d'entre eux étant une combinaison linéaire de quelques variables reconnues statistiquement comme les plus significatives. Ces variables relèvent de facteurs tant physiques qu'humains, et sont porteuses également de l'information sur l'aléa et sur la vulnérabilité. La combinaison linéaire de ces trois indicateurs élémentaires, qui ont été considérés d'un poids égal, a été corrigée de la part de contribution explicative de chaque indicateur. Elle a ainsi produit un indice global de risque. Cet indice, calculé sous SIG, a été traduit par une carte de risque dans les interfaces (Fig.3).

5. Conclusion

Dans l'article l'interface habitat-forêt est définie de façon précise dans le contexte du risque d'incendie. Une typologie d'interfaces est créée, fondée sur la combinaison de deux critères jugés pertinents pour le risque d'incendie, traduisant

des caractères prégnants des milieux humain, avec la structure de l'habitat résidentiel, et naturel, avec la structure de la végétation. La méthode de caractérisation et de cartographie des interfaces habitat-forêt est applicable sur de grandes surfaces et à une grande échelle. La carte des interfaces habitat-forêt contribue alors à produire une nouvelle carte du territoire, alors compartimenté en espaces dits « interfacés » (types d'interfaces habitat-forêt), et en espaces dits « non interfacés » (espaces bâtis hors interfaces et le reste du territoire). Une première relation, forte, entre les types d'interface habitat-forêt et l'importance des départs de feu et des taux de surfaces brûlées a pu être mise en évidence. Une méthode d'évaluation du risque d'incendie, innovante, a alors été développée. Elle s'appuie sur une analyse spatiale et statistique du territoire, fondée sur une nouvelle cartographie de types de territoire déduite de la cartographie des interfaces habitat-forêt. L'analyse a consisté à croiser les types de territoire et les caractéristiques environnementales, topographiques et socio-économiques avec l'historique des feux à travers la distribution spatiale des départs de feu, celle des surfaces brûlées et la fréquence de passage des incendies. Elle a permis de mettre en évidence l'importance de certaines variables pour leur contribution positive ou négative à l'explication de trois indicateurs de risque définis comme densité d'éclosion, densité d'incendie et taux de surfaces brûlées. La modélisation de ces indicateurs a contribué à la construction d'un indice global de risque et à sa cartographie qui permet de déduire facilement, et de manière assez directe, l'information synthétique sur les niveaux de risque à l'échelle du territoire.
Ainsi l'approche par les « interfaces habitat-forêt », intrinsèquement porteuses de l'information synthétique aléa/enjeux/vulnérabilité, a servi de clé d'entrée pour une évaluation directe et globale du risque, fondée sur l'observation et la description des territoires d'une part, et en particulier des interfaces habitat-forêt, et sur une analyse spatiale et statistique de ces territoires. Elle permet également de tirer des enseignements d'une meilleure connaissance du territoire et du risque d'incendie associé en termes de prévention.

Remerciements

Les auteurs remercient le Ministère de l'Écologie, de l'Énergie, du Développement durable et de la Mer, le Ministère de l'Agriculture, le Conseil Régional Provence-Alpes-Côte-d'Azur pour le financement de programmes de recherche ainsi que la Commission Européenne pour le financement du projet Européen Fireparadox n° FP6-018505.

6. Bibliographie

Cohen, J.D. Preventing disaster: Home ignitability in the wildland-urban interface. *Journal of Forestry, 98 (2000)* (3), pp 15-21.

Collins, T. W. Households, forests, and fire hazard vulnerability in the American West: a study case of a Califorlia community. *Environmental Hazards,* 6 (2005), pp 23-37.

Davis, J. B. “The wildland-urban interface : paradise or battleground ?”. *Journal of forestry 6 (1990)*, 88 (1), 26-31.

Dumas, E., Jappiot, M., Tatoni, T. Mediterranean urban-forest interface classification (MUFIC): A quantitative method combining SPOT5 imagery and landscape ecology indices. *Landscape and Urban Planning* . 84 (2008), 183–190.

Hardy, C.C. Wildland fire hazard and risk: Problems, definitions, and context. *Forest ecology and management* (2005), 211, pp 73-82.

Jappiot, M., Gonzales-Olabarria, J.R., Lampin-Maillet, C., Borgniet, L. Assessing wildfire risk in time and space. *In* Living with wildfires: What science can tell us? A contribution to the science-policy dialogue. (Birot., Y. Eds European Forest Institute), pp 41-47.2009.

Joliclercq, F. OFME-EGA - Diaporama : Quelle politique de prévention et d’aménagement du territoire régional ? Débroussaillement obligatoire et autoprotection des habitations. Retours d’expérience après incendie. http://www.ofme.org/affdoc.php3?ID=95&Page=1. (2003).

JRC. Statistics 1980-2006, data source JRC-IES Report n°7, 2006.

Lampin-Maillet, C., Jappiot, M., Long, M., Bouillon, C., Morge, D., Ferrier, J.P. 2010a. Mapping wildland-urban interfaces at large scales integrating housing density and vegetation aggregation for fire prevention in the South of France. *Journal of Environmental Management,* 91 (2010), pp 732–741.

Lampin-Maillet, C. Caractérisation de la relation entre organisation spatiale d’un territoire et risque d’incendie : Le cas des interfaces habitat-forêt du sud de la France. Thèse de doctorat de l’université Aix-Marseille, mention Lettres et Sciences humaines (Géographie- Structures et dynamiques spatiales). 325 pages + annexes, 2009.

Lampin-Maillet, C., Jappiot, M., Long, M., Morge, D., Ferrier, J.P. Characterization and mapping of dwelling types for forest fire prevention. *Computers, Environment and urban systems* 33 (2009), pp. 224-232.

Lampin-Maillet, C. *2007, Summer Fires in the European Mediterranean – The Cases of Greece, Italy and Spain*. Mediterranean yearbook. European Institute of the Mediterranean. Med.2008, Economy and Territory- Sustainable Development, p 243-247. http://www.iemed.org/anuari/2008/aarticles/EN243.pdf, 2008.

McGarigal, K. Landscape Pattern Metrics. Chapitre du livre Encyclopedia of Environmentrics, Volume 2, John Wiley & sons, Sussex, England. (2002), pp 1135-1142.

Sanchez-Guisandez, M., Cui, W., Martell, D.L. FireSmart Strategies for wildland urban interface landscapes. In Proceedings (Eds Xanthopoulos, G.) of the international workshop WARM, Forest fires in the wildland-urban interface and rural areas in Europe: an integral planning and management challenge. Athens, Greece. (2003), pp 121-130.

Sturtevant, B.R., Cleland, D.T. Human and biophysical factors influencing modern fire disturbance in northern Wisconsin. *International Journal of Wildland Fire*. 16 (2007), pp 398-413.

Summerfelt, P. The Wildland-Urban interface. What's really At risk? http://www.gffp.org/pine/risk/default.htm 4/14/03, (2001)

Syphard, A.D., Clarke, K.C., Franklin, J. Simulating fire frequency and urban growth in southern California coastal shrublands, USA. *Landscape Ecology*.22 (2007), 431-445.

Theobald, D.M., Romme, W.H. Expansion of the US wildland-urban interface. *Landscape and Urban Planning*. 83 (2007), 340-354.

Vince, S.W., Duryea, M.L., Macie, E.A., Hermansen, L.A.. Forests at the wildland-urban interface: conservation and management (2005) - Boca Raton, CRC Press).

Structure informatique pour la réponse aux plaintes liées à l'air au sein des logements

Zoulikha Bellia Heddadji[*,], Nicole Vincent[*], Séverine Kirchner[**] et Georges Stamon[*]**
Laboratoire LIPADE.
[*]Équipe SIP. 45, rue des Saints Pères. 75270 Paris Cedex 06
{bellia, vincent, stamon}@math-info.univ-paris5.fr
[**]Centre Scientifique et Technique du Bâtiment (CSTB)
84 avenue Jean Jaurès. Champs-sur-Marne. 77447 Marne-la-Vallée Cedex 2
severine.kirchner@cstb.fr

RÉSUMÉ. *Les effets de la pollution de l'air à l'intérieur des ouvrages de construction que nous occupons (logements, écoles, bureaux, hôpitaux, etc.) sur la santé publique sont au moins aussi importants que la pollution de l'air extérieur. Par conséquent, les pouvoirs publics accordent de plus en plus d'intérêt aux études sur la qualité des environnements intérieurs, notamment au sein des logements. Notre travail s'inscrit dans cette perspective. En effet, notre objectif est de mettre en œuvre une structure informatique qui réunit plusieurs applications dans le but de répondre automatiquement à une plainte d'un particulier écrite entièrement en langue naturelle et qui soit bien sûr liée à une situation de pollution de l'air au sein des logements. Par « réponses » nous entendons une précision concernant la nature du problème à l'origine des symptômes décrits ainsi qu'un ensemble d'actions correctives permettant de réduire les effets du problème sanitaire cité. L'approche suivie consiste en premier lieu à émettre l'hypothèse de l'existence d'une régularité des phénomènes de pollution intérieure pour créer des scénarios. Tout d'abord, nous apportons la preuve du bien-fondé de l'hypothèse initiale à partir d'un corpus représentatif de plaintes résolues sur le terrain. La motivation majeure de la constitution de scénarios est de réaliser des solutions génériques adaptées à chacun. Enfin, notre application implémentant des systèmes de recherche d'information structurée directe et sémantique se chargera de mettre en évidence le scénario auquel appartient une plainte à traiter et d'assigner à cette dernière la solution attribuée au scénario désigné. Nous nous sommes par ailleurs beaucoup intéressés à l'analyse de la qualité des différents résultats d'assignation de solution dans un contexte structuré des textes exploités selon différentes mesures de similarités.*

MOTS-CLÉS : *pollution de l'air intérieur, système de recherche d'information, dictionnaire des synonymes, classification de documents.*

1. Introduction

Les citoyens passent en moyenne plus de 80% de leur temps en environnements clos. La réalité est que cette moyenne est d'autant plus élevée chez les populations les plus fragiles: jeunes enfants, personnes âgées et/ou malades. Ces populations sont exposées à des contaminants forcément plus concentrés au sein de leurs lieux de vie. Ceci entraîne dans certains cas une aggravation de certains symptômes à l'intérieur (éternuements, toux, sensations de gêne respiratoire, etc.) et leur disparition une fois l'individu se trouve à l'extérieur. Ces polluants proviennent de sources diverses. En effet, ces contaminants peuvent être de sources naturelles comme les animaux de compagnie, ou bien des matériaux de construction intérieure (colle, plastique, solvant de peinture, etc.), ou de sources dites d'activité comme le bricolage ou le tabagisme, ou encore les activités ménagères (aérosols d'entretien, etc.). Le problème est en réalité complexe. En effet, la plupart des symptômes ne sont pas très spécifiques. De plus, les atmosphères intérieures d'un ancien appartement, d'une maison individuelle, d'un bureau loti au sein d'une tour ultramoderne par exemple ne peuvent pas avoir grand-chose en commun. Néanmoins, les populations sont de plus en plus sensibilisées par rapport à leurs conditions de vie. Les exemples récents dans la presse témoignent en effet d'une importante recrudescence des plaintes en lien avec l'air des différents lieux de vie. Il existe aujourd'hui un grand nombre de demandes de renseignements et d'investigations témoignant de l'étendue du phénomène à travers le pays. Ces plaintes se présentent le plus souvent sous forme de demandes d'interventions écrites ou téléphoniques auprès des autorités. Parmi ces autorités qui reçoivent les plaintes de particuliers en France nous avons pu distinguer la Direction Générale de la Santé (DGS), les Directions des Affaires Sanitaires et Sociales (DDASS), les Services Communaux d'Hygiène et de Santé (SCHS) et le Centre Scientifique et Technique du Bâtiment (CSTB). Dans le cadre de notre application, notre choix s'est porté sur les ouvrages de construction à usage d'habitation exclusivement, et cela par rapport au temps important passé par la population au sein de ces lieux de vie, et également par rapport à la nécessité de la prise en compte des personnes sanitairement sensibles (enfants, personnes âgées, etc.).

