SAFE-Next : Une approche systémique pour l'intégration des connaissances du domaine dans la fouille de données complexes

Walid Ben Ahmed*,***, Mounib Mekhilef*
Michel Bigand**, Yves Page***

*LGI – Laboratoire de Génie Industriel, Ecole Centrale Paris, Grande voie des Vignes 92295 Châtenay-Malabry cedex, France
{ walid, mekhilef@lgi.ecp.fr}
http://www.lgi.ecp.fr
**Équipe de Recherche en Génie Industriel, Ecole Centrale de Lille, BP 48, 59651 Villeneuve d'Ascq cedex, France
michel.bigand@ec-lille.fr
http://www.ec-lille.fr
***LAB (PSA-Renault), Laboratoire d'Accidentologie, de Biomécanique et d'études du comportement humain, 132, rue des Suisses-92000 Nanterre
yves.page@lab-france.com

Résumé. L'Extraction de Connaissances de Données (ECD) est un processus itératif dont la complexité dépend de la nature des données traitées, de la nature des connaissances à extraire ainsi que du domaine de l'application. Lorsque cette complexité est élevée, une forte implication de l'utilisateur est requise tout au long du processus et surtout dans la première et la dernière phase (i.e. préparation de données et interprétation des résultats). En combinant des approches d'ECD et d'Ingénierie de Connaissances (IC), nous développons une méthodologie descendante qui permet l'identification multi-vues des connaissances du domaine, leur formalisation sous forme de méta-données ainsi que leur incorporation dans le processus d'ECD. Notre approche est appliquée dans le domaine d'accidentologie pour l'extraction à partir des bases de données d'accidents de la route des connaissances exploitables pour le développement des systèmes de sécurité embarqués dans les véhicules.

1. Introduction

La sécurité routière est un enjeu social, politique, économique et technologique. Malgré les derniers progrès en terme de réduction du nombre de tués sur la route en France (moins 22% en 2003 par rapport à 2002), ce nombre reste élevé (5732).

L'accidentologie est un domaine dont l'objectif est l'étude des accidents pour construire des connaissances permettant d'améliorer la sécurité routière. L'un des supports de cette connaissance est le concept de scénario type d'accident (STA) qui est défini par l'INRETS[1] comme *un déroulement prototypique (prototype de déroulement) correspondant à un groupe d'accidents présentant des similitudes d'ensemble du point de vue de l'enchaînement des faits et des relations de causalité dans les différentes phases* [Brenac et Fleury, 1999]. Un exemple de scénario est donné dans la figure suivante :

[1] INRETS : Institut National de REcherche sur les Transports et leur Sécurité.

Scénario16: Concerne 33 situations accidentelles dans l'échantillon étudié (4.6%)
Cette situation accidentelle concerne un conducteur non prioritaire qui franchit un carrefour et ne voit pas le véhicule prioritaire arriver latéralement, car sa visibilité est réduite par la présence d'un masque permanent. Les conducteurs de cette catégorie roulent à une vitesse adaptée (85%), mais la fonction défaillante principale est la perception. Dans 67% des cas, l'accident a lieu en agglomération.

FIG. 1 - *Exemple de scénario type d'accident*

D'après la littérature [Brenac et Fleury, 1999] le concept scénario type d'accident est apparu pour résoudre la difficulté de synthèse dont la problématique est la suivante : *Comment passer de la compréhension de chaque accident, pris individuellement, à une vue plus générale des problèmes de sécurité sur l'échantillon d'accidents étudiés*.

Un STA, lorsqu'il est mis sous une forme suffisamment simple, peut aussi constituer un support de communication sur les accidents, dans le cadre d'actions de formation ou d'information [Brenac et Fleury, 1999]. Il est utilisé au LAB[1] comme outil de réflexion pour définir ou évaluer des *systèmes de sécurité embraqué dans le véhicule (SSE)*.

L'élaboration de ces scénarios peut être effectuée d'une manière purement experte [Brenac et Fleury, 1999] [Megherbi, 1999] [Van Elslande et Alberton, 1997]. Dans cet article, nous proposons une approche qui combine des approches automatiques d'Extraction de Connaissances de Données (ECD) (essentiellement de classification automatique) et des approches expertes.

Les bases de données d'accident que nous utilisons sont développées par le LAB en collaboration avec le CEESAR[2]. Le LAB développe des études détaillées sur la scène des accidents : une équipe de spécialistes est envoyée immédiatement sur les lieux de l'accident, qui recherche tous les indices pouvant aider à expliquer ce qui s'est passé. Cette équipe est composée d'un psychologue, d'un expert du véhicule et d'un expert de la route.

Ces informations sont ensuite stockées dans des bases de données appelées EDA (pour Études Détaillées des Accidents), constituées de données sur le *conducteur*, *l'environnement*, le *véhicule*, mais aussi des *photographies* et des *reconstructions cinématiques* et *comportementales* réalisées par des experts à l'aide de logiciels spécialisés.

Vu la complexité des données ainsi que celle des connaissances recherchées, une forte implication de l'expert dans toutes les phases d'ECD (i.e. préparation de données, datamining et interprétation des résultats) est cruciale. Ceci augmente d'une façon significative le coût de l'élaboration des STA. De plus, les STA sont à élaborer à la demande. Autrement dit, à chaque fois qu'il y a objectif d'étude (e.g. étudier la perte de contrôle, étudier la détection d'obstacles etc.), les accidentologistes appliquent un processus d'ECD avec ces différentes phases. Ceci rend l'utilisation du concept STA en accidentologie très coûteux. Pour contourner ce problème, nous proposons une méthode qui combine des approches d'Ingénierie de Connaissances (IC) et d'ECD et que nous intitulons **SAFE-Next** pour **S**ystemic **A**pproach **F**or Enhanced k**N**owledge **EXT**raction. Les objectifs de SAFE-Next sont :

- Identifier les différents types de connaissances que l'expert (i.e. l'accidentologiste) met en œuvre tout au long du processus d'ECD (essentiellement durant la première et la dernière phase du processus) pour construire des STA qui seront utilisés par des

[1] Laboratoire d'Accidentologie, de Biomécanique et d'études de comportement humain
[2] CEESAR : Centre Europe d'Etude de Risque et d'Analyse de Sécurité

concepteurs des systèmes de sécurité (utilisateurs finaux). Ainsi nous construisons des modèles multi-vue du domaine,
- Formaliser ces connaissances sous forme de *méta-données*,
- Incorporer ces connaissances dans les phases de préparation de données et d'interprétation des résultats.

Dans le deuxième paragraphe de cet article, nous présentons le fondement théorique de notre approche qui est la systémique. Nous justifions son utilisation pour l'élaboration d'un méta-modèle multi-vue du domaine. Nous menons dans cette partie aussi une réflexion sur la notion de « complexité » dans la fouille de données complexes, sur les sources de cette complexité et sur la définition de données complexes, autant de questions qui animent aujourd'hui la communauté de fouille de données. Le troisième paragraphe de cet article est consacré à un bref état de l'art sur l'élaboration de modèles du domaine pour la représentation des connaissances ainsi qu'à leur intégration dans un processus d'ECD. Nous positionnons notre approche ainsi que nos contributions par rapport à cet état de l'art. Dans le quatrième paragraphe nous présentons notre approche SAFE-Next qui consiste à construire des modèles multi-vues du domaine et leur formalisation sous forme de méta-données. Nous présentons dans le même paragraphe comment on utilise les méta-données pour l'incorporation des connaissances du domaine dans le processus d'ECD. Les avantages et les limites de SAFE-Next sont discutés dans le cinquième paragraphe.

2. Complexité, Accidentologie et fouille de données

Le fondement épistémologique de notre méthode SAFE-Next est le constructivisme qui reconnaît le caractère relatif de la connaissance et sa dépendance de la construction du sens par les individus en se basant sur leurs expériences. Un modèle dans une perspective constructiviste est défini comme une représentation de la réalité perçue par un certain nombre d'individus dans un contexte donné, d'où l'importance de la *notion de point de vue*. Le constructivisme est opposé au positivisme qui affirme un caractère stable, universel de la connaissance et indépendant du contexte.

2.1. La systémique : présentation générale

La méthodologie de modélisation que nous utilisons dans cet article dérive de l'épistémologie constructiviste et elle s'appelle la *systémique* ou la *systémographie* [Le Moigne, 1999]. C'est une approche de modélisation des systèmes complexes qui est connue aussi (surtout dans la littérature anglophone) sous le nom de la *cybernétique* [Ashby, 1965; Von Foerster, 1995] ou la *théorie de système général* [Bertalanffy, 1969; Le Moigne, 1999; Le Moigne, 1974; Morin et Le Moigne, 1999]. Son histoire remonte aux années 1940 et 1950. Von Bertalanffy, Wiener, Ashby, Von Foerster sont parmi les pères fondateurs de cette théorie.

Selon cette approche, les systèmes peuvent être classifiés en deux catégories : les systèmes complexes et les systèmes compliqués [Morin et Le Moigne, 1999] [Le Moigne, 1999]. Les systèmes compliqués sont des systèmes qui sont caractérisés par des comportements qui peuvent être prévus par l'analyse des interactions entre les composantes, ils sont déterministes (e.g. un ordinateur). Les systèmes complexes sont des systèmes non-déterministes dont les comportements ne peuvent pas être prévisibles par ce genre d'analyse. *La complexité n'est pas la complication*, nous dit Edgar Morin.

