HAL open science

Apprentissage automatique et fusion de connaissances spatio-temporelles

Robert Jeansoulin

To cite this version:

Robert Jeansoulin. Apprentissage automatique et fusion de connaissances spatio-temporelles. Entretiens Jacques-Cartier, la géomatique au service d'un environnement durable, Centre Jacques Cartier, Dec 2006, Lyon, France, France. 7pp. hal-00622321

HAL Id: hal-00622321

https://hal.science/hal-00622321v1

Submitted on 22 Feb 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Entretiens Jacques Cartier, 4-6 décembre 2006, ENS, Lyon – colloque Cemagref-CNRS-ENS

Apprentissage automatique et Fusion de connaissances spatio-temporelles

Répondre aux questions environnementales de manière durable.

RobertJeansoulin[1]

[1] CNRS Laboratoire d'Informatique Gaspard Monge (LIGM)
Université Paris Est Marne-la-Vallée (UPEMLV) - UMR8049

Résumé.
Dès que deux affirmations concernent le même objet, ou le même état du monde, on tend à les rapprocher, à évaluer leur compatibilité pour en déduire de nouvelles conséquences. Le partage de la localisation géographique implique ce besoin. Mais, les données relatives au monde réel sont porteuses d'incertitude et d'ambiguïté : elles sont observées de manière imparfaite et sont liées à des concepts souvent similaires, mais ambigus. Ceci impose de rapprocher les concepts avant de rapprocher les données, afin que le mécanisme de raisonnement tienne compte des erreurs possibles et garantisse la validité des inférences, à défaut de garantir la vérité. Nous présentons plusieurs résultats : - fusion des concepts (alignement d'ontologies), par apprentissage et analyse des concepts formels (treillis de Galois); - étude préalable des "causes possibles" par un apprentissage à base de réseaux bayesiens, sur des séries temporelles de données spatiales; - détection de conflits (incohérences logiques) entre données spatiales et leur révision pour restaurer la cohérence. L'apprentissage artificiel peut aider à traiter les situations où données numériques et informations qualitatives sont souvent mêlées.

1. Introduction.

La notion même de « durabilité » dans les questions d'environnement et d'aménagement, impose que l'ensemble des informations reconnues avérées, les hypothèses fondées et les a priori, théorisés ou non, soient partageables d'une manière rationnelle. Techniquement, le partage d'information environnementale, économique ou sociale, est devenu possible, non seulement par la technologie des réseaux, mais aussi par l'approche « données gratuites » (mais non-publiques) adoptée par les opérateurs de l'Internet, et les données publiques des services publics.
Le partage entraîne la cohabitation de fait, sur un même média, d'une quantité incroyablement volumineuse d'une information extrêmement diverse mais liée à une même sémantique, commune ou proche. Dans le cas de l'information géographique, les données sont liées à un même référent spatial, ou du moins supposé tel (déjà source d'erreurs).
Ceci a deux conséquences antagonistes :

- Il semble possible d'entreprendre des études sur des phénomènes complexes, grâce au volume et à la diversité des informations disponibles,
- Il semble vain de traiter efficacement une telle masse de données aussi désorganisées.

Donc si le partage simple d'information, -disons la *juxtaposition*-, semble à portée de main (ou de technologie), le partage rationnel, -disons la *fusion* pour suivre la terminologie de l'intelligence artificielle-, se révèle d'une complexité sémantique et calculatoire difficilement surmontable. De plus, la « durabilité » s'évalue dans le temps : une fois les informations partagées par la *fusion*, la succession temporelle sur une même localisation doit pouvoir être interprétée en terme de changement (ou non) et de causalité (ou non) : c'st à nouveau affaire d'apprentissage.

Au travers de quelques applications relatives aux évolutions environnementales ou sociétales, nous illustrons les questions auxquelles prétend répondre la fusion de connaissances, son principe général et les problèmes particuliers rencontrés.