2. Réflexion autour des approches possibles

Le système de réponse aux plaintes écrites que nous avons développé devait permettre aux utilisateurs de s'exprimer en langue naturelle comme ils le feraient dans le cadre d'une lettre classique. Cette option offre des avantages majeurs. Nous avons proposé d'utiliser la langue naturelle essentiellement pour éviter l'utilisation des questionnaires fermés. En effet, la pollution des environnements clos est un domaine récent et il reste encore des facteurs de risque peu ou mal connus. Par conséquent, la description des différentes situations possibles ne peut se limiter aux

éléments de formulaires fermés. Donc, le système proposé s'inscrit plus exactement dans la lignée des systèmes de recherche d'information.

Le système s'appuie sur une base archive évolutive constituée de plaintes écrites résolues. Cette base est évolutive puisqu'elle est destinée à se développer au fur et à mesure que de nouveaux cas seront saisis, résolus et vérifiés à travers l'applicatif proposé. Dans (Bellia, 2008), nous réalisons un état de l'art de plusieurs paradigmes pouvant être appliqués dans le cadre de l'approche de résolution des plaintes écrites. Parmi ces modèles de représentation des ressources et de raisonnement nous pouvons citer ici les systèmes experts et le raisonnement à partir de cas. Étant donné qu'aujourd'hui il n'existe pas de modèle formalisé permettant d'appréhender les circonstances de pollution de l'air au sein des lieux de vie, l'expérience acquise à ce jour sous forme d'archive est prédominante. Par conséquent nous n'avons pas pu définir une approche fondée sur le paradigme des systèmes experts. Le RàPC semblait au départ une idée très appropriée à la philosophie et à la logique d'action des experts. Cependant, la phase d'adaptation qui s'occupe de remettre dans le contexte du cas courant la solution d'un cas pertinent situé en mémoire par rapport à un cas à traiter n'a été que très peu abordée dans la littérature. Par ailleurs, une des rares études s'étant intéressée à la formalisation de la phase d'adaptation dans le cadre du RàPC textuel est celle de Luc Lamontagne (Lamontagne, 2004). Son étude formalise cette phase dans le cadre de la mise en place d'une application dédiée à la réponse automatique à des courriers électroniques. Une des étapes de ce procédé est la phase de sélection des extraits pertinents. Ces passages sont naturellement sujets à modification, et pour les distinguer, Lamontagne propose de réaliser une extraction des entités nommées, qui est une technique du TAL nécessitant des ressources externes appropriées. Malheureusement, nous ne possédons pas cette ressource. Généralement, et le plus souvent, ces ressources sont en effet incomplètes, complexes et difficiles à la mise en œuvre.

3.Notre approche

À partir de notre corpus représentatif des plaintes résolues nous avons constaté une régularité à partir des rapports d'experts alors que les textes des plaintes apparaissent différents lexicalement et sémantiquement. Par conséquent, nous avons souhaité connaître le nombre et la nature des classes de plaintes possibles. Ces classes résumeront le domaine de la pollution domestique à partir de l'échantillon représentatif du corpus qu'on a pu nous fournir tout au long de cette étude. Afin de vérifier notre hypothèse concernant la régularité thématique des plaintes, nous réalisons d'abord une segmentation automatique d'un échantillon représentatif de la base des textes. À partir du même corpus, nous demandons à des experts de regrouper les plaintes selon la nature du problème de pollution intérieure. Ensuite, des taux d'accord sont calculés entre les classes automatiques et les classes des experts. Les experts n'avaient pas conscience a priori de ces classes mais peuvent interpréter et réaliser des regroupements. Une fois l'hypothèse vérifiée nous établissons un rapport type de solution associé à chaque scénario parmi l'ensemble des classes extraites et validées par les experts. Nous pouvons considérer la réalisation des solutions génériques comme une alternative à l'adaptation du RàPC. Afin d'assigner une solution appropriée à une plainte à traiter, un système de recherche est utilisé. Cela revient à attribuer la nouvelle plainte écrite à un scénario (ou une classe de plaintes) (figure 1). Le module fonctionnel est chargé d'apparier le texte de la plainte nouvelle aux textes des plaintes résolues regroupées au sein des scénarios en mémoire. Pour cela, nous avons étudié et développé différents modèles de recherche d'information. Ce choix est essentiellement dû à la nature hétérogène des textes de notre corpus. En effet, l'hétérogénéité concerne d'une part la taille des documents, et d'autre part celle-ci concerne la quantité d'information "a priori" sémantiquement inconsistante (le bruit).

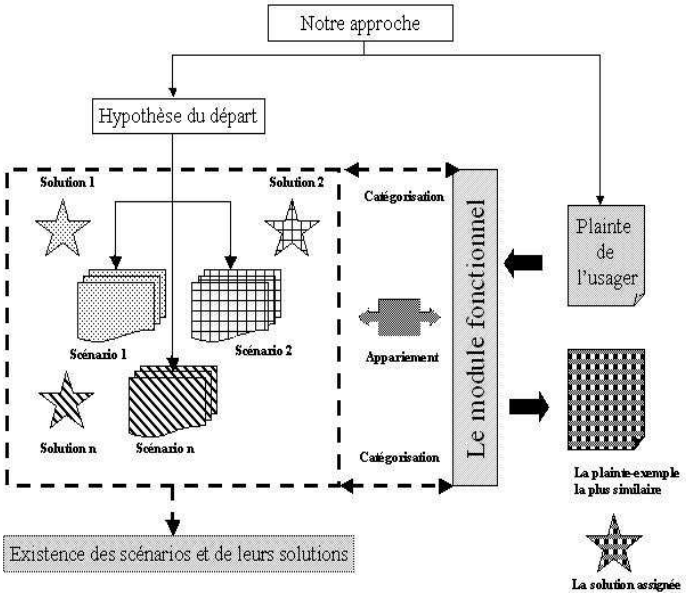


Figure 1. Architecture synoptique de l'approche proposée

3.1. *Prétraitement des textes des plaintes*

Une étape de lemmatisation et de filtrage des textes est nécessaire pour l'ensemble des plaintes résolues stockées en mémoire et les textes à traiter. Pour s'affranchir des différentes flexions des termes nous avons utilisé le lemmatiseur TreeTagger adapté au français (Schmidt, 1994). Suite à la lemmatisation d'un texte uniquement les lemmes sont maintenus en sortie. Un dictionnaire d'arrêt, validé par un groupe d'experts, est utilisé afin d'éliminer automatiquement les mots vides de sens à partir de la forme lemmatisée des textes.

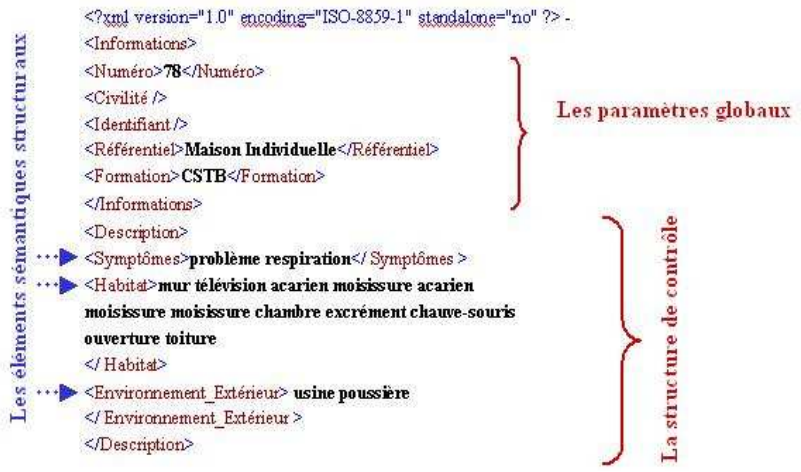


Figure 2. Exemple d'une plainte structurée au format XML

3.2. *Formalisme des données*

Concernant la structure discursive des plaintes, malgré le fait qu'aucune régularité conversationnelle n'est explicite sur l'ensemble des textes au départ, une certaine structure "rhétorique" [1] apparaît de manière fréquente à travers le corpus. Nos experts se sont unanimement prononcés en faveur de la collecte des plaintes en utilisant les items suivants :

- Description des symptômes: perception et description de la pathologie par le plaignant.
- Description de l'environnement extérieur: permettant d'évoquer par exemple l'existence d'une rue à fort trafic, d'usines, de travaux, de sources de pollens, etc.
- Description de l'habitat et des habitudes de vie: description de son équipement, le mobilier, les systèmes de chauffage, l'usage de produits chimiques, etc.

Par conséquent, l'interface usager de notre applicatif, permettant la saisie de la plainte, se présente sous forme d'un questionnaire permettant la saisie de ces 3 champs. Ces 3 champs correspondent plus formellement à des éléments structurants sémantiquement pertinents permettant de conserver les plaintes sous forme XML. Ainsi, les balises XML (symptôme, habitat, environnement_extérieur) délimitent le

[1]La Théorie de la Structure Rhétorique (en anglais RST) concerne la structure des usages langagiers, et plus spécifiquement la structure discursive des textes écrits.

contenu de chaque partie (ou rubrique) de la plainte (figure 2). Nous avons choisi d'abord de mettre en œuvre l'adaptation structurelle XML du modèle vectoriel (Zargayouna, 2005) qui est dérivé du modèle vectoriel classique de Salton (Salton, 1991) reconnu comme étant mieux adapté au traitement des textes longs. Par ailleurs, et pour le traitement des textes courts, nous avons implémenté le modèle de proximité floue.