Selon la systémique, pour appréhender la complexité d'un système, quatre aspects sont à analyser conjointement : le fait qu'il *existe*, qu'il *fonctionne* et se *transforme* en même temps et

qu'il a une *téléologie* (i.e. un objectif) [Le Moigne, 1999]. Ces aspects sont détaillés dans la paragraphe 4.2 de cet article. Selon la systémique aussi, l'*observateur* (i.e. le modélisateur) et l'*objet observé* (i.e. l'objet à modéliser) ne sont plus séparés et le résultat de l'*observation* dépend de leur interaction. Cette observation est perçue aussi comme une *action téléologique* (i.e. qui a un objectif) dans un *contexte* donné [Von Foerster, 1995]. Elle affirme que toutes nos connaissances sur les systèmes sont basées sur des représentations simplifiées (i.e. des modèles) qui ignorent les propriétés du système qui ne sont pas pertinentes pour l'observateur et pour la *téléologie de l'observation*.

2.2. Accidentologie et complexité

En accidentologie, chacun des experts a sa propre perception du même phénomène qui est l'accident de la route. Cette perception varie selon le contexte de l'étude (objectif, outils utilisés etc.), selon la discipline de l'expert (biomécanique, ergonomie cognitive, mécanique, psychologie, automatique etc.), selon la spécialité de l'expert et sa tâche (diagnostic, collecte de données etc.). Même si deux experts partagent la même discipline, la même spécialité et la même tâche, leurs perceptions de l'accident peuvent être différentes. Ceci peut être vérifié dans les résultats du projet ACACIA[1] [Dieng et al., 1996] où les modèles issus de deux experts de la même discipline sont différents entre eux et sont différents de ceux des experts d'une autre spécialité.

De plus, les résultats des études accidentologiques doivent tenir compte du point de vue de leur *utilisateur final*. Le même résultat (des STA, par exemple) est présenté différemment selon que cet utilisateur est un concepteur de SSE, un concepteur d'infrastructure, une société d'assurance etc.

2.3. Accident de la route et complexité

Miller [Miller, 1995] caractérise un système complexe par un système vivant, évolutif et ouvert à son environnement avec lequel il est en interaction continue. C'est un système qui fonctionne et se transforme en même temps. Ses éléments sont reliés par des boucles de feedback (ou boucles de rétroaction). Les interactions entre ces éléments donnent à l'ensemble des propriétés que ne possèdent pas les éléments pris séparément (principe du *holisme*) et qui sont imprévisibles [Le Moigne, 1999].

Sur la base de cette définition, le comportement du système Conducteur-Véhicule-Environnement (CVE) est complexe. C'est essentiellement l'imprévisibilité des interactions entre ses différents composants qui le rend complexe. Cette impossibilité de prévision est due notamment au fait que l'action humaine est fortement impliquée (surtout à travers le conducteur) dans la production de l'accident (80% des accidents sont dus à une erreur humaine) et que ce comportement humain est, jusqu'à présent, imprévisible et non déterministe. De plus, on observe dans le comportement du système CVE des boucles de rétroaction récursives. En effet, le conducteur effectue des tâches (perception, diagnostic, action etc.) qui génèrent de nouvelles situations (nouvelles contraintes, nouvel état etc.) qui à leur tour stimulent chez le conducteur des tâches de régulation (e.g. récupérer une situation) ou de nouvelles tâches etc.

[1] ACACIA : Acquisition des Connaissances pour l'Assistance à la Conception par Interaction entre Agents

2.4. Fouille de données et complexité

Ayant défini la notion de complexité pour un système, nous allons essayer de la définir et la caractériser dans le cas d'un processus de fouille de données.

Le processus d'ECD est défini comme : « *le processus d'identification de patrons (patterns) valables, nouveaux, potentiellement utiles et compréhensibles dans les données*» [Fayyad et al., 1996]. C'est un processus interactif et itératif, impliquant de nombreuses étapes avec des décisions faites par l'utilisateur. [Brachman et Anand, 1996] soulignent la nature interactive du processus d'ECD.

L'utilisation et l'intégration des connaissances expertes du domaine dans le processus d'ECD est très importante pour découvrir de nouvelles connaissances cachées interprétables et utilisables [Palmeri et Blalock, 2000]. Cette forte implication des experts dans le processus d'ECD est l'une des sources de la complexité du processus d'ECD. En effet, les boucles de rétroaction que nous avons définies dans le paragraphe 2.1 peuvent être observées durant le processus d'ECD. Chacune des phases du processus est dépendante des phases en amont et conditionne les phases en aval. Les fonctions effectuées durant un processus d'ECD (nettoyage de données, sélection de variables, application de techniques de datamining, etc.) génèrent des transformations (transformation de données, génération de résultats etc.) pouvant conduire l'utilisateur à modifier les fonctions déjà effectuées ; ainsi, de nouvelles fonctions pourraient être définies ou redéfinies. Le processus est donc en perpétuelle interaction avec son environnement (utilisateur, objectif de l'étude etc.). La forte implication de l'opérateur humain génère une certaine imprévisibilité aussi bien au niveau de la façon de faire qu'au niveau des résultats. Cette imprévisibilité peut être accrue par :

- *La nature des données utilisées* : les aspects multi-sources, multi-domaines, multi-formes (textuelles, images, vidéo etc.), multi-niveaux de granularité sont avec la quantité parmi les éléments qui peuvent augmenter la non-répétitivité ou non-reproductibilité, donc la complexité, du processus d'ECD ;
- *La spécificité du domaine ou du phénomène étudié* : quand le phénomène étudié est complexe (e.g. accidentologie), plusieurs points de vues issus de différentes disciplines et approches sont nécessaires pour appréhender cette complexité. Ceci augmente les difficultés relatives aux conflits de points de vue (différence de raisonnements, différence de terminologies, etc.) ;
- *La nature des tâches effectuées durant les différentes phases du processus d'ECD :* si ces tâches se basent sur l'expertise dans leur application ou l'interprétation de leurs résultats, le processus d'ECD devient plus complexe ;
- *L'interdépendance des étapes d'ECD* : plus les étapes sont dépendantes, plus les feedbacks et la récursivité sont intenses et plus l'implication de l'utilisateur et donc la complexité est accrue ;
- *L'objectif de l'étude et la nature des connaissances recherchées* : de ces deux éléments dépendent les tâches suivantes : trouver un support adéquat qui représente les connaissances recherchées, réduire au maximum la perte d'information et permettre l'exploitabilité des connaissances.

3. Etat de l'art, positionnement et contributions

Cet état de l'art comporte deux parties. La première présente, sans prétendre à l'exhaustivité, des approches en Ingénierie de Connaissances (IC) d'élaboration de modèles du domaine. La deuxième partie porte sur l'intégration des connaissances du domaine (dites aussi expertes) dans un processus d'ECD. Nos contributions par rapport à la littérature sont détaillées dans ces deux parties.

3.1. Elaboration de modèle conceptuel de connaissances en IC

Le Modèle Conceptuel de connaissances (MC) est une description abstraite des connaissances indépendamment de l'implémentation. Il consiste en deux composantes : le *modèle conceptuel de raisonnement*, appelé aussi *savoir-faire* ou *Méthodes de Résolutions de Problème* (MRP) : c'est une description abstraite du raisonnement de l'expert lors de la résolution de problème. Le *modèle conceptuel du domaine*, appelé aussi savoir : il s'agit d'une représentation et structuration des connaissances du domaine étudié à travers des concepts et des relations entre ces concepts [Aussenac-Gilles et al., 1992].

Deux approches de modélisation de ces connaissances peuvent être distinguées [Duribreux-Cocquebert et Houriez, 2000] : la première est *ascendante* (*Bottom-Up*) ou *dirigée par les données* (*data-driven*) et elle se base sur une étape d'*élicitation* des connaissances des experts (entretiens, analyse de document etc.) suivie d'une étape de conceptualisation. La méthode KOD (Knowledge Oriented Design) [Vogel, 1989] et la méthode MIKE (Model-based and Incremental Knowledge Engineering) [Angele *et al.*, 1996] sont des exemples de cette famille. L'un des avantages de cette approche est qu'elle laisse la liberté aux experts d'exprimer leurs perceptions ainsi que leurs tâches sans les contraindre. De plus, les modèles élaborés correspondent aux points de vue existants. Cependant, elle présente un certain nombre d'inconvénients [Dieng *et al.*, 1996] : le coût élevé en terme de temps et d'expertise dans le processus d'élicitation ainsi que dans le processus de validation aussi bien pour les experts du domaine que pour les ingénieurs de connaissances. La seconde difficulté de cette approche est relative à la gestion des conflits de perception entre les experts aussi bien au niveau opérationnel (e.g. différentes façons d'effectuer une tâche de diagnostic de l'accident) qu'au niveau conceptuel (e.g. différents modèles de l'accident utilisés, différents modes de raisonnement etc.). Les modèles issus de cette approche souffrent aussi de manque d'abstraction, de généricité et sont donc difficilement réutilisables [Duribreux-Cocquebert et Houriez, 2000]. Nous avons soulevé d'autres inconvénients de l'approche ascendante : Nous avons constaté un risque d'incomplétude des représentations fournies par le modèle final du domaine. En effet, comme la construction du modèle est basée sur l'élicitation, si l'un des experts ayant un point de vue particulier ne participe pas à la phase d'élicitation, ce point de vue risque de ne pas apparaître dans le modèle final. Donc, les modèles élaborés sont très dépendants des experts qui ont participé à leur élaboration. De plus, si dans l'entreprise, où les modèles sont à construire, il manque des compétences ou des points de vue existant ailleurs et qui sont importants pour le domaine, ces points de vue ne peuvent pas être représentés dans les modèles élaborés par l'approche ascendante.