2. Questionnements autour de la fusion de connaissances

La prise de décision est presque toujours localisée (aménagement, situation de crise), en contexte multi-acteurs et multi-critères : les bases d'un accord sont à trouver, si possible de manière rationnelle et en fonction de données acceptées en commun.
Exemples : les sondages, les enquêtes d'utilité publique, les expériences qualifiées de « ppgis » (public participation GIS).
Les raisons des échecs résident souvent dans l'usage de termes différents et les défauts d'appariement des classes et des nomenclatures ; les acteurs ne parlent pas des mêmes objets, ce qui a pour résultat que : (1) en situation inégalitaire, les moins "empowered" (en capacité de décision) se sentent exclus et/ou s'excluent ; (2) en situation plus égalitaire, les acteurs n'ont pas la même perception de ce qui peut changer, ni de l'importance du changement (attendu ou constaté).
Le problème que nous nommons fusion dans cet article, est nommé de manières diverses, mais concerne par exemple ce qui est soulevé dans un articlerécent The Future of Remote Sensing: Interoperability and Integration : 'Multi-source Data Integration is Needed', publié dans la revue professionnelle GEO Informatics Magazine, le 28-11-2006. Les industriels et les économistes de ce marché-là posent donc la fusion comme un problème non résolu et de la plus haute importance.

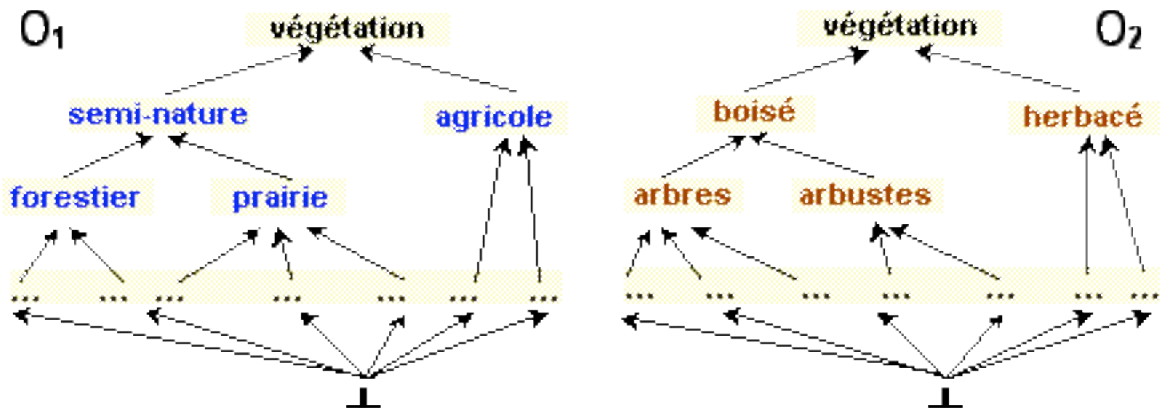
Essayons de définir l'objectif général de la fusion de connaissances.
Les trois principaux défis sont :
→Fusion des données : nombreux Modèles, beaucoup de Données, … une seule Réalité ;
→Qualité compatible : La qualité est variable, ce qui a un impact sur la fusion ;
→Associations cohérentes : surmonter les incertitudes nécessite une phase d'apprentissage.

3. Quelques problèmes de la fusion de connaissances

3.1. Ontologie / ontologies : pas une simple question de vocabulaire.

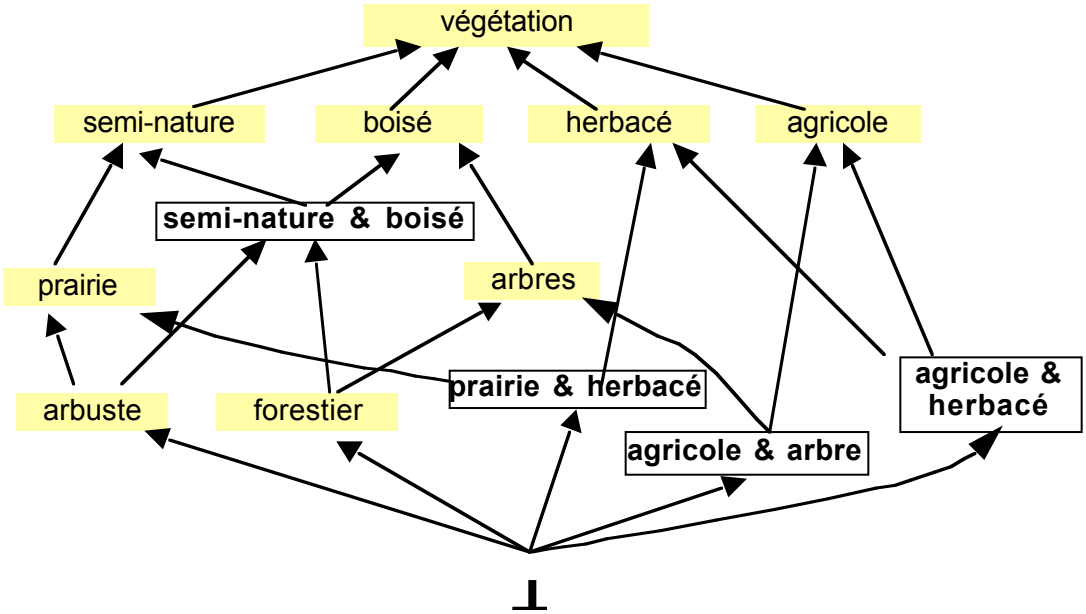
Un préliminaire sur le mot ontologie : « vieux » terme pour une « vieille » idée, réactivé parfois de manière restrictive (cf. évolution du terme sur Wikipedia dans les années 2002-2004).
Une définition de W.O. Quine (circa 1950) : parler d'ontologie c'est parler de tout ce sur quoi portent les quantificateurs : « On **what** is there » — « On what **there** is ». C'est à dire que le « substrat spatio-temporel » dans la représentation de la connaissance est aussi important que ce qu'il porte (relation d'ordre, inclusion, symétrie, etc…).