3.3. *Les modèles de recherche développés et utilisés dans le cadre du module fonctionnel d'appariement*

3.3.1. Adaptation structurelle de Zargayouna

Pour la modélisation et l'appariement des documents XMLisés, un terme du vocabulaire utilisé a un poids TF-ITDF (« term frequency-inverse tag and document frequency ») au sein de chaque balise. Par conséquent une matrice[2] des poids des termes est le modèle du document (formules 1, 2, 3). D'abord une similarité locale est estimée entre deux rubriques de deux documents à l'instar du modèle classique en calculant la valeur du cosinus de l'angle formé par leur deux vecteurs représentatifs. Le score de similarité globale correspond à l'agrégation des similarités locales. Cela pourrait être une somme, une moyenne, une moyenne pondérée, etc.

$$TF-ITDF(t,b,d)=TF(t,b,d)\,ITF(t,d)\,IDF(t,b) \quad [1]$$

Sachant que:

- TF(t,b,d) « term frequency » est la fréquence d'apparition du terme t dans la balise b du document d.

$$ITF(t,d)=\log(|D_B|/DF(t,b)) \quad [2]$$

- $|D_b|$ correspond au nombre de documents où le modèle de balise (ou rubrique) b est renseigné.

$$IDF(t,d)=\log(|B_d|/TagF(t,d)) \quad [3]$$

- $|B_d|$correspond au nombre total des modèles de balises (ou rubriques) renseignés dans le document d.

[2]Le nombre des lignes de la matrice des poids des termes correspond à la taille des termes du vocabulaire considéré et les colonnes correspondent au nombre total des rubriques retenues pour la composition des documents.

3.3.2. Le modèle de proximité floue

Inspirée du modèle booléen et le modèle booléen pondéré, l'approche de Mercier (Mercier, 2004) repose sur l'hypothèse que plus les occurrences des termes d'une requête se retrouvent proches dans un document, plus ce document est pertinent par rapport à cette requête. Pour cela, elle calcule aux différentes positions x au sein d'un document un niveau de pertinence μ pour chaque terme t de la requête (formule 4). Si un terme de la requête est rencontré au sein du document la position de son occurrence prend la valeur 1. Plus on s'éloigne de cette position, plus cette valeur

$$\mu_t^d(x) = max_{i \in d^{-1}(t)}(max(\frac{k-|x-i|}{k},0)) \quad [4]$$

diminue de 1/k (au prorata d'un indice k qui est à fixer a priori du calcul de la mesure. Il est pris souvent égale à 10 pour évaluer la proximité au sein de la phrase, ou bien entre 20 et 100 pour évaluer la densité d'apparition d'occurrences de termes au sein du chapitre, etc.). Dans le cas où plus d'une occurrence d'un terme de la requête apparaît dans le document, la valeur maximale des pertinences issues des différentes occurrences est fixée (tableau 1)[3]. Au final, pour évaluer la pertinence du document et par analogie à l'ancienne mesure, le niveau de coordination, Mercier retient la valeur maximale des pertinences locales aux différentes positions dans le cas où la requête est purement disjonctive (elle tient compte du minimum dans le cas où la requête est purement conjonctive, une combinaison ordonnée de ces opérateurs sinon). Ensuite, elle tient compte de la moyenne des taux de pertinence par rapport à la taille $|d^{-1}|$ du document. Ce modèle a été défini dans un cadre non structuré. Par conséquent, nous avons proposé une adaptation de ce modèle dans le cadre de nos textes XMLisés en réalisant une moyenne (simple et pondérée) des similarités locale. Par ailleurs, ce modèle est asymétrique, puisque la requête constitue la référence. En effet, $Sim(D_1, D_2)$ n'a pas toujours la même valeur que $Sim(D_2, D_1)$. En effet, nous calculons la densité de la requête dans le document et non pas l'inverse. Cette spécificité est maintenue en tant que mesure de similarité mais nous employons également la formule résultant d'une symétrisation de l'ensemble des modèles asymétriques développés au sein de cette structure en réalisant une moyenne des valeurs réciproques.

[3]k=10 dans l'exemple du tableau.

x	0	1	2	3	4	5	6	7	8	9	10
d		A		B			C		A	B	C
μ_A^d	0.9	1	0.9	0.8	0.7	0.7	0.8	0.9	1	0.9	0.8
μ_B^d	0.7	0.8	0.9	1	0.9	0.8	0.7	0.8	0.9	1	0.9
μ_C^d	0.4	0.5	0.6	0.7	0.8	0.9	1	0.9	0.8	0.9	1
μ_{AETB}^d	0.7	0.8	0.9	0.8	0.7	0.7	0.7	0.8	0.9	0.9	0.8
$\mu_{(A\ ET\ B)\ OU\ C}^d$	0.7	0.8	0.9	0.8	0.8	0.9	1	0.9	0.9	0.9	1

Tableau 1. Les valeurs de proximité floue locale dans l'exemple de Mercier

3.3.3. Le modèle vectoriel étendu en vue bidimensionnelle et sémantique

La fonction SemW(t,b,d) (formule 5) réévalue la pondération des termes en tenant compte des liens sémantiques entre termes au moyen d'une ontologie du domaine. Au moyen de cette nouvelle évaluation des scores, le poids d'un terme qui n'apparaît pas directement dans une unité sémantique (balise ou rubrique dans notre cas) peut être augmenté en fonction des scores TF-ITDF des termes (t_i dans la formule 5) appartenant à ce contexte (balise) et qui sont sémantiquement liés au terme recherché (t dans la formule 5).

$$SemW(t,b,d) = TF-ITDF(t,b,d) + \frac{\sum_{i=1..n} Sim(t,t_i)\,TF-ITDF(t_i,b,d)}{n} \quad [5]$$

Sachant que n désigne le nombre de termes t_i proche sémantiquement de t existants au sein du modèle de balise b.

3.3.4. Adaptation sémantique du modèle de proximité floue

Le modèle de Mercier est limité par la relation de co-occurrence directe des termes et ne tient pas compte des éventuels liens sémantiques qui peuvent exister entre les termes de la requête et ceux du document. L'intégration d'une mesure sémantique entre termes dans ce modèle nous a semblé nécessaire. Le principe de notre modèle augmenté est d'observer au sein du document non seulement les positions prises par les termes de la requête mais aussi les termes qui leur sont sémantiquement proches (formule 6). Sachant que t est le terme de la requête, x est la position à laquelle on souhaite évaluer le taux de pertinence, Sem(t) est l'ensemble des termes sémantiquement proches possédant au moins une occurrence dans d. $d^{-1}(Sem(t))$ est l'ensemble des positions prises par les termes sémantiquement proches de t (donc l'ensemble Sem(t)) au sein du document d. $Sim(t_i, t)$ désigne le taux de similarité entre le terme t et le terme t_i avec qui il partage un sens commun.

$$\Psi s_t^d(x) = max_{i \in d^{-1}(Sem(t))}(max(\frac{k - |x-i| Sim(t_i, t)}{k}, 0)) \quad [6]$$

3.3.5. Le modèle de recherche fondé sur la superposition des ondes d'information

Rappelons qu'une plainte est saisie en langue naturelle par les soins de l'usager du système. L'utilisation de requêtes booléennes pour supplanter l'expression naturelle des besoins de manière générale, même enrichies de connecteurs variés, nécessite une mise en forme manuelle soigneuse et coûteuse des requêtes. En effet, nous n'avons pas connaissance d'outils permettant de traduire automatiquement une expression en langue naturelle vers son interprétation sous forme booléenne. Dans l'étude expérimentale (Dinet, 2000) concernant l'usage des requêtes booléennes en vue de recherche d'information, Dinet positionne la logique impliquée par les opérateurs booléens par rapport à la logique utilisée habituellement par un individu. Par exemple, l'opérateur ET implique une inclusion dans le langage naturel alors qu'il implique une exclusion (des résultats) dans le langage documentaire. Par exemple, lorsqu'une personne demande un « croissant » ET un « café », elle espère avoir les deux. Par contre, dans le cadre d'une recherche documentaire, demander « croissant » ET « café » correspond à une restriction du champ de réponses. De même, l'opérateur OU implique une inclusion dans le langage documentaire alors qu'il implique une exclusion (une restriction) dans le langage naturel. Parfois, dans la vie courante, il faut choisir: « boire » OU « conduire ». Vis à vis de cette différence entre les deux langages, nous nous sommes intéressés à adapter le modèle de Mercier. Ainsi, nous nous sommes inspirés d'un principe que nous appelons les ondes d'information. Nous présentons un nouveau modèle de recherche en émettant une hypothèse. Cette dernière consiste à supposer que les termes de la requête émettent des ondes au sein des positions possibles dans les documents. À l'aide de ce modèle, la superposition des ondes émises par les termes de la requête au sein des

documents détermine la densité de la requête dans ces documents. Dans le contexte de notre adaptation, les ondes à traiter proviennent de sources connues et sont engendrées dans les documents dont on souhaite évaluer la pertinence. Ces sources correspondent aux occurrences des termes de la requête. Le support des ondes correspond à l'intervalle discret borné [1, $|d^{-1}|$], où d^{-1} désigne l'ensemble des positions pouvant être prises au sein du document d. L'amplitude de l'onde d'information émise par un terme t de la requête à une position x du document correspond au degré d'influence du terme t au niveau x. L'amplitude maximale d'un terme t à une position x est de 1 lorsqu'il existe une occurrence de t en x. L'amplitude diminue au prorata de la distance. Notre modèle (formule 7) ne tient pas compte du paramètre k permettant d'évaluer la pertinence locale dans le modèle de Mercier. Ce paramètre qui définit la taille de la zone d'influence d'un terme, correspond ici à la taille du support d^{-1}. Ceci est mieux adapté à notre application, puisque la taille de nos rubriques varie et nous ne pouvons dépendre d'un paramètre fixé a priori.

$$\zeta_t^d(x) = \sum_{i \in d^{-1}(t)} \left(max\left(1 - \frac{|x-i|}{|d^{-1}|}, 0\right)\right) \quad [7]$$

La mesure d'appariement prend en considération le principe de superposition des ondes aux différentes positions actives. La somme des amplitudes des interférences aux positions actives, correspond au niveau de pertinence de la requête par rapport au document. Tenir compte des positions actives uniquement est important. En effet, nous souhaitons faire en sorte que la variation des densités des termes de la requête ne soit pas fondée uniquement sur l'effet de bord imposé par la taille du document. De la même façon que nous avons adapté sémantiquement le modèle de proximité floue, nous présentons dans la formule 8, le modèle augmenté du modèle de l'onde d'information. Les modèles fondés sur le principe de l'onde d'information peuvent être appliqués dans des contextes quelconques (structurés ou non-structurés). Dans le cas des textes structurés une agrégation est toujours nécessaire.