La deuxième famille de méthodes d'ingénierie des connaissances est *descendante* (*Top-Down*), appelée aussi approche *dirigée par les modèles* (*Model-Driven*). La méthode CommonKADS en est l'exemple le plus connu [Schreiber *et al.*, 1994] [Wielinga *et al.*, 1994]. Pour acquérir les connaissances, cette approche propose l'utilisation de *modèles génériques* [Wielinga *et al.*, 1994] préexistant, connus aussi sous le nom de *modèles squelettiques* [Motta et

al., 1990]. L'avantage de cette approche est la généricité et la réutilisabilité des modèles qui en résultent. Cependant, il est nécessaire de disposer d'une large bibliothèque de modèles génériques. De plus, l'adaptation d'un modèle existant à une application spécifique peut être une tâche difficile. Des comparaisons plus détaillées entre ces deux approches peuvent être retrouvées dans [Motta *et al.*, 1990] et [Duribreux-Cocquebert et Houriez, 2000]. Des approches mixtes ont été proposées dans [Duribreux-Cocquebert et Houriez, 2000] et [Aussenac-Gilles et al., 1992].

Aujourd'hui, il existe plusieurs modèles génériques pour guider l'élicitation des connaissances relatives à la MRP tels que les modèles de diagnostic, d'évaluation, de conception etc. La bibliothèque de CommonKADS [Breuker et Van de Velde, 1994] en présente d'autres exemples. Cependant, les travaux sur des modèles génériques d'élicitation des connaissances du domaine sont très rares. Un *modèle de connaissances du domaine* est défini selon CommonKADS [Wielinga et al., 1994] comme « *une structuration des ontologies du domaine selon un point de vue donné* ». Toutefois, ComonKADS n'offre pas d'outils d'aide à l'identification et la formalisation de ces points de vue. Elle n'offre pas non plus d'outils qui permettent de comparer des modèles préexistants pour en choisir ceux qui sont adaptés au domaine et à l'application. Nous montrons dans cet article que l'approche que nous proposons, SAFE-Next, répond à ces deux points. SAFE-Next peut être utilisée dans une approche descendante, ascendante ou mixte.

3.2. Intégration des connaissances dans le processus d'ECD

Nous nous intéressons particulièrement à l'incorporation de connaissances expertes dans la première et la dernière phase du processus d'ECD (i.e. préparation de données et interprétation des résultats). Dans la littérature, très peu de travaux existent sur des méthodologies formalisées et génériques permettant cette tâche. La variabilité et la diversité des domaines, des données, et des objectifs ont rendu difficile tout travail générique sur des méthodes d'intégration de connaissances.

3.2.1. Intégration des connaissances du domaine en préparation de données

La qualité des résultats d'un processus d'ECD dépend en grande partie de la qualité des données utilisées, d'où l'importance de l'étape de préparation de ces données [Fayyad *et al.*, 1996]. En effet, les données initiales peuvent être incomplètes, bruitées, aberrantes et incohérentes [Famili et al., 1997; Han et Kamber, 2000; Pyle, 1999]. Un processus de pré-traitement et de préparation comporte plusieurs étapes [Han et Kamber, 2000] telles que le nettoyage, l'intégration, la transformation et la réduction de ces données.

Chacune de ces tâches de préparation est effectuée selon l'objectif de l'étude, selon la nature des données (i.e. textes, images, vidéo etc.), le type de techniques de datamining à utiliser, le domaine de l'étude etc. Ceci requiert une forte implication des experts, d'où l'importance de développer des méthodes permettant d'intégrer d'une manière efficace leurs connaissances dans ces différentes tâches. Plusieurs techniques statistiques sont utilisées (e.g. filtrage de données en utilisant des estimateurs statistiques, modélisation du bruit, les méthodes de clustering, les régressions etc.), mais nous nous intéressons ici uniquement aux techniques permettant l'intégration de l'expertise.

Parmi les méthodes que nous avons identifiées dans la littérature et qui permettent l'intégration des connaissances dans la préparation des données, nous citons la représentation hiérarchique des données. En effet, cette représentation effectuée par les experts permet d'identifier et de représenter les relations entre les données [Han et Fu, 1996] [Dhar et Tuzhilin, 1993]. Les techniques de visualisation [Ahlberg et Wistrand, 1995] [Keim et Kriegel, 1996] sont aussi utilisées pour l'intégration des connaissances du domaine. Elles permettent, par exemple,

aux experts la sélection et la transformation des variables dans la phase de prétraitement des données. Nous proposons une autre technique basée sur l'utilisation de méta-données.

3.2.2. Intégration des connaissances du domaine en interprétation des résultats

L'interprétation des classes issues de l'étape datamining du processus ECD nécessite l'implication des points de vue du domaine pour pouvoir en extraire du sens et les rendre exploitables. A travers notre recherche bibliographique, nous avons constaté là encore que la majorité des travaux développent des critères statistiques qui sont orientés *validation des classes* plutôt qu'*interprétation des classes*. Une revue des critères statistiques de validation des classes peut être trouvée dans [Halkidi et al., 2002; Manco et al., 2004].

En ce qui concerne la tâche d'*interprétation des classes*, les techniques les plus utilisées sont des techniques d'exploration visuelles combinées avec les statistiques descriptives. La visualisation des résultats est généralement interactive. Parmi ces méthodes de visualisation, citons : la représentation des données multidimensionnelles par un diagramme plan [Zhang et al., 2002], le graphique matriciel (multiplot) [Grinstein et al., 2001], les techniques utilisant des icônes [Oliveira et Levkowitz, 2003]. Une revue de ces techniques peut être trouvée dans [Sachinopoulou, 2001].

Nous n'avons pas donc trouvé des travaux génériques sur des méthodes pour l'intégration de connaissances du domaine dans la phase d'interprétation des résultats issus de datamining. Nous montrons dans cet article que l'utilisation des méta-données est une technique efficace pour intégrer les connaissances expertes dans la phase d'interprétation.

4. SAFE-Next : Une approche systémique pour l'intégration des connaissances du domaine dans la fouille de données complexes

La description des accidents dans la base de données que nous utilisons s'effectue sous deux formes. La première est textuelle et consiste en des entretiens avec les impliqués dans l'accident. La deuxième est structurée et consiste en des tables *accidents x attributs*. Ces attributs (e.g. *« age du conducteur », « type d'infrastructure », « jour/nuit »,* etc.)[1] sont collectés sur le lieu de l'accident et complétés par les informations textuelles dans les entretiens. L'élaboration des STA, comme nous l'avons déjà noté, consiste à appliquer une classification automatique sur les tables pour créer des classes homogènes d'accidents et ensuite à interpréter ces classes.

Notre approche SAFE-Next se veut générique et indépendante du domaine de l'accidentologie. Elle est applicable dans le cas de classification d'*instances* (e.g. accidents, clients, incidents) décrites par des *attributs* dans une base de données. Son objectif est d'aider un *expert du domaine* (e.g. un accidentologiste, analyste en marketing, agent en maintenance) dans cette *tâche de résolution de problème* (i.e. classification) pour construire des *connaissances* (e.g. des scénarios d'accidents, des scénarios de consommation, des scénarios de défaillances) sur le *phénomène étudié* (e.g. accident de la route, comportement du client, défaillance). Ces connaissances doivent être exploitables par un *utilisateur final* du même domaine (e.g. un accidentologiste) ou d'un autre domaine (e.g. concepteur de systèmes de sécurité).

[1] La BD contient 1300 accidents caractérisés par 947 attributs. Chaque attribut contient plusieurs modalités. Exemple : *age du conducteur={jeune, adulte, vieux}*.

4.1. Architecture globale de SAFE-Next

SAFE-Next propose une architecture de modélisation conceptuelle des connaissances à plusieurs niveaux d'abstraction :

1. *Niveau 1 :* ***méta-modèle***. Ce niveau d'abstraction consiste à identifier les différents types de points de vue qui sont nécessaires pour la construction des connaissances recherchées ;
2. *Niveau 2 :* ***modèle***. Un point de vue au niveau méta-modèle est « *instrumenté* » à l'aide de plusieurs modèles du domaine. Ces modèles peuvent être préexistants dans le cas d'une approche descendante ou ils sont développés à partir des données dans une approche ascendante. Ces modèles correspondent à des points de vue existants dans le domaine ;
3. *Niveau 3 :* ***méta-données***. Il s'agit de *données sur les données* : Chacun des modèles au niveau précédent est instrumenté par plusieurs attributs. Concrètement, il s'agit d'une classification experte des attributs selon les différents modèles ;
4. *Niveau 4 :* ***attribut***. Les instances (accidents dans notre cas) dans la BD sont caractérisées par des attributs (e.g. « age du conducteur », « type d'infrastructure » etc.) ;
5. *Niveau 5 :* ***ontologie***. Une ontologie est composée de concepts et de relations entre concepts. Chaque concept est caractérisé par un ou plusieurs attributs ;
6. *Niveau 6 :* ***instance***. Ce niveau correspond à la spécification d'un cas par l'attribution de valeurs aux attributs.