Alignement d'ontologies, Apposition de concepts, Vocabulaires différents : utilisation d'un même terme pour différentes significations, même signification mais termes différents, relations d'ordre différentes entre les termes. Les données mesurées même avec beaucoup de sérieux, sont toujours le résultats de nombreux choix intermédiaires : voir exemple Figure 1.



1. Figure 1 : d'une même réalité indifférenciée, un agronome tirera la nomenclature O_1, alors qu'un forestier tirera la nomenclature O_2, pour décrire la même végétation.

La construction de la « relation apposée » conduit à un treillis (arbre de nomenclature) unique mais qui n'informe pas sur ce qui aurait pu être commun entre les deux spécialistes. Par contre une approche « d'alignement d'ontologies » conduit à un treillis plus complexe mais plus informatif. En partant des observations qui ont conduit à des nomenclatures différentes, on peut identifier (paires de Galois), celles qui peuvent être appariées [PTT].
On aboutit ainsi à un treillis unique comme dans la Figure 2.



2. Figure 2 : dans l'ontologie « alignée », on retrouve les nomenclatures de l'agronome et du forestier mais aussi leurs relations hiérarchiques et de nouvelles « classes mixtes » qu'ils attribuent aux mêmes observations.

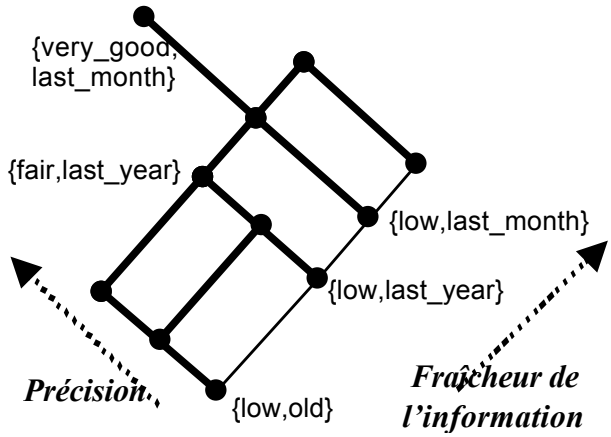
3.2. Incertitude et qualité de l'information.

Toujours dans le même exemple, on constate que diverses situations peuvent survenir :

- la *correspondance certaine*, qu'on nomme « dépendance fonctionnelle » en bases de données ; c'est le cas pour les classes mixtes si elles sont terminales, c'est à dire que leur successeur est la classe végétation dans l'exemple : ce cas n'est pas présent dans notre exemple;

- la <u>*correspondance possible*</u> lorsque la classe « A&B » a pour successeurs les deux classes « A » et « B ». Il y a 4 cas dans l'exemple précédent. On peut faire un traitement a priori de l'incertitude, si nous avons des hypothèses sur les proportions entre A et B, mais nous pouvons aussi calculer des poids en fonction, par exemple des surfaces respectives de A et B qui interviennent dans les observations qui ont conduit à construire la classe « A&B ». On peut conclure à un « mélangee de classes » si les poids sont proches, ou à une séparation s'ils sont très différents.

D'une manière générale, on peut généralement faire intervenir divers « éléments de qualité » tels que décrits dans la série de normes ISO 191xx. Ces divers éléments peuvent être dotés d'un ordre partiel qui peut aussi être représenté par un treillis.
Exemple: par numéro de parcelle, on peut avoir une idée a priori de la qualité des données :



#1 {précision: bonne, fraîcheur: 1 an, couverture: totale pour forêt, partielle sinon}
#2 {précision: bonne, fraîcheur: 1 mois, couverture: totale pour tout}
#3 {précision: moyenne, fraîcheur: 1 mois, couverture: partielle}
etc…

ces informations de qualité se combinent aux informations du treillis précédent pour aider la détermination des correspondances.