$$\zeta s_t^d(x) = \sum_{i \in d^{-1}(Sem(t))} \left(max\left(Sim(t_i, t)\left(1 - \frac{|x-i|}{|d^{-1}|}\right), 0\right)\right) \quad [8]$$

3.4. Choix de la ressource sémantique

Pour la gestion de la sémantique des textes des plaintes, nous avons choisi d'utiliser un dictionnaire généraliste des synonymes de la langue française DICTIONNAIRE. Ce choix s'est imposé d'une part par rapport au fait qu'aucune ressource terminologique française du domaine de la pollution de l'air et du bâtiment

n'existe à ce jour et d'autre part par rapport à la nature hétérogène du corpus. En effet, la figure 3 témoigne de l'évolution du nombre de mots nouveaux au sein d'un corpus de 655 plaintes. Par définition, un dictionnaire est la ressource la plus exhaustive possible tant au niveau des termes à considérer qu'au niveau des liens possibles entre mots lorsqu'il s'agit des dictionnaires des synonymes. Dans DICTIONNAIRE, la distance entre deux vedettes est calculée en fonction du nombre de leurs synonymes communs (Manguin, 2005). Cette technique correspond plus exactement à la distance de Jaccard traditionnellement utilisée pour évaluer le taux de similarité entre échantillons. Nous avons par ailleurs utilisé une heuristique des codes appliquée à la langue française. En effet, les différentes flexions des termes de la langue française partagent le même sens. Par cette méthode, le texte est formé uniquement de racines ou plus exactement selon la théorie d'Enguehard de concepts primitifs (Enguehard, 1992). Un concept est déterminé à partir de la forme canonique correspondant à un terme. C'est-à-dire, un concept ne contient que la sous-chaîne des caractères rassemblant les premières lettres qui le composent jusqu'à l'obtention de deux voyelles non consécutives. La traduction des termes par des codes ne se fait pas sans heurts, surtout dans le cadre d'une langue aussi riche par ses variations morphologiques que le français. Néanmoins, nous avons choisi d'implémenter le principe de racinisation d'Enguehard, d'une part par rapport à l'aisance de sa mise en œuvre et d'autre part par rapport à la réussite de son implémentation dans d'autres applications (Serradura, 2002).

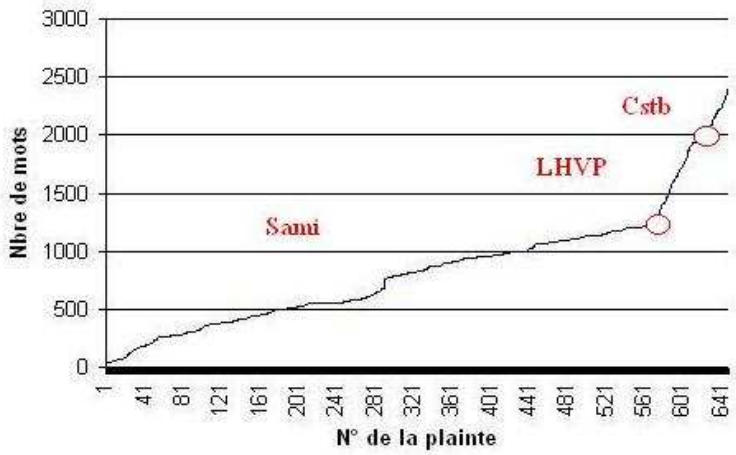


Figure 3. Évolution du nombre de mots différents (connus par TreeTagger) en fonction du nombre des plaintes analysées

Pour estimer la similarité entre deux termes ayant une même racine, nous avons réalisé un échange de synonymes entre les vedettes de DICTIONNAIRE

correspondant à ces termes avec une influence de ½ du synonyme provenant du second terme. Par application de l'indice de Jaccard, nous avons pu généraliser et dire que le taux de similarité entre les termes de toute paire issue d'un même concept-racine est de ½. L'échange de synonymes n'est pas toujours possible dans le cadre des dictionnaires. En effet, il existe une relation de pseudo transitivité dans la synonymie, étant donné que la polysémie est un frein au lien de transitivité. Si X est synonyme (ou traduction) de Y, et que Y est synonyme (ou traduction) de Z, alors ou bien X et Z sont synonymes, ou bien Y (l'élément transitoire) est polysémique. Prenons un exemple dans DICTIONNAIRE: « fenêtre » est synonyme de « baie », « baie » est synonyme de « golf », « fenêtre » n'est pas synonyme de « golf », et pour cause; « baie » est polysémique. Dans le contexte précis de notre échange de synonymes entre termes issus d'un même concept primitif selon le théorème d'Enguehard, nous pouvons dire que la transitivité est possible dans ce cas de figure. En effet, le problème de polysémie est levé en considérant les deux termes de la même famille en tant qu'une seule racine et donc partageant le même sens.

4. Expérimentation et évaluations

4.1. Évaluation des modèles de recherche directe par approche comparative

Nous avons réalisé des courbes rappel-précision afin de mettre en confrontations les modèles de recherche ayant opéré sur un même jeu de requêtes (15 plaintes XMLisées de taille variée et d'organismes différents) sur un même corpus expérimental (100 documents). Dans le cadre de notre application, la précision est privilégiée par rapport au taux de rappel. En effet, ce constat est de mise par rapport au fait que l'assignation de solution à une plainte à traiter est effectuée en fonction de l'élément positionné en tête de liste dans le classement du modèle de recherche employé. Par conséquent, pour juger de la performance des systèmes à partir des courbes rappel-précision nous analysons les positions des courbes les unes par rapport aux autres aux premiers taux de rappel.

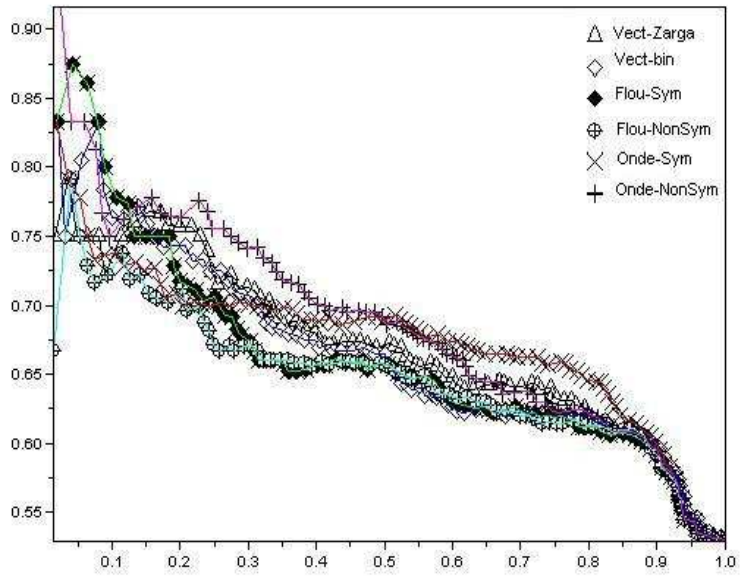


Figure 4. Évaluation de l'ensemble des modèles directs

4.2. Évaluation générale des modèles de recherche non sémantique

Selon le besoin de notre application, le modèle fondé sur l'onde d'information non symétrisé est le meilleure (figure 4). Derrière, nous remarquons le modèle de proximité floue implémenté sous sa forme symétrique. Cette expérience qui prouve l'avantage pratique du modèle de proximité floue symétrisé par rapport à sa version asymétrique témoigne de l'intérêt de notre contribution par rapport à l'application de la moyenne des valeurs réciproques du modèle de Mercier dans le cadre de notre structure. Concernant la suite du classement, et par rapport à des soucis de visibilité concernant la position des courbes les unes par rapport aux autres (figure 4), nous avons réalisé une moyenne des précisions jusqu'au taux de rappel de 5%. Le tableau 2 [4] indique le classement des modèles conformément à la moyenne des précisions au taux de rappel retenu. En résumé, nous retenons du tableau les conclusions suivantes :

- Les modèles implémentés dans le cadre de notre corpus et de notre application et pour qui l'avantage de l'intégration de la sémantique est

[4]On y trouve également le modèle Vect-bin. Ceci correspond au modèle vectoriel bidimentionnel binaire. Une dimension est à 1 si le terme correspondant existe, 0 sinon.

probante sont: le modèle vectoriel de Zargayouna et le modèle flou asymétrique.

- Par rapport au contexte non-sémantique et le contexte général ces approches ne sont pas les plus appropriées aux besoins de notre application.
- Pour les modèles les plus en accord avec le classement des experts, en l'occurrence le modèle des superpositions des ondes d'information et le modèle flou, l'intégration de la sémantique n'apporte pas une amélioration probante par rapport à leurs applications directes.

		Onde		Flou		Vect-bin	Vect-Zarga
Direct	Mode	NoSym	Sym	NonSym	Sym		
	Moyenne	0.875	0.813	0.73	0.833	0.792	0.771
	Rang	1er	5ème	12ème	3ème	7ème	9ème
Sémantique	Moyenne	0.854	0.75	0.771	0.833	0.771	0.813
	Rang	2ème	8ème	9ème	3ème	9ème	5ème

Tableau2. Classement général des modèles conformément aux moyennes de précision

4.3. Construction de la base de scénarios

Cette étape consiste à créer des classes de plaintes automatiquement en utilisant les mesures de similarité développées. Les modèles algébriques implémentés, en l'occurrence le modèle vectoriel étendu en vue bidimensionnelle direct et sémantique, utilisent DICTIONNAIRE en tant qu'espace de représentation, et ce, pour chaque rubrique renseignée. Ces modèles nous permettent d'utiliser la méthode des k-moyennes (Macqueen, 1967) pour créer des classes. Concernant les modèles de densité tels que le modèle de proximité floue et le modèle fondé sur le principe de l'onde d'information, l'algorithme est adapté puisque nous ne pouvons créer un centroïde. En effet, les points sont en réalité des textes pour les modèles flous et n'ont pas de coordonnées pour la constitution de centroïdes. Par conséquent, à chaque itération de l'algorithme des nuées dynamiques le centroïde des classes est à chaque phase l'élément le plus au centre du nuage des points (donc la plainte maximisant la somme des similarités par rapport aux plaintes du groupe considéré).

Nous avons effectué des catégorisations allant de 3 à 8 classes, et ce, au moyen des différents modèles de recherche implémentés. Trois experts du CSTB ont regroupé,

dans un nombre de classes de leur choix, les 100 éléments du corpus de tests selon les thématiques qu'ils constatent. Pour comparer les résultats des différentes classifications (tableau 3), nous avons utilisé l'indice de Rand-corrigé ((Saporta, 2004) et (Hubert, 1985)). Pour évaluer concrètement ces valeurs, nous avons besoin des résultats d'une comparaison de référence. Pour cela, nous avons appliqué l'indice de Rand-corrigé pour confronter les jugements des experts entre eux.