4.2. Identification des points de vue du méta-modèle

Von Foerster (1995) dans son livre « *Cybernetics of Cybernetics* » (ou *cybernétique du second ordre* qu'on appelle aussi *la systémique*) montre que la modélisation d'un système doit faire la conjonction entre *l'objet observé*, *l'observateur* et le *contexte de l'observation* (cf. §2.1). Nous nous basons sur ce postulat pour construire notre méta-modèle. Nous identifions alors trois points de vue :

1. ***Point de vue objet observé*** : ce point de vue est proche du *point de vue domaine* dans CommonKADS [Wielinga *et al.*, 1994]. Cependant, l'aspect multi-points de vue dans la connaissance du domaine n'est pas considéré dans CommonKADS et c'est ce que les auteurs de cette méthode reconnaissent dans leur livre [Schreiber *et al.*, 1994] : « *Having multiple experts is a considerable risk factor. In the context of this book we cannot go into details about multi-expert situation …* ». Nous proposons alors d'enrichir ce point de vue par quatre sous-points de vue génériques, propres au phénomène étudié et qui, selon la systémique [Le Moigne, 1999], sont nécessaires pour modéliser un phénomène complexe. Ces sous-points de vue systémique permettent la définition de nouveaux concepts du domaine ainsi que de nouvelles structurations. Ce sont les sous-points de vue ontologique, fonctionnel, transformationnel et téléologique :

   (a) Sous-point de vue *ontologique* ou *structurel* (ce qu'est le système) : il représente les composants du système (conducteur, infrastructure, trafic, conditions ambiantes et véhicule), les différentes interactions entre ces composants ainsi que leurs taxonomies. Ce sous-point de vue correspond exactement au *point de vue domaine* de CommonKADS,

   (b) Sous-point de vue *fonctionnel* (ce que fait le système) : il représente le processus global du fonctionnement du système complexe Conducteur-Véhicule-Environnement (CVE),

   (c) Sous-point de vue *transformationnel et génétique* (comment évolue le système) : il décrit l'aspect dynamique, évolutionnel et génétique (dans le sens de la genèse et non celui de l'hérédité) du comportement du système CVE lors d'un accident,

(d) Sous-point de vue *téléologique* ou *motivationnel* (quels sont l'objectif et la motivation du système) : il permet l'analyse de l'accident à travers les objectifs du système CVE lors de la conduite en général et lors d'un accident en particulier.

Nous notons ici qu'en ce qui concerne les connaissances du domaine, CommonKADS offre uniquement une vue ontologique des concepts et de leurs relations à travers le *schéma du domaine*. Nous proposons donc d'enrichir la représentation du schéma du domaine en incluant des représentations fonctionnelle, transformationnelle et téléologique des relations entre les différents concepts du domaine. Cela nous permettra d'obtenir un *schéma multi-vue du domaine* que nous représentons à travers des méta-données.

2. ***Point de vue observateur*** : l'*observateur* pour nous est l'*expert* qui effectue la *méthode de résolution de problème* (MRP). Le point observateur intègre *l'objectif* de l'expert ainsi que sa MRP. En effet, selon qu'on applique du diagnostic ou de la classification automatique, les points de vue ainsi que les modèles utilisés sont différents. Ce point de vue peut être analysé en fonction des quatre points de vue systémique :
   (a) Le sous-point de vue ontologique concerne tous les concepts qui sont relatifs aux connaissances dynamiques, propres à la tâche de résolution de problème,
   (b) Le sous-point de vue transformationnel correspond à la structure d'inférences dans CommonKADS. En effet, la structure d'inférence peut être perçue comme un processus de transformation des entrées et des rôles de connaissances pour aboutir au résultat recherché. Elle exprime donc l'aspect transformationnel et génétique,
   (c) Les *méthodes de tâches* permettent la description, à travers les structures de contrôle d'exécution des inférences, le fonctionnement du raisonnement de l'expert durant le processus d'élaboration des STA. Elles peuvent donc être utilisées pour instrumenter le sous-point de vue fonctionnel,
   (d) Le sous-point de vue téléologique correspond à l'objectif de l'expert. Les *tâches* dans le modèle des *connaissances de tâche* illustrent cet axe car elles permettent de décrire et décomposer les objectifs de l'expert durant le processus d'élaboration des STA.
3. ***Point de vue contexte de l'observation*** : le contexte de l'observation intègre l'environnement dans lequel se déroule la modélisation. Ce point de vue peut aussi être analysé selon les axes systémiques :
   (a) Le *modèle d'agents*, qui décrit les caractéristiques des agents impliqués (humains, systèmes d'information et autres entités), est un exemple de modèle pouvant instrumenter le sous-point de vue ontologique du contexte. Le modèle de l'organisation (i.e. structure, processus, personnel et ressources) appartient aussi à ce sous-point de vue,
   (b) Le *modèle de communication* représentant le système d'interaction entre les agents est un exemple d'instrumentation du sous-point de vue fonctionnel du contexte,
   (c) Le *modèle de l'organisation* permet aussi l'instrumentation du sous-point de vue transformationnel. En effet, il est utilisé pour l'analyse des facettes majeures de l'organisation (structure, processus, personnel et ressources) afin de déterminer la faisabilité des solutions Systèmes à Base de Connaissance (SBC) et d'évaluer leur impact sur l'organisation,
   (d) Le *modèle des tâches* qui permet d'identifier et d'analyser les tâches globales de l'organisation est un exemple de modèle permettant l'instrumentation du sous-point de vue téléologique du contexte.

4.3. Instrumentation des points de vue

L'instrumentation des trois points de vue présentés dans le paragraphe précédent consiste à identifier et élaborer des modèles correspondant à chacun de ces points de vue. Comme le suggère la systémique, ces points de vue sont dépendants et leur instrumentation se fait dans un processus itératif.

Pour l'instrumentation du *point de vue observateur*, nous avons choisi le *modèle de la connaissance de la tâche* qui, selon CommonKADS [Wielinga *et al.*, 1994], détermine les objectifs de l'application et les moyens pour les réaliser. Nous avons utilisé le modèle générique de la tâche de classification automatique (i.e. préparation de donnée, application d'un algorithme de classification et interprétation des résultats). Trois études d'élaboration des STA ont été effectuées séparément et ont été suivies d'entretiens avec les experts. Cela nous a permis de représenter les différentes *tâches* et *inférences* faites par les accidentologistes lors de l'élaboration de STA ainsi que l'identification des différents modèles du domaine qu'ils utilisent implicitement ou explicitement au cours de la tâche d'élaboration de STA.

En ce qui concerne l'instrumentation du *point de vue contexte de l'observation*, les modèles d'*agent*, d'*organisation*, de *communication* et de la *tâche* définis dans CommonKADS [Wielinga *et al.*, 1994] peuvent être utilisés. Parmi ces modèles, c'est le modèle de l'agent « *concepteur* » qui est pertinent pour notre application. Rappelons que le concepteur est l'*utilisateur final* des connaissances que nous allons construire (i.e. les STA). Deux modèles de représentation de l'accident chez les concepteurs des SSE ont été élaborés. La présentation de ces modèles ainsi que celui de la connaissance de tâche sort du cadre de cet article.

L'instrumentation des deux points de vue *Observateur* et *contexte* nous a permis d'identifier les connaissances du domaine nécessaires pour notre application. Ceci nous a donc permis d'instrumenter le point de vue *objet observé* (i.e. ontologique, fonctionnel, transformationnel et téléologique). Concrètement, nous nous sommes basés sur ces deux points de vue pour sélectionner des modèles de représentation de l'accident parmi les modèles qui existent déjà dans la littérature [Brenac, 1997; Ferrandez et al., 1996; Kurucz et al., 1977; Van Elslande et Alberton, 1997]. Nous nous sommes basés aussi sur ces deux points de vue pour développer de nouveaux modèles. Les modèles que nous avons choisis ainsi que ceux que nous avons développés sont donc ceux dont l'expert a besoin lors de la tâche d'élaboration des STA et ceux qui répondent aux exigences de l'*utilisateur final*. Ces différents modèles ont été assignés à l'un des sous-points de vue systémiques.

La figure Fig. *2* illustre la présentation des niveaux 1 et 2 de SAFE-Next : identification des points de vue et leur instrumentation. *Ce méta-modèle intègre trois types de points de vue :* objet observé (ontologique, fonctionnel, transformationnel et téléologique), observateur et contexte. Le point de vue observateur *a été instrumenté par le* modèle de la connaissance de la tâche *d'élaboration des STA*. *Le* point de vue contexte *a été instrumenté par des modèles issus de* l'agent concepteur (utilisateur final). *Les modèles choisis pour l'instrumentation des points de vue systémiques tiennent compte des deux autres points de vue.*

Pour faciliter la compréhension de la suite de l'article, nous avons choisi de présenter deux exemples de modèles pour l'instrumentation des sous-points de vue fonctionnel et transformationnel, que nous appelons durant la suite de l'article points de vue fonctionnel et transformationnel. Plus de détails sur les différents modèles sont présentés dans [Ben Ahmed *et al.*, 2003b].

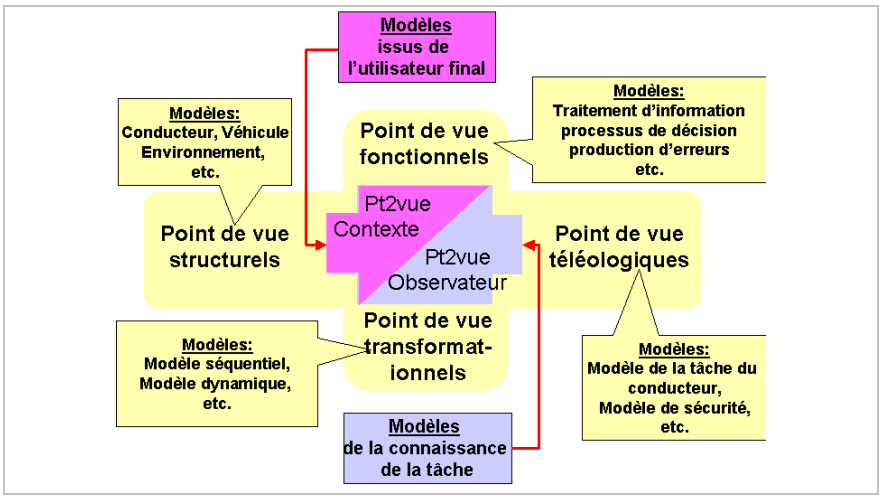

FIG. 2 –*Le méta-modèle dans SAFE-Next.*

4.3.1. Instrumentation du point de vue fonctionnel

Le modèle que nous avons choisi pour instrumenter le point de vue fonctionnel et qui est utilisé par les experts dans l'élaboration de STA est basé sur le modèle de traitement d'information de Rasmussen [Rasmussen, 1986] adapté à la tâche de conduite par Van Elslande dans [Van Elslande et Alberton, 1997].