3. Figure 3 : la qualité est d'autant meilleure qu'on « monte » dans le treillis, mais certains choix sont préférables selon le type de qualité privilégié par l'application.

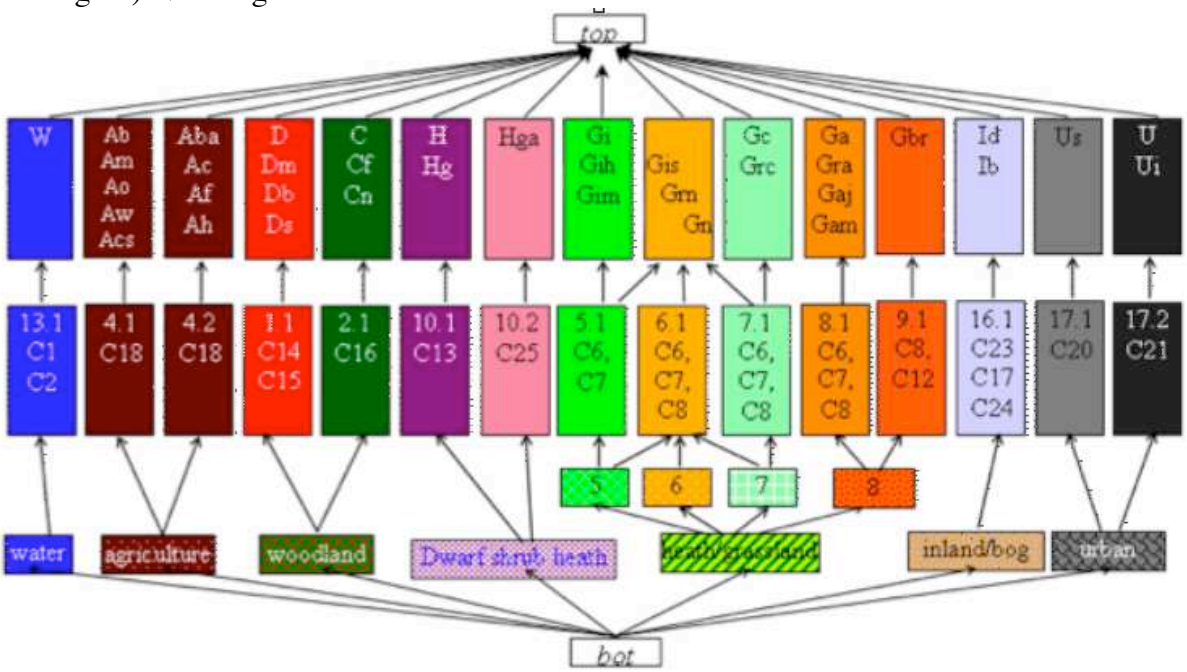
4. Fusion de données et apprentissage.

Essai de formalisation. Soit un certain espace géographique $\mathcal{X}$ et un vocabulaire $\mathcal{I}$.
Nommons "couverture" tout ensemble $\mathcal{C}(\mathcal{X})$ de parties de $\mathcal{X}$ dont l'union est égale à $\mathcal{X}$: un cas particulier est la "partition", dans laquelle les parties sont disjointes.
Nommons "nomenclature" tout sous ensemble $\mathcal{N}(\mathcal{I})$ de $\mathcal{I}$.
Un contexte formel sur $\mathcal{X}$ et $\mathcal{I}$ est un triplet composé d'une couverture, d'une nomenclature et d'un ensemble de relations binaires entre éléments de la couverture et de la nomenclature, c'est à dire sur leur produit cartésien. Notation : $\mathcal{D} = \{ \mathcal{C}(\mathcal{X}), \mathcal{N}(\mathcal{I}), R \subseteq \mathcal{X} \times \mathcal{I} \}$
Soient deux contextes formels $\mathcal{D}_1 = \{\mathcal{C}_1(\mathcal{X}), \mathcal{N}_1(\mathcal{I}), R_1\}$ et $\mathcal{D}_2 = \{\mathcal{C}_2(\mathcal{X}), \mathcal{N}_2(\mathcal{I}), R_2\}$
La fusion de $\mathcal{D}_1$ et $\mathcal{D}_2$ est un nouveau contexte formel : $\mathcal{D} = \{\mathcal{C}_{1x2}(\mathcal{X}), \mathcal{N}_{1\&2}(\mathcal{I}), R_{12}\}$.
Pour $\mathcal{C}_{1x2}(\mathcal{X})$ un choix simple est la partition construite sur l'intersection de toutes les parties de $\mathcal{C}_1(\mathcal{X})$ et $\mathcal{C}_2(\mathcal{X})$, notons-la $\mathcal{P}_{1x2}(\mathcal{X})$.
Pour $\mathcal{N}_{1\&2}(\mathcal{I})$, nous avons vu comment l'analyse de Galois peut nous aider, enfin :
$R_{12} = \{ (X',N') \mid X' \in \mathcal{P}_{1x2}(\mathcal{X}) ; N' = \min(\cup_i (N_i \mid \exists (X,N_i) \in R_1 \Phi R_2), X' \subseteq X)) \}$, il existe plusieurs choix possibles pour l'opération Φ entre R_1 et R_2 : par exemple $\cup$ et $\cap$ (ce qui est peu