	Vectoriel	Vect-sém	Vect-binaire	Vect-binaire-sém
Nbre de classes	5	4	3	4
Expert N°1	0,1925	0,2962	0,3426	0.0631
Expert N°2	0,2412	0,2938	0,3953	0.1382
Expert N°3	0,2457	0,2948	**0,4539**	0.1563
	Flou-Sym	Flou-NonSym	Flou-Sém-Sym	Flou-Sém-NonSym
Nbre de classes	3	4	3	3
Expert N°1	0,1676	0,3289	0,0997	0.3831
Expert N°2	0,2887	0,3705	0,1970	**0.4325**
Expert N°3	0,2924	0,3634	0,1976	**0.4366**
	Onde-Sym	Onde-NonSym	Onde-Sém-Sym	Onde-Sém-NonSym
Nbre de classes	5	5	4	5
Expert N°1	**0,4414**	0,3701	0,3215	**0.9140**
Expert N°2	**0,4642**	0,3949	**0,4007**	**0.9264**
Expert N°3	**0,4925**	**0,4206**	**0,4396**	**0.9328**

Tableau 3. Évaluation des partitions automatiques par application de l'indice de Rand-corrigé

4.4. Comparaison entre les partitions des experts

Les comparaisons entre l'ensemble des partitions de référence et les partitions automatiques construites à l'aide du modèle de l'onde d'information sémantique dans sa version asymétrique donnent des indices de correspondance relativement élevés. À savoir que les taux d'accord entre les partitions des experts varient entre 0.5927 et

0.7717. On constate une nette domination de notre modèle de l'onde d'information sémantique non symétrisé dans le cadre de notre application avec une catégorisation à 5 classes. En analysant de plus près les classes des experts, nous pouvons dire qu'il existe au sein de notre corpus 5 scénarios, d'ailleurs indiqués le plus clairement par l'expert N°1: « Moisissures », « Fibres », « Contamination chimique », « Moisissures et acariens », « Moisissures et contamination chimique ». En effet, pour le reste des experts il est question notamment d'une classe supplémentaire notée « sans cause apparente ». Ces situations sont décrites à l'exemple de toute plainte convenable au traitement et à la prise en compte, néanmoins nos experts les ont situées dans cette classe "de rejet" tenant compte des rapports qui les accompagnaient.

4.5. Évaluation de l'assignation automatique des solutions

Évaluer notre système revient à évaluer le procédé des assignations automatiques de solutions à des plaintes écrites. À partir de 96 nouvelles plaintes non résolues, nous réalisons une affectation de solution à l'aide de notre application. L'exercice a été effectué parallèlement par 3 experts du CSTB. Le tableau 4 affiche les taux de réussite d'assignation de solutions au moyen des différents systèmes de recherche développés. Une assignation automatique est considérée erronée dans le cas où l'affectation ne correspond à l'avis d'aucun expert.

	Direct		Sémantique	
Modèle vectoriel de Zargayouna	81,93%		79,52%	
Nombre de classes	5		7	
Modèle vectoriel binaire	83,13%		78,31%	
Nombre de classes	4		8	
Modèle flou	NonSym	Sym	NonSym	Sym
	81,48%	87,95%	83,13%	89,16%
Nombre de classes	6	2	4	1
Modèle de l'onde d'information	NonSym	Sym	NonSym	Sym
	86,75%	75,90%	87,95%	74,70%
Nombres de classes	3	9	2	10

Tableau 4. Taux de réussite des assignations de solutions au moyen des modèles automatiques

Nous nous sommes basés sur le taux de désaccord entre les avis des experts pour avoir une mesure de référence par rapport aux scores du système. Les taux des

accords entre les experts considérés deux à deux vont de 88,54% à 88,75%. Nos résultats semblent globalement (toutes les méthodes) favorables à l'automatisation des réalisations des solutions aux plaintes écrites.

5. Conclusion

Dans ce travail, nous avons cherché à étudier le degré de faisabilité de l'approche automatique de résolution de plaintes écrites en français et en langue naturelle. Ces plaintes décrivent des problèmes de santé dus à la qualité de l'air au sein des ouvrages d'habitation. La principale limite de notre travail résidait dans le manque de ressources terminologiques adaptées aux différentes ingénieries connexes à la pollution domestique (santé, bâtiment, ventilation, etc.). Ces critiques nous ont motivé à dresser, un certain nombre de perspectives dont l'utilisation d'une ressource sémantique hiérarchisée nouvelle en français WOLF [5]. En effet, cette ressource permettrait probablement de mettre en évidence des liens sémantiques manquant par rapport aux liens de correspondance estimés à partir de leurs configurations synonymiques. De plus, une comparaison entre des premiers résultats d'appariement selon les mêmes modèles de recherche utilisés mais dans un contexte non structuré a démontré une amélioration nette des résultats dans ce dernier cas de figure. En effet, structurer le corpus sous forme de documents XML n'apporte pas toujours une amélioration par rapport aux documents plats. Cela est dû essentiellement à la quantité d'information-bruit "non pertinente" dans des rubriques au plus faible poids, rapprochant ainsi des éléments sans grand intérêt. Par ailleurs, la gestion de la négation est un point très important auquel on devrait s'intéresser. En effet, la non prise en compte de la négation par le module fonctionnel était dû à ses formes d'expression multiples. Ceci est à l'origine de l'inefficacité de l'intégration de la dimension sémantique qui amplifie le sens opposé des textes dans le cas où les formes de négation ne sont pas prises en compte. En effet, l'intégration de la sémantique augmente le sens des concepts visés par les textes. Dans le cas où un terme exprimé, suite à une négation, est retenu sous sa forme positive, l'utilisation de la sémantique va ramener l'ensemble des termes proches de la forme inverse exprimée à la base. De manière plus générale, la non-représentation de la négation, des paramètres numériques, des données temporelles et spatiales est une entrave inévitable à la performance des modèles fondés sur des formalisations sommaires (filtrées) de la langue naturelle.

[5]WOLF peut être téléchargée à l'adresse suivante: alpage.inria.fr/~sagot/wolf.html

6. Bibliographie

Beigbeder M., Mercier A., « Application de la logique floue à un modèle de recherche d'information basé sur la proximité », *Proceedings de la 12es rencontres francophones sur la Logique Floue et ses applications*, CEPADUES, p. 231-237, 2004.

Bellia Z. H., Modélisation et classification de textes. Application aux plaintes liées à des situations de pollution de l'air intérieur, Thèse de doctorat, Université de Paris DESCARTES, 2008.

Dinet J., « La pertinence des outils d'experts au service des non-experts en recherche d'informations: un exemple avec les opérateurs booléens », *Revue de l'EPI, n° 99*, 2000.

Enguehard C., ANA, Apprentissage Naturel Automatique d'un réseau sémantique, Thèse de doctorat, Université de Compiègne, Compiègne, France, 1992.

Hubert L., Arabie P., « Comparing Partitions », *Journal of Classification*, vol. 2, p. 193-218, 1985.

Lamontagne L., Une approche CBR textuel de réponse au courrier électronique, Thèse de doctorat, Faculté des arts et des sciences, Montréal, Canada, 2004.

Macqueen J., « Some Methods for Classification and Analysis of Multivariate Observations »,*Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, p. 281-296, 1967.

Manguin J.-L., « La dictionnairique Internet : l'exemple du dictionnaire des synonymes du CRISCO », *CORELA Cognition, Représentation, Langage, Numéro spécial*, 2005.

Salton G., « The Smart Project in Automatic Document Retrieval », *Proc. SIGIR*, ACM Press, p. 356-358, 1991.

Saporta G., Youness G., « Une méthodologie pour la comparaison de partitions », *Revue de Statistique Appliquée*, vol. 1, p. 97-120, 2004.

Schmidt H., « Probablistic Part-of-Speech Tagging Using Decision Trees », *Actes de the FirstInternational Conference on New Methods in Natural Language Processing (NemLap-94)*, Manchester, England, p. 44-49, 1994.

Serradura L., Slimane M., Vincent N., « Classification semi-automatique de documents Web à l'aide des Chaînes de Markov Cachées », *in* , F. Sèdes (ed.), *actes de INFORSID 2002, 20e congrès informatique des organisations et des systèmes d'information et de décision*, p. 215-228, juin, 2002.

Zargayouna H., Indexation sémantique de documents XML, Thèse de doctorat, Université Paris XI Orsay, 2005.

Gestion intelligente d'entrepôts de données énergétiques : quels défis ?

Lucie Copin*,,**** — Anne Laurent*,** — Hervé Rey**** — Maguelonne Teisseire*,*** — Xavier Vasques****

**LIRMM - UMR 5506 - CNRS*

{lucie.copin, laurent, teisseire}@lirmm.fr

***Université Montpellier 2*

****UMR TETIS*

maguelonne.teisseire@teledetection.fr

*****IBM France - PSSC*

{lucie.copin, reyherve, xavier.vasques}@fr.ibm.com

RÉSUMÉ. *Le projet RIDER (Réseau et Inter connectivité Des Energies classiques et Renouvelables) rassemble un consortium de laboratoires de recherche et d'industriels dans le but de concevoir des plateformes énergétiques intelligentes. Nous nous intéressons ici à la conception de modèles d'architecture d'entrepôts de données massives et hétérogènes au sein d'une plateforme intelligente temps réel de gestion énergétique multi-bâtiments*[1]*. Nous nous basons sur des données issues de capteurs (température, consommations électriques), d'applications ou de sources externes.*

Dans cet article, nous présentons les nombreux enjeux associés au projet et proposons une revue globale des travaux associés et de leurs limites dans les domaines des flux de données, du data warehousing, des processus ETL. Nous nous attachons plus précisément à ce dernier point, peu abordé jusqu'à aujourd'hui dans la littérature bien que crucial.

ABSTRACT. *The RIDER (Réseau et Inter connectivité Des Energies classiques et Renouvelables) project brings together a consortium of research laboratories and national and international companies in order to develop smart energy platforms. Here we focus on the design of data warehouse architectural models for the treatment of large heterogeneous data, needed for real-*

1. Travail soutenu par le PSSC Customer Center Montpellier, IBM France et partiellement financé par le FUI (appel à projets n°9) et par la bourse IBM "PhD Fellowship Award".

time intelligent management platforms of buildings with multiple energy sources. The data are collected from sensors (temperature, power consumption), applications or external sources. In this paper, we present the many challenges associated with this project and present a comprehensive review of related works and their limitations in the areas of data streams, data warehousing and Extract Transform Load (ETL) processes. We also focus on ETL data exceptions that have not been studied extensively in the literature.