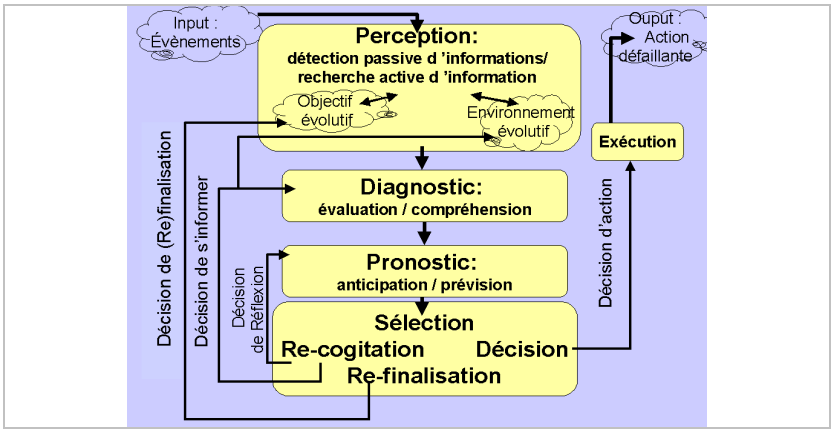

FIG. 3 – *Un modèle fonctionnel de l'accident (adapté de [Van Elslande et Alberton, 1997])*

Ce modèle décrit le traitement d'information par le conducteur selon les tâches suivantes : (i) *Etape perceptive*, caractérisée par les fonctions de recherche et de détection des informations, (ii) *Etape de diagnostic*, caractérisée par une fonction de compréhension qui concerne l'interprétation des données détectées, (iii) *Etapes de pronostic*, caractérisée par

une fonction d'anticipation qui concerne l'évolution attendue d'une situation déjà identifiée et une fonction de prévision qui renvoie aux attentes développées sur la rencontre d'un événement non encore présent dans la scène visuelle, (iv) *Etape décisionnelle* faisant référence à la prise de décision, (v) *Etape motrice* qui correspond à la fonction d'exécution des actions décidées lors de la prise de décision.

4.3.2. Instrumentation du point de vue transformationnel

Le modèle que nous avons choisi pour instrumenter le point de vue transformationnel et qui est utilisé par les experts dans l'élaboration de STA intègre le modèle séquentiel et causal de représentation de l'accident développé par l'INRETS [Brenac, 1997] (Fig. 4).

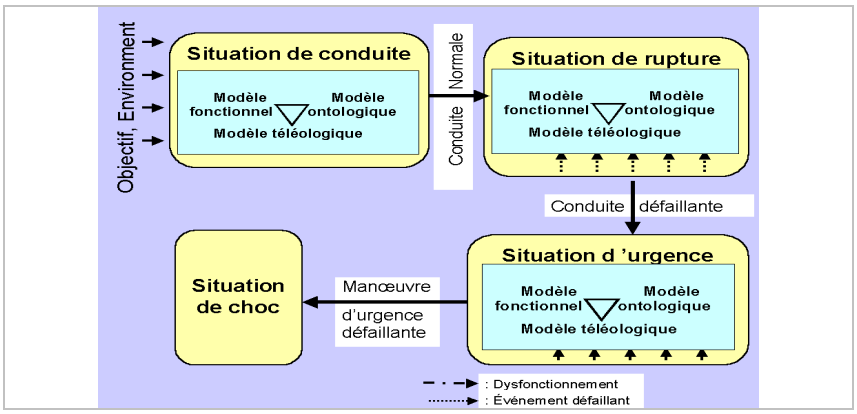

FIG. 4 - *Un modèle transformationnel de l'accident*

Ce modèle décrit l'aspect transformationnel (dynamique) de l'accident à travers les cinq phases suivantes (Fig. 4) : (i) *Situation avant la conduite* : description de l'état permanent du conducteur et du véhicule, (ii) *Situation de conduite normale* : elle correspond au comportement « normal » ou « stable » du système CVE, (iii) *Situation de rupture*: il s'agit d'une rupture qui se produit par rapport à la situation précédente caractérisée par un événement qui fait basculer le système CVE vers une situation d'urgence, (iv) *Situation d'urgence* : c'est une situation dégradée dans laquelle le conducteur doit mettre en œuvre une tentative de récupération, (v) *Situation de choc* : elle marque l'échec des manœuvres d'évitement entreprises dans la situation d'urgence.

4.4. Elaboration des méta-données : instrumentation des modèles par les attributs

Les méta-données, comme leur nom l'indique, sont *des données sur les données*. Notre approche pour la construction des ces méta-données consiste concrètement à prendre, un par un, les attributs (947) caractérisant un accident dans les BD et les assigner aux composantes des différents modèles identifiés dans le niveau 2 de l'architecture conceptuelle de SAFE-Next (cf. paragraphe 4.1). Autrement dit, il s'agit de conjoindre le niveau 3 et le niveau 2 de SAFE-Next. Ce travail a été effectué par les six experts participant à l'étude (2 équipes de 3 personnes). Avec les experts, nous avons identifié 101 attributs comme attributs administratifs. Le reste (846) a été projeté sur les différents modèles.

Pour faciliter la compréhension de cette partie, nous avons choisi de travailler sur un exemple concret d'accident (une instance). Nous présentons ci-dessous une brève description textuelle de cet accident.

```
Accident n°503 : Peugeot 206 /Renault Express. L'accident a lieu le
vendredi 20 septembre 2003, vers 13h40, au niveau du rond point à
l'intersection de la RN 25 et de la RN 29, hors agglomération de
Longueau. Le temps est ensoleillé. La chaussée est sèche. Mr X., âgé de
51 ans est électrotechnicien. Au volant de sa Peugeot 206, il vient de
quitter son travail, à la zone industrielle et se dirige vers Paris
(trajet de 420 kms) afin de rentrer chez lui, comme chaque fin de
semaine. Il est précédé d'une file de voitures quand il aborde le rond
point. Il ne connaît quasiment pas ce chemin et cherche sa route en
lisant les panneaux. Il roule à environ 20 km/h quand il arrive au
niveau du rond point. Il regarde sur sa gauche furtivement et ne voit
pas de véhicule. Il passe, alors qu'arrive au même moment un Renault
Express qu'il percute sur l'arrière droit. Le conducteur de la Peugeot
206 ne réagit qu'après-coup en freinant, n'ayant pas vu L'Express. Il
était en effet préoccupé par son itinéraire. Les conducteurs sont
ceinturés. Mr X. est indemne. Mr Y. est très légèrement blessé. Les
dépistages d'alcoolémie sont négatifs.
```

FIG. 5 - *Description textuelle d'un cas réel d'accident*

Nous ne représentons dans cet article que les méta-données qui correspondent aux points de vue fonctionnel et transformationnel.

4.4.1. **Méta-données transformationnelles**

La construction de ces méta-données consiste à assigner chacun des 846 attributs (e.g. « *age* », « *type_infrastructure* », « *conditions_ambiantes* », etc.) à l'une des composantes du modèle transformationnel présenté dans le paragraphe précédent (i.e. conduite normale, rupture, urgence, choc) (Fig. 4). Cette classification a été faite manuellement par les experts selon le patron suivant (FIG 6) :

```
<Vue transformationnelle, Concept transformationnel, Attributs>
avec
« Concept transformationnelle » ∈ {Situation avant la conduite, Conduite
normale, Rupture, Urgence, Choc}.
```

FIG 6 - *Patron de l'élaboration des métadonnées transformationnelles*

L'attribut « *age* », par exemple, est assigné au concept transformationnel «*Situation avant la conduite* ». l'attribut « *perception* » est assignée au concept transformationnel « *situation de rupture* », etc. Ces classifications des attributs selon les concepts transformationnels permettent de formaliser des connaissances expertes sur ces attributs, d'où la notion de méta-données. Ces connaissances sont stockées et elles sont consultables et modifiables par les différents experts. Chaque expert peut aussi créer ses propres méta-données et accéder aux méta-données des autres.

L'instanciation des méta-données transformationnelles pour un cas d'accident consiste en la projection de cet accident sur les méta-données. En d'autres termes, il s'agit de trouver les valeurs des attributs correspondant à cet accident. Cette instanciation est effectuée automatiquement selon le patron de point de vue suivant :

```
< N° Accident, Vue transformationnelle, Concept transformationnel, Attribut,
Valeur(s)>.
```

La représentation de l'accident n°503 présenté dans la Fig. 5 selon ce patron donne le tableau présenté dans la Fig. 7.

PointDeVue	Concepts	Attributs	Valeurs
Vue_Transformationnelle	Situation_avant_conduite	Age Modele_du_vehicule Situation_Professionnelle	51 Peugeot 206 Electricien
	Situation_Conduite_normale	Type_de_linfrastructure Conditions_de_surface Manœuvre_avant_accident Vitesse_declaree	Intersection Sec aborde un rond point 20
	Situation_de_Rupture	Perception Difficulte_de_vision Type_du_masque	Defaillance Soleil Barriere anti-choc
	Situation_durgence	Action_avant_le_crash Action_apres_le_crash	Pas de reaction Freinage
	Situation_de_Choc	Premier_impact Deploiement_de_lairbag Cote_impacte	contre un vehicule Oui Frontal

FIG. 7 - *Représentation de l'accident n°503 selon le point de vue transformationnel*

Une représentation en XML de la projection de l'accident n°503 sur les méta-données transformationnelles est donnée dans la Fig. 10.