efficace en général), ou toute combinaison possible guidée par le treillis de qualité: *θ-fusion*

4.1. Application « cartes de couverture du sol » (land-cover).

Couverture totale du territoire européen pour estimer les surfaces agricoles, naturelles ou urbaines, par un recensement régulier (environ 10 ans) : « land-cover mapping » LCM.
Exemple de la Grande Bretagne : avec LCM-GB (1990) + LCM2000
Difficile convergence sur une nomenclature commune et pérenne : le débat a conclu à la nécessité d'introduire de nouvelles classes en 2000. Les données sont issues de classification automatique d'images + corrections manuelles → variété spatiale de la qualité.
Alignement des deux ontologies LCMGB et LCM2000, basée sur plus de 5000 parcelles d'une même région). Voir Figure 4.



4. Figure 4 : treillis d'information commun aux ontologies des deux campagnes de LCM.

4.2. Application « réseaux causaux » (land-cover change).

Une fois les nomenclatures établies et les classes attribuées aux parcelles avec une qualité associée, la correspondance diachronique parcelle par parcelle peut être analysée en terme d'évolution : y-a-t-il des relations de cause à effet ? y-a-t-il une hiérarchie des causes ?
L'approche basée sur les « réseaux bayesiens » est bien adaptée à ce type de traitement car elle permet d'inférer des règles à partir des données, tout en acceptant aisément l'ajout de règles connues a priori, selon les objets scientifiques manipulés.

Principe du réseau Bayésien.
Soit $U = \{X_1, X_2, \ldots, X_n\}$ n variables aléatoires de domaines finis : $Val(X_i)=\{x_{i1} \ldots , x_{ip}\}$
Loi de probabilité totale [cf. *Théorème Bayes*]: $P(X_1,X_2, \ldots X_i,\ldots X_n) = \prod_{i=1,n} P(X_i / X_1, \ldots X_{i-1})$
Soit G un graphe dirigé acyclique avec X_i pour noeuds, les arcs représentant les dépendances entre les X_i : $\forall X_i$, soit $Par(X_i) \subset \{X_1, \ldots, X_n\}$ = {parents directs de X_i dans G}

Soit l'ensemble de probabilités conditionnelles

$\Theta = \{P(X_i=x_{ij} / Par(X_i)=par_{ik}) / x_{ij} \in Val(X_i), par_{ik} \in Val(Par(X_i))\}$

BN=(G, Θ), constitue le réseau bayésien qui modélise la loi de probabilité totale du domaine *U*.

Objectif : modifier nos croyances sur des probabilités conditionnelles, à partir de croyances initiales et d'un ordre partiel donné entre les variables (prédécesseurs et successeurs).

Application évolution paysage

Un paysage agricole (vallée du Gers) est étudié à deux dates : l'occupation du sol a été acquise en 1995 et 1996. Divers paramètres stables sont connus par ailleurs : type de sol (pédologie), altitude aire et indice de forme de la parcelle, rang amont-aval dans la vallée, etc.

Le traitement par les réseaux de Markov a permis d'établir des règles d'inférence et certaines règles (forçage de l'ordre partie) ont été introduite (voir Figure, partie gauche).