MOTS-CLÉS : *ETL, données énergétiques, rejets, fouille de flux de données, entrepôts de données*

KEYWORDS: *ETL, energy data, data exceptions, stream mining, data warehouse*

---

1. Introduction

Le projet RIDER (Réseau et Inter connectivité Des Energies classiques et Renouvelables) rassemble un consortium de laboratoires de recherche et d'industriels dont IBM, dans le but de partager leur compétences en recherche et développement de plateformes énergétiques intelligentes.
Nous nous intéressons ici à la conception de modèles d'architecture d'entrepôts de données massives et hétérogènes au sein d'une plateforme intelligente temps réel de gestion énergétique multi-bâtiments. Nous nous basons sur des données issues de capteurs (température, consommations électriques par exemple), d'applications ou de sources externes (données météo).
Les enjeux associés à ce projet sont donc nombreux : gros volume et importance de la capitalisation des données, saisonnalité des données, fort taux d'arrivée s'apparentant aux données en flux, hétérogénéité des données (e.g. données de capteurs, données issues d'usages, préférences utilisateurs, etc.), sites monitorés de natures diverses avec différents paramètres énergétiques (eg. data center, espace résidentiel, chaîne de production), et enfin importance stratégique de permettre la prise de décisions en temps réel. Il est nécessaire de s'extraire du positionnement de l'intelligence en sortie de l'entrepôt, et de définir une nouvelle méthode qui privilégie la réactivité du système d'information (voir figure 1).

Nous nous concentrons pour répondre à ces défis sur les étapes d'alimentation de l'entrepôt, et plus précisément sur les parties transformation et chargement des données. Notre objectif est d'enrichir le procédé d'extraction des données qui génère des "rejets" c'est-à-dire des données ne respectant pas les contraintes établies et qui sont extraites sans être intégrées à l'entrepôt pour correction. Au lieu de transmettre ces rejets pour retraitement manuel et réinsertion dans le cycle comme c'est le cas aujourd'hui, nous proposons de les traiter en continu, en classifiant ces exceptions selon leur nature et en déclenchant en temps voulu les actions pertinentes correspondantes.

Dans cet article, nous présentons tout d'abord les objectifs et les problématiques du projet, puis nous détaillons l'état de l'art dans les différents domaines concernés que sont la modélisation d'entrepôts de données, la gestion des flux de données, le design et l'optimisation de processus ETL et l'extraction de règles, enfin nous décrivons notre positionnement vis-à-vis des problématiques énoncées ainsi que les pistes de recherche qui s'offrent à nous après ces travaux préliminaires.

2. Objectifs et problématique

L'objectif du projet RIDER est de concevoir un système d'information permettant d'optimiser l'efficacité énergétique d'un bâtiment ou groupe de bâtiments (incluant data centers, bâtiments tertiaires et habitats). Dans ce cadre nous utilisons des données provenant de sources variées : d'équipements matériels ou logiciels (capteurs, systèmes de gestion énergétique) des bâtiments étudiés, des utilisateurs eux-mêmes (préférences, retour sur utilisation), de données externes (e.g. données météo, données

constructeur sur les composants observés), dans un but d'analyse et d'optimisation de l'activité énergétique des bâtiments ou sites concernés.
Le contexte implique donc de gérer des données présentées sous forme de flux (e.g. données de capteurs) et sous une forme plus statique (e.g. données constructeur ou préférences utilisateurs). Dans la suite de cet article, nous utiliserons le terme "système" pour désigner le bâtiment ou le site observé ainsi que son activité énergétique.
Des systèmes de natures très différentes doivent être gérés (bâtiments résidentiels, bureaux, data centers, sites hétérogènes, bâtiments publics etc.) pour lesquels certaines règles d'efficacité énergétique seront communes et d'autres spécifiques, il faut donc connaître le contexte d'apparition des données en temps réel afin de pouvoir les traiter.

Si certaines caractéristiques énoncées semblent classiques pour des systèmes d'entrepôts de données, plusieurs questions sont associées aux difficultés inhérentes à ce contexte en opposition aux projets entrepôts classiques :

– Comment améliorer l'efficacité de l'intelligence appliquée sur les données ?
Ce type d'application nécessite de s'extraire du modèle de positionnement de l'intelligence en sortie de l'entrepôt, et de définir une nouvelle méthode qui privilégie la réactivité du système d'information (voir figure 1).

– Comment répartir le travail pour répondre au mieux à la nécessité de réactivité en temps réel et au besoin d'une analyse approfondie du système (bâtiment ou ensemble de bâtiments) observé sur le long terme ?

– Comment mettre à jour cette intelligence selon l'évolution de ce système ?
Les données extraites doivent être analysées à deux niveaux : en temps réel d'abord, pour accéder au besoin de réactivité du système d'information sur l'activité énergétique immédiate, et a posteriori pour analyser l'activité à plus long terme, détecter des tendances et les prendre en compte si nécessaire. Le contexte de données énergétiques implique de gérer un système en évolution, la prise de décision doit donc s'adapter en temps réel à son évolution globale sans pour autant devenir instable. Une problématique importante va donc être de mettre à jour en temps réel les contraintes qui régissent le traitement des données extraites avant même leur intégration à l'entrepôt.

– Comment s'adapter à des systèmes de nature et d'échelle très différentes, à partir du même cœur d'architecture ?
Le besoin de généricité de l'architecture à concevoir est primordial, car nécessaire à l'application aux différents pilotes puis aux différents sites de mise en place du projet. Dans le contexte de RIDER, la variabilité à la fois des sources de données et de l'entrepôt associé nous oblige à définir non pas un, mais plusieurs processus ETL car l'enjeu est de s'adapter systématiquement aux données dont on pourra disposer (selon par exemple les types de capteurs disponibles sur les sites observés) pour proposer les analyses correspondantes.

Ces questions constituent les principaux verrous scientifiques. Répondre à ces questions implique d'agir sur toutes les étapes du cycle de vie des données, depuis les sources jusqu'à la fouille de données dans l'entrepôt. Un état de l'art, résumé dans les sections suivantes, a donc été réalisé afin de déterminer les points critiques néces-

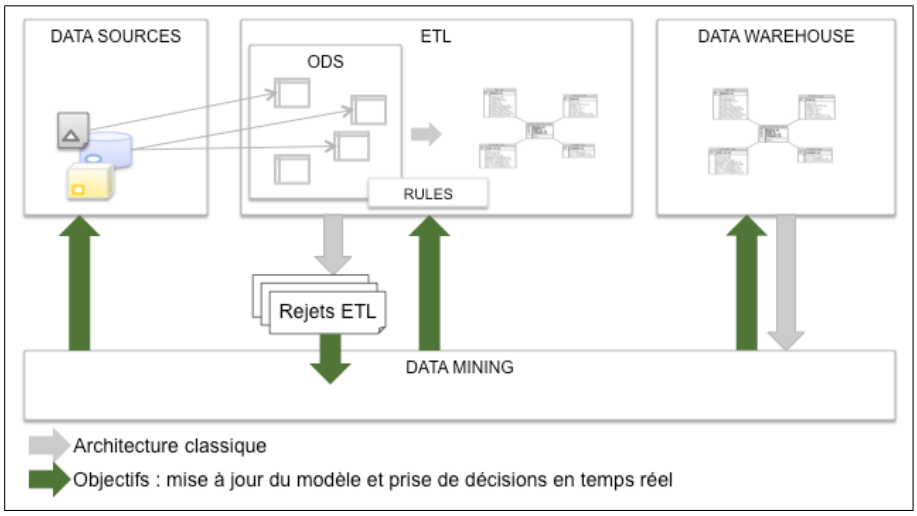


Figure 1 – Architecture classique et objectifs

sitant un travail de recherche supplémentaire. En effet certains aspects du traitement de données sont des domaines très explorés comme par exemple la fouille de données sur entrepôt ou plus récemment la fouille de flux (section 3), tandis que d'autres sont encore largement à explorer comme les aspects ETL (section 4).

3. Panorama : entrepôts, flux de données et temps réel

Pour concevoir une architecture complète de traitement des données énergétiques, nous nous interrogeons d'abord sur la nécessité d'une gestion particulière des données en terme de *data warehousing* ou entreposage. Face à des données multidimensionnelles complexes, l'architecture conçue doit permettre de disposer de données à jour ou quasi à jour à chaque instant, et intégrer aussi bien les données "statiques" (mises à jour une fois ou moins par jour) que les données "dynamiques" qui arrivent en flux doivent être intégrées pour pouvoir être analysées de façon efficace.

3.1. *Data warehousing*

(Rizzi *et al.,* 2006) présente une vue d'ensemble relativement récente de l'état de la recherche sur la conception et la modélisation d'entrepôts de données. Il souligne les éventuelles ouvertures et / ou domaines déjà explorés. Cet article souligne notamment le manque d'approches relatives au lien entre couches conceptuelle et logique et à la maintenance des processus ETL, ainsi que les possibilités non encore exploitées de conception basée sur les modèles et les ontologies. Ces pistes rejoignent de près nos

objectifs en termes de modularité et généricité. Cet article souligne aussi l'émergence d'approches *near-real time data warehousing* ou entrepôt de données en quasi-temps réel, sans toutefois en décrire les limites.

3.2. *Flux de données*

Le concept de temps réel est étroitement lié avec la notion de flux de données or à notre connaissance StreamCube (Han *et al.,* 2005) est la principale approche existante proposant une solution de construction d'un cube de données à partir d'un flux. Nous revoyons donc ici une sélection parmi les approches récentes adaptant les techniques de fouille de données existantes aux flots, sans lien avec la construction d'entrepôts de données.
Plusieurs contraintes inhérentes à la nature des flux sont à gérer dans notre contexte : un flux est potentiellement infini, on ne peut donc pas le stocker ; chaque élément du flux doit être traité en une passe ; le temps de traitement d'un élément doit être inférieur au temps d'arrivée de l'élément suivant ; le traitement des flux de données multidimensionnelles constitue encore aujourd'hui un challenge supplémentaire.
On retrouve une technique de résumé spatial et temporel adapté aux flux dans (Chiky, 2009), des techniques de construction de clusters sont proposées dans CLUSTREAM (Aggarwal *et al.,* 2003) et (Beringer *et al.,* 2005). Si CLUSTREAM permet de prendre en compte l'évolution des données en autorisant l'exploration de différentes parties d'un même flot, l'objectif de (Beringer *et al.,* 2005) est en revanche de grouper les flots de données dont l'évolution est similaire et synchrone. *On Demand Classification* (Aggarwal *et al.,* 2004) construit une classification sur les éléments d'un flot à partir de statistiques résumées et peut être utilisée entièrement en ligne. Dans (Mendes *et al.,* 2008) deux techniques d'extraction de séquences fréquentes sur flux de données sont présentées, permettant de garantir un taux d'erreur maximal et d'optimiser la gestion de la mémoire par rapport à l'existence ou non de faux négatifs.
Dans le monde académique et industriel, plusieurs outils de traitement des flux ont été développés (IBM Infosphere Streams, StreamBase, TelegraphCQ, Aleri/Coral8). Dans un premier temps il sera utile de pouvoir s'appuyer sur ces produits avant d'éventuellement affiner les besoins en termes de traitement de flux.
Le développement d'applications produisant ou gérant les données en flux et le besoin croissant d'analyse en temps réel qui les accompagne a provoqué l'émergence d'approches encore peu nombreuses pour la gestion d'entrepôts en quasi temps réel. Nous examinons SARESA (Nguyen *et al.,* 2005) qui est à notre sens la plus développée à ce jour.