4.4.2. Méta-données fonctionnelles

La construction de ces méta-données consiste à assigner chacun des 846 attributs à l'une des composantes du modèle fonctionnel (i.e. perception, diagnostic, pronostic, décision, action). La même procédure utilisée pour la construction des méta-données transformationnelles a été utilisée pour la construction des méta-données fonctionnelles (et des autres méta-données). La classification des attributs selon ce point de vue a été faite selon le patron suivant :

```
<Vue Fonctionnelle, Concept Fonctionnel, Attributs>
avec :
« Concept fonctionnelle » ∈ {Perception, Diagnostic, Pronostique, Décision,
Action}.
```

FIG. 8 - *Patron de l'élaboration des métadonnées transformationnelles*

Certains attributs peuvent concerner plusieurs étapes fonctionnelles. Par exemple, « l'état d'alcoolémie », peut avoir une influence sur toutes les étapes fonctionnelles. Pour cela une classe « *globale* » a été définie pour regrouper ces attributs.

Une représentation automatique de chacun des accidents est effectuée selon le patron suivant :

```
<N° Accident, Vue Fonctionnelle, Concept Fonctionnel, Attribut, Valeur(s)>
```

La projection de l'accident n°503 sur les métadonnées donne la Fig. 9. Une représentation en XML de cette instanciation, semblable à la Fig. 10, est effectuée pour ce point de vue.

PointDeVue	Concepts	Attributs	Valeurs
Vue_Fonctionnelle	Etape_de_Perception	Action_perception Defaillance_perception Visibilite Difficulte_de_vision	regarde a gauche Contr.visuel_inadapte Reduite Soleil
	Etape_de_Diagnostic	Estimation_de_danger Vitesse_estimee	etat de securite 20
	Etape_de_Pronostic	Defaut_danticipation	Oui
	Etape_de_Decision	Temps_de_reaction_necessaire	2s
	Etape_dAction	Action_avant_crash Action_apres_crash	Pas de reaction Freinage

FIG. 9- *Représentation de l'accident n°503 selon le point de vue fonctionnel*

```
<?xml version="1.0"?>
<Base_de_donnees_daccidents>
<Accident>
<N_Accident> N 503 </N_Accident>
<Point2vue>
  <Nom_Point2vue> Vue_Transformationnelle </Nom_Point2vue>

  <Concept>
      <NomDuConcept>Situation_avant_la_conduite</NomDuConcept>
      <Attributs>
          <Age>51</Age><Situation_Professionnelle>Electricien</Situation_Professionn
           elle> <Modele_du_vehicule> Peugeot 206 </Modele_du_vehicule>
      </Attributs>
  </Concept>
  <Concept>
      <NomDuConcept> Situation_de_Conduite_normale </NomDuConcept>
      <Attributs>
          <Type_de_linfrastructure>
           Intersection</Type_de_linfrastructure><Conditions_de_surface>Sec
           </Conditions_de_surface><Vitesse_declaree>20</Vitesse_declaree>
      </Attributs>
  </Concept>
  <Concept>
      <NomDuConcept> Situation_de_Rupture </NomDuConcept>
      <Attributs>
          <Perception> Defaillance </Perception> <Difficulte_de_vision> Soleil
           </Difficulte_de_vision> <Type_du_masque> Barriere anti-choc
           </Type_du_masque>
      </Attributs>
  </Concept>
  <Concept>
      <NomDuConcept> Situation_durgence </NomDuConcept>
      <Attributs>
          <Action_avant_le_crash>Pas de reaction</Action_avant_le_crash>
           <Action_apres_le_crash> Freinage </Action_apres_le_crash>
      </Attributs>
  </Concept>
  <Concept>
      <NomDuConcept> Situation_de_Choc </NomDuConcept>
      <Attributs>
          <Premier_impact> contre un vehicule </Premier_impact>
           <Deploiement_de_lairbag> Oui </Deploiement_de_lairbag><Cote_impacte>
           Frontal </Cote_impacte>
      </Attributs>
  </Concept>
</Point2vue>
</Accident>
</Base_de_donnees_daccidents> >
 </Accident>
</Base_de_donnees_daccidents>
```

FIG. 10 – *Projection de l'accident n°503 sur les métadonnées transformationnelles : une représentation en XML*

4.5. Incorporation des connaissances du domaine dans le processus ECD

Les méta-données que nous avons élaborées ont pour objectif de permettre l'intégration de connaissances expertes dans la première et dernière étape du processus d'ECD. Rappelons que la technique d'ECD utilisée ici est la classification automatique.

4.5.1. Utilisation des méta-données pour la sélection des variables

En utilisant les méta-données, les utilisateurs peuvent orienter leurs sélections d'attributs pour une étude donnée selon un ou une combinaison de points de vue. Pour l'étude de problème de perception, par exemple, l'un des experts a formulé la requête suivante :

```
(Fonctionnel ∨ Transformationnel ∨ Téléologique) ∧ (Conducteur/Véhicule
∨ Conducteur/Environnement) ∧ (Situation de rupture ∨ Situation
d'urgence) ∧ (Perception ∨ Diagnostic ∨ Pronostiques)
```

Nous avons développé une interface permettant d'exprimer automatiquement cette requête dans le langage SQL. Pour l'exemple choisi, notre interface donne la requête suivante (Fig. 11) :

```
SELECT
   Attribut_tabl.(systemic_aspect),          Attribut_tabl.(Components_interaction),
   Attribut_tabl.(Functional_step), Attribut_tabl.(Transformational_step),
FROM Attribut_table
WHERE
   (((Attribut_tabl.(systemic_aspect))="Functional") OR
   ((Attribut_tabl.(systemic_aspect))="Transformational") OR
   ((Attribut_tabl.(systemic_aspect))="Teleological") AND
   ((Attribut_tabl.(Components_interaction))="Driv/Vehicl") OR
   ((Attribut_tabl.(Components_interaction))="Driv/Envir") AND
   ((Attribut_tabl.(Functional_step))="Perception") OR
   ((Attribut_tabl.(Functional_step))="Diagnostic") OR
   ((Attribut_tabl.(Functional_step))="Prognostic") AND
   ((Attribut_tabl.(Transformational_step))="Accident_situation") OR
   ((Attribut_tabl.(Transformational_step))=" Emergency_situation"));
```

FIG. 11- *Requête générée par l'interface SAFE-Next*

Ainsi, les attributs sont sélectionnées selon le choix de l'utilisateur en combinant plusieurs points de vue existant dans le domaine et qui ont été identifiés par le méta-modèle (paragraphe 4.2) et formalisés par les experts (paragraphe 4.4). La requête ci-dessus permet de réduire le nombre de variables de 947 à 67.

Pour évaluer l'apport de l'utilisation des méta-données dans la phase de sélection des variables, nous avons comparé des études qui ont été menées sans et avec leur utilisation. Parmi ces apports, rappelons les points suivants (voir [Ben Ahmed et al., 2003a] pour plus de détails) : (i) L'utilisation des méta-données nous a permis de détecter des points de vue qui ont été ignorés dans la sélection purement experte sans utilisation des modèles du domaine ; (ii) Le nombre de variables est réduit automatiquement, mais en utilisant des connaissances expertes déjà implémentées par les méta-données ; (iii) L'expert peut toujours effectuer une sélection de variables manuelle et dans ce cas là les méta-données peuvent lui servir d'outil pour vérifier la pertinences de ces variables.

4.5.2. Utilisation des méta-données pour l'interprétation des résultats de classification

La sortie de l'application d'un algorithme de classification automatique se présente généralement sous forme d'un tableau. La Fig. 12 présente le début d'un tableau contenant le résultat d'une classification que nous avons effectuée sur un échantillon de 717 accidents en utilisant 25 attributs pour obtenir 18 classes. Ce tableau contient 1771 lignes et 7 colonnes à analyser par l'expert (donc près de 12500 informations).

Libellés des variables	Modalités caractéristiques	% de la modalité dans l'échantillon	% de la modalité dans la classe	% de la classe dans la modalité	Valeur-Test
LocChoc	Hors_chaussee	26.64	96.72	30.89	12.26
typeChoc	Tonneau_renversement	21.76	78.69	30.77	9.92
Obp	Obp=sol	18.97	68.85	30.88	8.91
vehiprior2	Vehicule_seul	29.15	72.13	21.05	7.16
sitacc	prob_contrôle_ve_21#	32.50	73.77	19.31	6.79
critini	Guidance_infrastr_5#	15.62	44.26	24.11	5.51
evini	gene_exterieur/d_14#	5.72	22.95	34.15	4.68
manident	Section_courante_17#	24.83	49.18	16.85	4.19
typacc	typacc=pilotabilite	55.51	80.33	12.31	4.09
atm	atm=Clair/normal	55.79	80.33	12.25	4.04
surf	surf=Sec	62.62	85.25	11.58	3.89
typlieu	typlieu=H-Agg sur RD	47.98	70.49	12.50	3.58
fondef	fondef=Action	9.07	22.95	21.54	3.30
manident	Changement_file_16#	6.14	18.03	25.00	3.26
typdef	Realisation_inco_20#	33.61	52.46	13.28	3.04
mask	mask=Pas_de_masque	65.13	81.97	10.71	2.86
critini	Perte_contrôle_tr_8#	17.85	32.79	15.63	2.83
mecdef	mecdef=Panique	5.72	14.75	21.95	2.57
mecdef	Activite_annexe_27#	7.67	16.39	18.18	2.23
evini	Drogue_medicamen_11#	2.51	8.20	27.78	2.21

FIG. 12 - *Sortie du logiciel SPAD d'une classification des accidents*

Les attributs dans le tableau des résultats (Fig. 12) portent sur des informations sur différents aspects et différentes composantes de l'accident et ne sont pas structurées selon un point de vue donné. A l'aide de fichiers XSL nous avons permis la sélection, pour chacune des classes, des attributs discriminants et caractérisants (voir [Ben Ahmed *et al.*, 2003a] pour plus de détails) et ce en utilisant des requêtes de sélection basées sur des critères statistiques. De plus, ces fichiers XSL permettent de projeter chacune des classes sur les différents points de vue représentés par les méta-données. En effet, une même classe peut être projetée selon le point de vue ontologique, fonctionnelle, transformationnel et téléologique. Dans une étude de validation de l'apport de SAFE-Next dans l'interprétation des résultats, nous avons montré les deux points suivants [Ben Ahmed *et al.*, 2003a] : (i) L'interprétation des résultats est moins coûteuse en terme de temps car leur projection sur les différents points de vue experts ainsi que la considération des différents critères statistiques est automatisée (quelques minutes au lieu d'un temps estimé de 3 semaines) ; (ii) Lors de l'interprétation des résultats, la projection sur les méta-données offre à l'accidentologiste une représentation multi-points de vue experts grâce à la possibilité de décrire automatiquement les groupes selon les différents axes de la systémique. Ceci permet de faciliter l'extraction de sens de ces classes (FIG. 13).