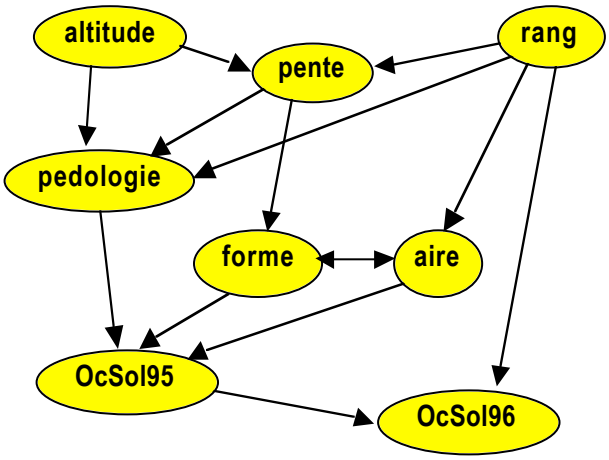
L'acquisition de données prévue pour une troisième date n'a pas été possible. Nous nous sommes contentés d'appliquer les règles construite à la première date, de comparer leur résultat avec la seconde, ce qui permet au moins de valider la portée de l'approche d'inférence basée sur un nombre limité de règles.

Introduction d'un ordre partiel explicite :

{rang, altitude}
≥ {pente, pédologie, aire, forme}

{rang, altitude, pente, pédologie, aire, forme}
> {ocSol95, ocSol96}

{ocSol95}
> {ocSol96}



5. En guise de conclusion.

L'Apprentissage automatique, incluant ce qu'on nomme "data mining" est incontournable dans les études environnementales. De nombreuses approches sont possibles, nous en avons rapidement exploré deux : une statistique (Bayes Net) ou une plus qualitative (FCA, formal concepts analysis) sans parler d'autres logiques (révision).

Une unification est-elle possible ? des tentatives sont en cours (tentatives théoriques) entre argumentation et optimisation (c'est un des chantiers actuels de l'IA). Il y a un équilibre à trouver entre "apprentissage par coeur" et "biais" trop fort.

Nous pouvons formuler quelques choix du principe d'inférence sous-jacent selon l'application :

- ***Loi d'inertie : le plus probable est ce qui maintien l'existant (Bayes)***
- ***Loi de précaution : éviter le pire cas (meilleure intersection)***
- ***La réalité est raisonable : préserver la cohérence (révision)***
- ***Lex parsimoniae d'Occam***[1]

[1] *entia non sunt multiplicanda praeter necessitatem*

Répondre aux questions environnementales de manière durable ?

A objectifs différents, stratégies différentes et principes inductifs différents :

Il y a beaucoup de bénéfices à attendre dans toutes les situations de « fusion » où l'incertitude est multiple (choix des ontologies, des règles d'association, des instruments d'observation, etc.)
Mais il reste un gros problème calculatoire ! et encore du travail de recherche…

Références bibliographiques.

Dossier Raisonnement Temporel et spatial. C Bessière, J Euzenat, R Jeansoulin, G Ligozat, S Schwer, J Revault. *Bulletin de l'Association Française pour l'Intelligence Artificielle* 29, 2-13. 1997.

Belief revision of GIS systems: the results of REV!GIS. S Benferhat, J Bennaim, R Jeansoulin, M Khelfallah, S Lagrue, O Papini. *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, 452-464. ECSQARU 2005, Barcelona. Proceedings LNCS 3571 (L. Godo Ed.).

Data fusion—from a logic perspective with a view to implementation. G Edwards, R Jeansoulin. *International Journal of Geographical Information Science* 18 (4), 303-307. 2004.

An application of problem and product ontologies for the revision of beach nourishments. D Van de Vlag, B Vasseur, A Stein, R Jeansoulin. *International Journal of Geographical Information Science* 19 (10), 1057-1072. 2005.

Révision et information spatiale. R Jeansoulin, O Papini. *Le temps, l'espace et l'evolutif en sciences du traitement de l'information,* 294-304. Editions Cepadues, 2000.

Integrating information under Lattice structure. V Phan-Luong, T TPham, R Jeansoulin. *Foundations of Intelligent Systems*, 83-87. 14th International Symposium on Methodologies for Intelligent Systems, ISMIS'03. Maebashi, Japan, 2003.

Land cover change detection: a quality-aware and semantic-based approach. R Jeansoulin, TT Pham, AJ Comber. *IPMU Conference, Special Session 04*, Perugia, Italy. 2004.

Modeling landuse changes using bayesian networks. S Benferhat, MA Cavarroc, R Jeansoulin. *22nd IASTED Intl. Conf. Artificial Intelligence and Applications*. IEEE Computer Society and Centre de Recherche Public Henri Tudor, Kirchberg, Luxembourg. 2004