3.3. *Entrepôts quasi-temps réel*

Dans *SARESA - Sense And REsponse Architecture* (Nguyen *et al.,* 2005) les données traitées sont modélisées sous forme d'événements et le traitement quasi temps-réel est mis en œuvre par division en cinq étapes ou modules traversés successivement

(annexes : figure 3). Bien que proches de nos besoins ces travaux comportent certaines limites pour le contexte concerné : la différence de charge de traitement entre des flux d'événements et des flux de données continues n'a pas été examinée mais elle pourrait présenter un coût élevé. De plus le problème du traitement d'un flot continu de données dont, à l'inverse d'événements, on ne sait pas si elle représentent ou non un changement dans le système observé et l'utilisation de la fouille de données réalisée sur l'entrepôt pour la prise de décision temps réel ne sont pas abordés.
Cependant, est née de ces observations l'intuition que pour atteindre l'objectif de prise de décisions temps réel, nous pouvons approcher l'intelligence au plus près des sources de données, en agissant sur le processus ETL.

4. Plus précisément : ETL, état de l'art et limites

L'ETL (Extract, Transform, Load) est un processus dont le but est d'intégrer des données provenant de bases opérationnelles dans un entrepôt et/ou des *data marts*. Les trois grandes phases identifiées sont :

– l'**extraction** qui consiste à récupérer les données dans une ou plusieurs bases opérationnelles et à les stocker provisoirement ;

– la **transformation** dont le but est de convertir les données ainsi stockées vers une forme respectant les contraintes appliquées sur l'entrepôt (le nettoyage des données est parfois distingué de la transformation en tant qu'étape à part entière) ;

– le **chargement** qui est l'action de transférer les données ainsi formatées vers l'entité de stockage.

Dans cette section nous décrivons l'état de l'art et ses limites sur les différents aspects des processus ETL : la modélisation, l'optimisation, les règles et enfin le domaine sur lequel nous souhaitons nous concentrer dans la suite de notre travail, les exceptions ou rejets.
Dans le contexte de la gestion de données énergétiques le processus ETL doit être adapté systématiquement au site observé et aux données qu'il peut fournir. On cherche donc à concevoir non pas un processus ETL mais une série de processus, qu'il faudra optimiser pour des raisons de coût mais aussi d'efficacité énergétique.

Nous recherchons donc dans les travaux existants les possibilités en termes d'automatisation de conception, d'optimisation et d'évolutivité des processus ETL.

4.1. *Modélisation*

A notre connaissance il n'existe pas aujourd'hui de méthode de modélisation de processus ETL standard ou générale. Des propositions ont été faites notamment dans (Vassiliadis *et al.,* 2009) qui propose une taxonomie des activités ETL.
(Lujan-Mora *et al.,* 2008) propose de décrire en UML les traitements réalisés sur chaque attribut des sources de données en se basant sur un profil spécifique. L'uti-

lisation de la norme UML adoptée très largement est un avantage, cependant modéliser des transformations complexes demande la création d'un schéma extrêmement étendu ; la méthode demandera de plus une redéfinition manuelle des transformations à chaque nouvelle génération d'un processus. (Skoutas *et al.,* 2009) utilise des techniques de web sémantique pour définir un processus ETL à partir d'une ontologie du domaine, en exploitant la théorie de transformation des graphes. (Skoutas *et al.,* 2009) et (Ambite *et al.,* 2007) sont à notre connaissance les deux propositions existantes de semi-automatisation du design d'ETL.

4.2. *Optimisation*

(Vassiliadis *et al.,* 2009) décrit l'application des design patterns aux processus ETL représentés sous forme de graphe, comme une méthode d'amélioration de l'efficacité de ces processus. Cette approche demande à être développée mais, n'étant basée sur aucun standard au départ (type UML), elle demandera un travail d'adaptation pour être conciliée avec d'autres approches. L'approche *Backward Constraint Propagation* (Liu *et al.,* 2009) propose d'optimiser un processus ETL par propagation des contraintes, qui s'appliquent sur l'entrepôt cible, vers les données sources. L'objectif est ici d'appliquer les contraintes au plus près des sources pour gagner en efficacité. L'implémentation devrait nous permettre de prendre du recul par rapport aux gains réels générés par l'optimisation réalisée.

A notre connaissance, les approches existantes proposent des solutions répondant en partie à nos besoins de semi-automatisation ou optimisation de la conception initiale d'un processus ETL, cependant aucune proposition n'a été faite sur la mise à jour du processus en temps réel à partir de l'analyse des données traitées.

Pour être capable de modifier l'action réalisée sur les données il faut agir sur la logique du traitement, c'est-à-dire sur les règles qui s'appliquent sur l'ETL.

4.3. *Règles*

Dans le cadre d'un processus ETL, une règle est constituée d'un ensemble fini d'attributs sur lesquels la contrainte est imposée et d'une transformation unique qui implémente l'application de la contrainte. Ces contraintes permettent de spécifier les actions de nettoyage, transformation, intégration, éventuellement enrichissement et calcul d'indicateurs supplémentaires sur les données. Une règle ETL dépend généralement d'une règle métier correspondante.

Prenons l'exemple d'un capteur de température dont on sait qu'il ne peut retourner qu'une valeur entre 0 et 100°C. Si ce n'est pas le cas la donnée doit être stockée dans une table particulière "ERROR" (règle métier), on peut alors en déduire l'implémentation d'une règle ETL correspondante s'appliquant sur les attributs de mesure des sources de données relatives aux capteurs de température, et dont l'opération de transformation sera un filtre appliqué sur ces données qui transfère les lignes dont les

valeurs de mesure sont inférieures à 0 ou supérieures à 100 vers la table correspondante.

A l'heure actuelle, les règles des ETL en exploitation sont en général définies par un expert (ou plusieurs). Elles sont proposées par exemple par l'administrateur de l'entrepôt de données, qui connaît les contraintes d'intégrité qui y sont associées.
(Skoutas *et al.,* 2009) propose la génération de processus ETL à partir d'une ontologie décrivant le système et fait ainsi le lien sémantique entre les règles ETL et les contraintes liées aux entités "réelles".

(Chiang *et al.,* 2008) propose de détecter des règles métier potentielles et données corrompues à partir de règles d'association fréquentes vérifiées ou "presque" vérifiées. *Exemple : si 99% des personnes mariées dont le statut marital est "époux" sont des hommes, on serait amené à rejeter la règle* `\[homme, marié -> époux]` *à cause des 1% de femmes "époux", mais on peut facilement reconnaître ici qu'il s'agit d'une erreur dans les données sources.* Une validation "manuelle" reste nécessaire pour vérifier la pertinence des règles potentielles extraites et la charge de travail dépendra alors fortement du seuil de fréquence utilisé.

(Rodic *et al.,* 2009) propose de générer une partie des règles ETL à partir des contraintes d'intégrité exprimées sur le schéma de l'entrepôt de données, et d'autre part de contrôler et d'optimiser l'éxécution de ces règles. Une limite ici est que les opérations complexes (tri, jointure) ne sont pas gérées ce qui impose une charge supplémentaire pour développer "à la main" les traitements supplémentaires.
Ces travaux nous proposent des pistes pour déterminer les règles s'appliquant sur nos données. Nous souhaitons maintenant caractériser les informations extraites du procédé ETL sur la base de ces règles avant de pouvoir appliquer sur elles des traitements, dans le but de mettre en œuvre tant la mise à jour de ces règles que la prise de décision en temps réel.

4.4. *Exceptions / Rejets*

Les règles spécifient si une donnée traitée par l'ETL est intégrée directement dans l'entrepôt, intégrée avec modification ou accompagnée d'une indication si elle est potentiellement incorrecte, ou rejetée totalement du processus.
Si nous reprenons l'exemple de notre capteur de température, la valeur extraite de celui-ci peut être :

– intégrée telle quelle dans l'entrepôt,

– modifiée (e.g. discrétisée),

– intégrée avec un message d'erreur (si valeur hors de la fourchette définie dans la règle),

– ou rejetée totalement selon la règle définie (e.g. si la valeur renvoyée est un caractère ou si la valeur ne correspond pas au comportement attendu du capteur dans le contexte).

On obtient ainsi à l'exécution du processus ETL une suite de « rejets » c'est-à-dire de données ne respectant pas les règles établies. Le traitement de ces rejets peut potentiellement nous apprendre des choses sur l'état du système ou son évolution or à l'heure actuelle ces données rejetées sont analysées manuellement pour action : modification et réintroduction dans le processus, suppression de données corrompues, détection d'anomalies sur la source des données, etc. Le domaine des rejets ou "data exceptions" dans les ETL est à notre connaissance très peu exploré aujourd'hui dans la littérature. Il est cependant très lié au domaine des règles ETL, et présente des points d'intérêt non négligeables en rapport avec les problématiques auxquelles nous sommes confrontés.
Nous explorons la possibilité d'analyser puis traiter automatiquement ces rejets, dans le but de faire évoluer les règles correspondantes de validité des données définies sur le processus ETL. Selon la finesse des règles appliquées, les rejets correspondent au minimum aux données qui ne respectent pas la syntaxe définie, et qui doivent donc être corrigées ou tout simplement éliminées ; et au maximum on peut imaginer rejeter pour analyse ou action immédiate toutes les données qui s'écartent des comportements typiques, déviances ou évènements notables.
Cette proposition présente donc plusieurs intérêts potentiels : (1) raccourcir la chaîne de prise de décision et donc améliorer la rapidité du processus, (2) adapter le système d'information par l'analyse des comportements déviants pour mise à jour des règles.