	Conducteur	Véhicule	Environnement	Cond/Veh	Cond/Env	Env/Veh
Modèle Ontol.	Realisation_incor Changement_file gene_exterieur typacc=pilotabilite Activité_Annexe Drogue_medicament mecdef=Panique fondef=Action	Tonneau_renversement Obp=sol	atm=Clair/normal surf=Sec typlieu=H-Agg_ou_RD mask=Pas_de_masque Activité_Annexe Section_courante	prob_contrôle_veh surf=Sec Perte_contrôle_trans	Guidance_infrastr surf=Sec Changement_file	Obp=sol surf=Sec Vehicule_se

	Etat long terme LT	Cond. normale CT	Rupture	Urgence	Choc
Modèle Transf.		atm=Clair/normal surf=Sec typlieu=H-Agg_ou_RD Section_courante	prob_contrôle_veh Guidance_infrastr mask=Pas_de_masque Realisation_incor Vehicule_seul Changement_file Activite_annexe	mecdef=Panique	Obp=sol surf=Sec Tonneau_renversement Obp=sol Hors_chaussee

	Perception	Diagnostique	Pronostique	Décision	Action	Globale
Modèle Fonct.	mask=Pas_de_masque			mecdef=Panique	fondef=Action Realisation_incor Perte_contrôle_trans	Activite_annexe Drogue_medicament/fon

	Navigation	Guidage latéral	Guidage longit.	Controle latéral	Controle longi
Modèle Téléo.	mask=Pas_de_masque	Guidance_infrastr Section_courante	Guidance_infrastr	Perte_contrôle_trans Tonneau_renversement	

FIG. 13 - *Projection automatique des résultats de classification sur les méta-données*

5. Discussion

La représentation graphique de l'approche SAFE-Next est donnée dans la Fig. 14 :

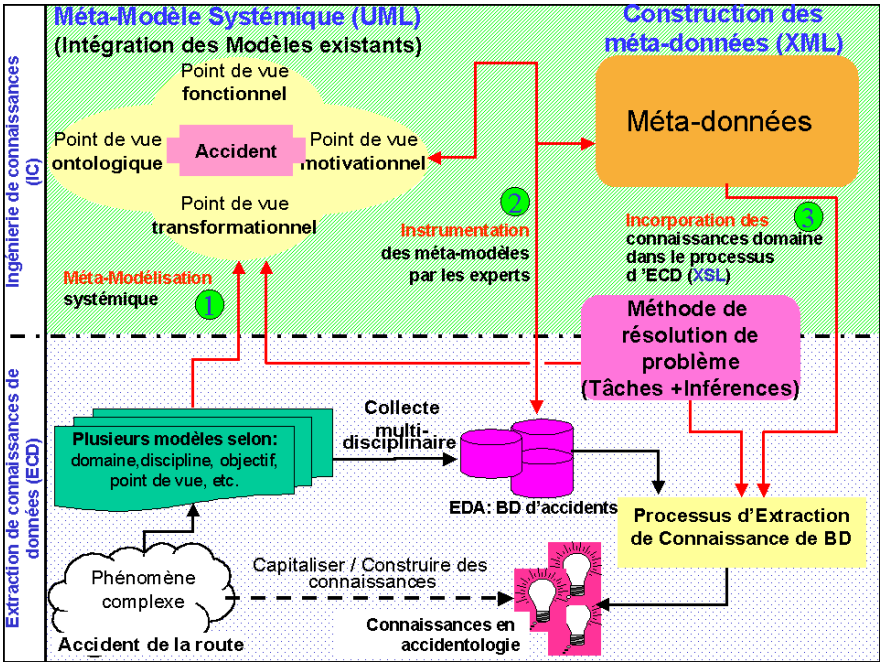

FIG. 14 - *Architecture globale de la méthode **SAFE-Next***

SAFE-Next combine une approche **IC** et une approche **ECD**. Son objectif est l'intégration des connaissances du domaine dans le processus d'ECD. La première étape consiste à construire un méta-modèle multi-vue du domaine en utilisant l'approche systémique (paragraphe 4.1 et 4.2).

Dans cette première étape, on tient compte des points de vue existants dans le domaine et du modèle de résolution de problème qui intègre les objectifs des utilisateurs du modèle du domaine à construire. La deuxième étape consiste à instrumenter le méta-modèle par les experts pour construire des méta-données (paragraphe 4.4). La troisième étape consiste à incorporer les connaissances du domaine dans le processus ECD en utilisant les méta-données (paragraphe 4.5).

5.1. Apport du méta-modèle multi-vue

Le méta-modèle multi-vue dans SAFE-Next offre les avantages suivants : (i) la structuration des points de vue utilisés dans l'entreprise et la création de liens entre ces différents points de vue ; (ii) l'identification de points de vue qui sont importants dans le domaine, mais qui ne sont pas utilisés par les experts impliqués dans l'étude.

En ce qui concerne la structuration des points de vue, celle-ci est obtenue grâce à la systémique et ses quatre axes. Structurer les différents modèles du domaine selon ces différents points de vue permet de vérifier le lien entre ces différents points et les modèles qui les représentent (ce lien est discuté dans le paragraphe suivant). Ceci permet d'éviter d'utiliser des modèles redondants et surtout d'éviter le risque d'oublier un point de vue pertinent, d'où le deuxième avantage.

En ce qui concerne ce deuxième point, il est relatif à la capacité de l'approche d'identifier de nouveaux points de vue. Ceci est assuré par l'utilisation de l'approche systémique qui commence par *modéliser le phénomène étudié lui-même (accident de la route) pour identifier les points de vue nécessaires à l'étude au lieu de modéliser la perception de ce phénomène par les experts qui participent à l'étude en se basant sur une étape d'élicitation*. Cette approche permet de réduire le risque d'incomplétude en terme de points de vue. Par exemple, le modèle de la tâche [Perron, 1997] qui analyse l'accident à travers l'objectif du conducteur (contrôle, navigation ou guidage) n'a pas été identifié dans le projet ACACIA [Dieng *et al.*, 1996] alors qu'il l'a été dans notre cas d'étude. Dans une approche ascendante, si l'un des experts ayant un point de vue particulier ne participe pas à la phase d'élicitation, ce point de vue risque de ne pas apparaître dans le modèle final. Les points de vue systémiques permettent de contourner ce problème.

5.2. Liens entre les différents points de vue

Puisque chacun des points de vue est représenté par un ou plusieurs modèles, le lien entre ces points de vue est assuré par le lien entre les différents modèles. Ce lien est d'abord sémantique car tous les modèles sont imbriqués les uns dans les autres. De plus, ce lien est formalisé grâce à l'utilisation des attributs pour instancier les modèles à l'aide de l'assignation de chacun des attributs à une ou plusieurs composantes des différents modèles. En effet, un seul attribut peut caractériser plusieurs composantes appartenant à des modèles différents. L'utilisation d'une approche basée sur les attributs permet aussi de représenter les distributions des attributs selon les différents points de vue. La Fig. 15, par exemple, représente la distribution des attributs dans la base de données en fonction des différents concepts du modèle fonctionnel (i.e. perception, diagnostic, pronostic, décision et action).

A partir de cette distribution (cf. Fig. 15) nous pouvons constater que certaines composantes du modèle fonctionnel sont sous-représentées. En effet, très peu d'attributs dans la base EDA caractérisent les phases de *diagnostic*, de *pronostique* et de *décision*. Une étude a été alors lancée au sein du LAB pour enrichir ce modèle en définissant de nouveaux attributs. L'utilisation d'une approche basée sur les attributs a permis aussi de montrer certaines divergences de perception des experts de chacun des attributs utilisée dans la BD

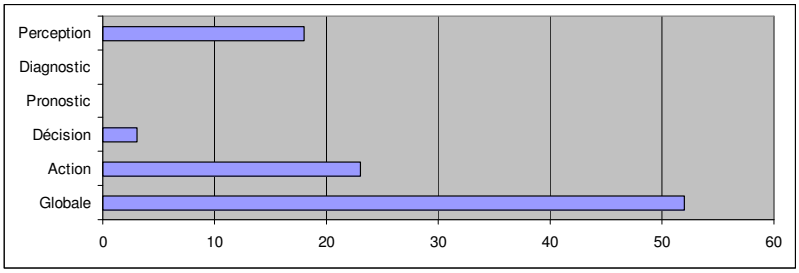

FIG. 15 - *Distribution des attributs selon le point de vue fonctionnel*

5.3. Généricité de SAFE-Next

Nous sommes en train d'étudier l'application de la méthode SAFE-Next à un autre domaine (qualité – maintenance industrielle), de manière à mettre en évidence l'indépendance de la méthode par rapport au domaine spécifique de l'accidentologie.