5. Positionnement et futurs travaux

Nous voulons "approcher l'intelligence" au plus près des sources de données. C'est donc à l'étape positionnée entre les sources de données et l'entrepôt que l'on va s'intéresser : la phase ETL. Nous cherchons ici comment modéliser et implémenter un processus ETL qui s'adapte aux sources de données disponibles qui peuvent varier fortement, qui permette de traiter les données "statiques" ainsi que les données en flux, et qui autorise la mise en œuvre de la réactivité temps réel.
Comme nous l'avons abordé dans la section précédente, nous proposons donc de nous concentrer plus précisément au sein du processus ETL sur le mécanisme de "data exceptions" ou rejets. Pendant le processus ETL des données sont extraites, qui ne respectent pas les règles d'insertion dans l'entrepôt fixées a priori. Ces rejets sont à l'heure actuelle retraités manuellement et réinsérés dans le cycle. Nous proposons de les traiter en continu, en classifiant ces exceptions selon leur nature, et en déclenchant en temps voulu les actions pertinentes correspondantes.
Parmi ces actions, l'analyse des comportements déviants (*e.g. la tendance d'un appareil usagé à consommer plus d'énergie dans un même contexte*) qui seraient détectés comme "erronés" dans un système d'information dont les règles sont statiques, nous permettra de détecter au plus tôt des changements dans les comportements observés et de mettre à jour automatiquement si besoin est les règles correspondantes pour les adapter aux comportements énergétiques (*e.g. au lieu de systématiquement signaler l'activité électrique de l'appareil on pourra ajuster la règle en fonction de l'âge de l'appareil étudié*). Ce mécanisme s'appliquera sur les données traversant l'ETL en

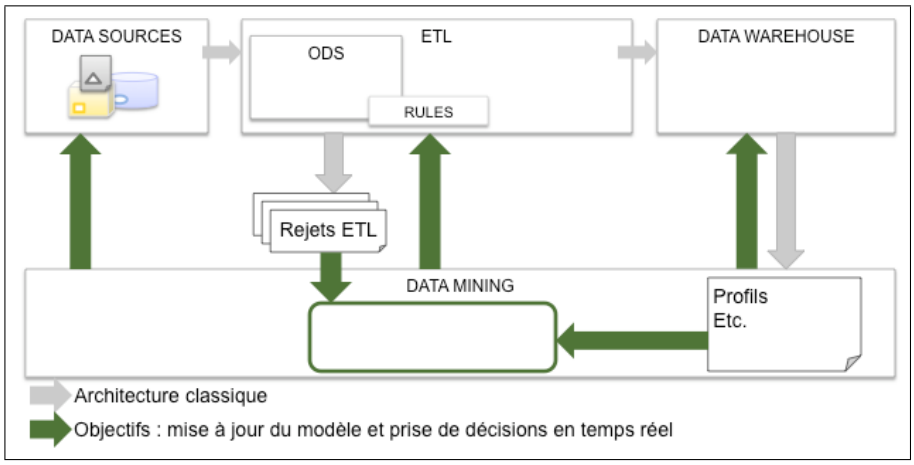


Figure 2 – Détail de l'architecture de traitement des rejets ETL

temps réel mais s'appuiera sur les résultats extraits de l'entrepôt grâce à des procédés déjà existants, notamment la recherche de profils, afin de fournir un positionnement des données entrantes par rapport au comportement historique du système (figure 2). L'architecture envisagée s'appuiera sur des flux de données énergétiques collectées, mises en forme et intégrées par le processus ETL dans une structure de *data warehousing*.
La première étape sera de caractériser ces exceptions selon le type de transformation ETL (ou règle) qui les provoque en fonction des étapes du processus : validation des données, nettoyage et transformation (annexes : tableaux 1, 2 et 3).

6. Conclusion

Nous avons dans cet article décrit les problématiques auxquelles nous faisons face dans notre démarche de conception d'une architecture d'entrepôts de données énergétiques. Plusieurs d'entre elles trouvent des réponses, totales ou partielles, dans des travaux existants sur lesquels nous pourront nous appuyer pour bâtir une architecture complète répondant à tous ces objectifs. Cependant certains aspects ne sont pas encore totalement adressés et méritent d'être approfondis dans de prochains travaux : l'hétérogénéité des données, l'adaptation du système d'information à l'évolution des comportements et l'amélioration des possibilités de prise de décision intelligente en temps réel. Nous nous consacrerons dans nos prochains travaux à fournir des solutions à ces questions, et notre attention va se porter spécifiquement sur la réactivité et la mise à jour en temps réel du processus à travers le traitement des rejets ETL, abordant ainsi un domaine encore non exploré directement par la littérature.

7. Bibliographie

Aggarwal C. C., Han J., Wang J., Yu P. S., « A Framework for Clustering Evolving Data Streams », *Proceedings of the 29th VLDB Conference, Berlin, Germany*, vol. 29, p. 81-92, 2003.

Aggarwal C. C., Han J., Yu P. S., « On Demand Classification of Data Streams », *Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining*, p. 503-508, 2004.

Ambite J. L., Kapoor D., *Automatically Composing Data Workflows with Relational Descriptions and Shim Services*, vol. Volume 4825/2007 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, p. 15-29, october, 2007.

Beringer J., Hüllermeier E., « Online Clustering of Parallel Data Streams », *Data & Knowledge Engineering*, vol. 58, p. 180-204, 2005.

Chiang F., Miller R. J., « Discovering Data Quality Rules », *PVLDB'08*, p. 1166-1177, August, 2008.

Chiky R., Résumé de flux de données distribués, PhD thesis, Telecom ParisTech, Janvier, 2009.

Han J., Chen Y., Dong G., Pei J., Wah B. W., Wang J., Cai Y., « Stream Cube : An Architecture for Multi-Dimensional Analysis of Data Streams », *Distributed and Parallel Databases*, vol. 18, n° 2, p. 173-197, September, 2005.

Liu J., Liang S., Ye D., Wei J., Huang T., « ETL Workflow Analysis and Verification Using Backward Constraint Propagation », *CAiSE '09 : Proceedings of the 21st International Conference on Advanced Information Systems Engineering*, p. 455-469, 2009.

Lujan-Mora S., Vassiliadis P., Trujillo J., « Data Mapping Diagrams for Data Warehouse Design with UML », *Proceedings of the OTM Confederated International Workshops and Posters on On the Move to Meaningful Internet Systems*, 2008.

Mendes L. F., Ding B., Han J., « Stream Sequential Pattern Mining with Precise Error Bounds », *Proc. 2008 Int. Conf. on Data Mining (ICDM'08), Pisa, Italy, Dec. 2008*, vol. 8, IEEE Computer Society, p. 941-946, December, 2008.

Nguyen T. M., Schiefer J., Tjoa A. M., « Sense & response service architecture (SARESA) : an approach towards a real-time business intelligence solution and its use for a fraud detection application », *DOLAP '05 : Proceedings of the 8th ACM international workshop on Data warehousing and OLAP*, ACM, p. 77-86, 2005.

Rizzi S., Abello A., Lechtenbörger J., Trujillo J., « Research in Data Warehouse Modeling and Design : Dead or Alive ? », *Proceedings of the 9th ACM international workshop on Data warehousing and OLAP*, p. 3-10, 2006.

Rodic J., Baranovic M., « Generating Data Quality Rules and Integration into ETL Process », *Proceeding of the ACM twelfth international workshop on Data warehousing and OLAP*, p. 65-72, 2009.

Skoutas D., Simitsis A., Sellis T., « Ontology-Driven Conceptual Design of ETL Processes Using Graph Transformations », *Journal on Data Semantics*, vol. XIII, p. 120-146, 2009.

Vassiliadis P., Simitsis A., Baikousi E., « A Taxonomy of ETL Activities », *in Proceedings of DOLAP'09*, p. 25-32, 2009.

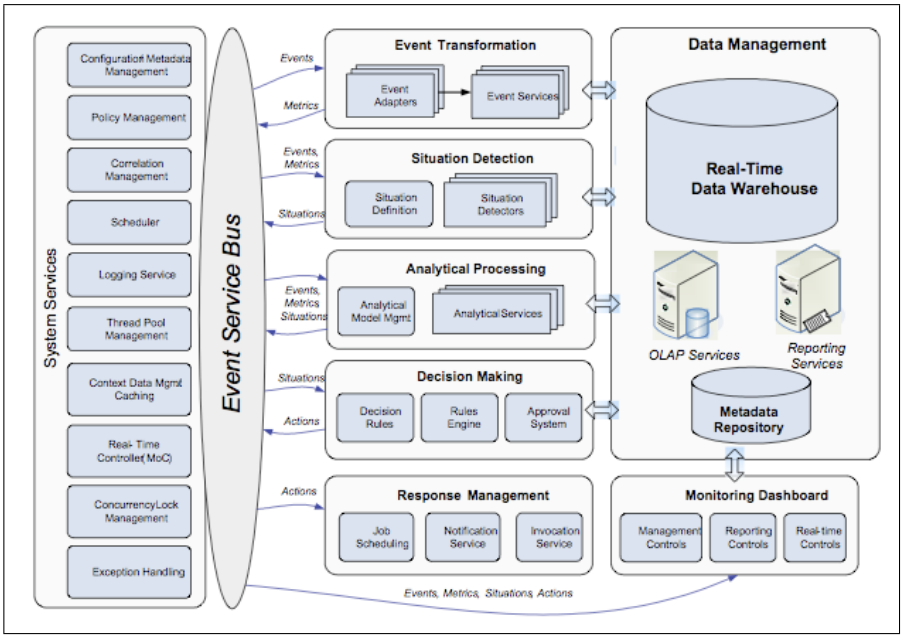


Figure 3 – Architecture SARESA

Opération	*Exceptions générées*
Vérification du bon transfert de données	Pas d'exception - renvoie une erreur de source
Vérification de la date des données	Enregistrements dont la date est incorrecte

Tableau 1 – Classification des données rejetées lors de l'étape de validation

Opération	*Exceptions générées*
Vérification du format	Enregistrement dont la valeur est hors de la fourchette définie, ou dont le format est invalide (ex : dates)
Vérification de consistance	Enregistrement sans correspondance avec les autres sources ou dans les hiérarchies définies
Vérification de complétude	Enregistrements dont certains attributs manquent (selon les contraintes spécifiées)
Vérification de la précision	Enregistrements dont les valeurs ont été évaluées (si le nombre total de données évaluées dépasse le seuil de précision fixé)

Tableau 2 – Classification des données rejetées lors de l'étape de nettoyage

Opération	*Exceptions générées*
Traduire les valeurs codées	Valeurs de code non reconnues
Encoder les valeurs	Valeurs non reconnues
Deriver les valeurs calculées	Échec du calcul
Filtre	Enregistrements filtrés (qui ne respectent pas la condition)
Tri	-
Jointure de sources multiples	Enregistrements sans correspondances
Agrégation	Enregistrements sans valeur sur la colonne d'agrégation ? (ont été nettoyés avant a priori)
Génération de clés de substitution	Erreurs sur la génération de clé
Transposition, pivot	-
Division des données	Enregistrements qui ne correspondent aux contraintes d'aucun *data* set cible
Look up et validation (Dimensions à évolution lente)	-
Validation des données (rejet nul, partiel ou total)	Données rejetées

Tableau 3 – Classification des données rejetées lors de l'étape de transformation

Edités en mai 2010 par :

Sandro Bimonte
André Miralles
François Pinet