L'architecture du modèle des connaissances dans SAFE-Next, avec ses différents niveaux d'abstraction (méta-modèle, modèle, méta-donnée, attribut, ontologie, instance), est générique (cf. paragraphe 4.1). L'utilisation de SAFE-Next pour la modélisation des connaissances est donc indépendante du domaine de l'accidentologie.

Les différents niveaux de son architecture peuvent être utilisés séparément pour d'autres objectifs. Par exemple, les niveaux 1-2-5 peuvent être utilisés dans cet ordre dans cadre d'une approche descendante pour la construction des ontologies du domaine guidée par les différents points de vue. L'utilisation des mêmes niveaux, mais dans un ordre inverse (i.e. 5-2-1) permet dans le cadre d'une approche ascendante de structurer les ontologies et donc construire des modèles multi-vues du domaine. L'utilisation de ces niveaux dans l'ordre 1-2-5-4-6 peut être utilisée pour la définition d'attributs et la construction d'une base de données ou enrichir une BD existante par la définition de nouveaux attributs.

En ce qui concerne l'utilisation de SAFE-Next pour l'incorporation des connaissances dans le processus d'ECD, elle est particulièrement adaptée aux tâches de sélection d'attributs d'étude et d'interprétation de classes.

6. Conclusion

Dans un processus d'ECD et essentiellement lors des phases de sélection des attributs d'étude et d'interprétation des résultats (e.g. classes d'instances), l'analyste a besoin d'intégrer des connaissances du domaine. Typiquement, une classe d'accidents (ou un accident en général) peut être interprétée selon plusieurs modèles (e.g. modèle séquentiel, modèle de traitement d'information, etc.). Notre approche SAFE-Next permet d'identifier et d'organiser les différents points de vue et modèles existant dans le domaine et de les utiliser dans un processus d'ECD.

La première étape dans SAFE-Next consiste à construire un méta-modèle multi-vues du domaine identifiant et regroupant les modèles du domaine selon les quatre points de vue de la systémique (i.e. ontologique, fonctionnel, transformationnel et téléologique) et en tenant compte du point de vue MRP et du point de vue utilisateur final des connaissances construites. La deuxième étape consiste à instrumenter les modèles par les attributs qui caractérisent les instances dans la BD. Concrètement, elle consiste à classifier les attributs selon les différents concepts qui composent les modèles. L'instrumentation du modèle séquentiel, par exemple, consiste à affecter

chacun des attributs à une ou plusieurs étapes de ce modèle (i.e. « situation avant la conduite », « conduite normale », « rupture », « urgence » et « choc »). Cette classification multi-vues des attributs est effectuée par les experts. Elle permet donc de formaliser des connaissances expertes sous forme de méta-données. La troisième étape de SAFE-Next consiste à incorporer ces connaissances dans le processus ECD en utilisant les méta-données pour une sélection multi-vues d'attributs ou pour une interprétation multi-vues des résultats.

Nous avons montré que SAFE-Next non seulement identifie et structure les points de vue existant dans une organisation à un moment donné, mais identifie en plus les points de vue manquants et qui peuvent exister ailleurs ou qu'il faut développer en interne ou en externe. Pour les points de vue déjà existants, cette approche peut aider à identifier des manques de représentation en terme d'attribut (Fig. 15) et donc la définition de nouveaux attributs.

Actuellement SAFE-Next, en plus de son utilisation au LAB, est en train d'être appliquée chez Renault pour la définition d'ontologie multi-vue des systèmes de sécurité embarqués dans les véhicules. De plus, comme SAFE-Next offre une vision synthétique et multi-vue des connaissances dans un domaine, son utilisation pour la formation des nouveaux recrutés en accidentologie est en cours de discussion. Une autre utilisation prévue de SAFE-Next est la construction d'ontologie dans le domaine de la maintenance industrielle.

Références

[Ahlberg et Wistrand, 1995] C. Ahlberg et E. Wistrand. IVEE : An information Visualization & Exploration Environement,. *Proceedings on Information Visualization*, p 66-73, 1995.

[Angele *et al.*, 1996] J. Angele, D. Fensel et R. Studer. Domain and task modeling in MIKE, Chapman & Hall, 1996.

[Ashby, 1965] W. R. Ashby. An introduction to cybernetics, ed. Hall, C., London, 1965.

[Aussenac-Gilles *et al.*, 1992] N. Aussenac-Gilles, J.-P. Krivine et J. Sallantin. L'acquisition des connaissances pour les systèmes à base de connaissances. *Editorial de la Revue Intélligence Artificielle*, vol. 6(1-2), p. 7-18, 1992.

[Ben Ahmed *et al.*, 2003a] W. Ben Ahmed, M. Mekhilef, M. Bigand et Y. Page. Intégration des connaissances du domaine pour la fouille de données complexes. *Extraction et Gestion de Connaissances*, Université Blaise Pascal, Clermont Ferrand, France, 2003a.

[Ben Ahmed *et al.*, 2003b] W. Ben Ahmed, M. Mekhilef, M. Bigend et Y. Page. MRC : un Modèle de Représentation des Connaissances en accidentologie. *Ingénierie de Connaissance (IC'03)*, ESIEA, Laval, France, 2003b.

[Bertalanffy, 1969] L. v. Bertalanffy. General system theory: foundations, development, applications, George Braziller, New York, 1969.

[Brachman et Anand, 1996] R. Brachman et T. Anand. The Process of Knowledge Discovery in Databases: A Human-Cen-tered Approach. *In Advances in Knowledge Discovery and Data Mining, 37-58, eds. U. Fayyad, G. Piatet-sky- Shapiro, P. Smyth, and R. Uthurusamy*, Menlo Park, Calif, 1996.

[Brenac, 1997] T. Brenac. *L'analyse séquentielle de l'accident de la route: comment la mettre en pratique dans les diagnostics de sécurité routière, Outil et méthode,* Rapport de recherche n°3, INRETS, 1997.

[Brenac et Fleury, 1999] T. Brenac et D. Fleury. Le concept de scénario type d'accident de la circulation et ses applications. *Recherche Transport Sécurité*, vol. 63, 1999.

[Breuker et Van de Velde, 1994] J. Breuker et W. Van de Velde. CommonKADS Library For Expertise Mdeling, Reusable Problèmle Solving Components, IOS Press, Amesterdam, 1994.

[Dhar et Tuzhilin, 1993] V. Dhar et A. Tuzhilin. Abstract-driven pattern discovery in databases. *IEEE Transactions on Knowledge and Data Engineering*, vol. 6, p. 926-938, 1993.

[Dieng *et al.*, 1996] R. Dieng, A. Giboin, C. Amerge, O. Corby, S. Despres, L. Alpay, S. Labidi et S. Lapalut Building of a Corporate Memory for Traffic Accident Analysis. *Proceedings of the 10th Knowledge Acquisition for Knowledge-Based Systems Workshop (KAW'96)*, Banff, Canada, p 35.31-35.20, 1996.

[Duribreux-Cocquebert et Houriez, 2000] M. Duribreux-Cocquebert et B. Houriez. Application industrille d'une approche mixte de modélisation des connaissances. in *Charlet et al. (Eds.), Ingénierie des connaissances, évolution récentes et nouveaux défis*, p. 25-42, Eyrolles, Paris, 2000.

[Famili *et al.*, 1997] A. Famili, W. M. Shen, R. Weber et E. Simoudis. Data Preprocessing and Intelligent Data Analysis. *Inetelligent Data Analysis*, vol. 1, p. 3-23, 1997.

[Fayyad *et al.*, 1996] U. Fayyad, G. Piatetsky-Shapiro et P. Smyth. The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications Of The ACM*, vol. 39(11), 1996.

[Ferrandez *et al.*, 1996] F. Ferrandez, D. Fleury, Y. Girard, L. Alpay, C. Amerge et O. Corby. *Acquisition et modélisation des connaissances expertes en situation de coopération : application à un système d'aide à l'analyse des accidents de la route,* Rapport final de Recherche, PREDIT, 1996.

[Grinstein *et al.*, 2001] G. Grinstein, M. Trutschl et U. Cvek High-Dimensional Visualizations. *Proceedings of the Visual Data Mining Workshop, KDD'2001*, 2001.

[Halkidi *et al.*, 2002] M. Halkidi, Y. Batistakis et M. Vazirgiannis. Cluster Validity Methods: Part I. *SIGMOD Record*, vol. 31(2), p. 40-45, 2002.

[Han et Fu, 1996] J. Han et Y. Fu. Attribute-oriented induction in data mining. *Advances in Knowledge Discovery*, Cambridge, MA, p 399-421, 1996.

[Han et Kamber, 2000] J. Han et M. Kamber. Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2000.

[Keim et Kriegel, 1996] D. A. Keim et A. P. Kriegel. Visualization techniques for mining large databases: acomparision. *IEEE transaction on knowledge and data Engineering*, vol. 8(6), 1996.

[Kurucz *et al.*, 1977] Kurucz, Morrow, Fogarty, Janicek et Klapper. The Effectiveness of ABS in Real Life Accidents. *14th international technical conference on Enhancing Safety Vehicles*, Munich, Allemagne, 1977.

[Le Moigne, 1974] J. L. Le Moigne. La théorie du système général, P. U. F., Paris, 1974.

[Le Moigne, 1999] J.-L. Le Moigne. La modélisation des systèmes complexes, Dunod, 1999.

[Manco *et al.*, 2004] G. Manco, C. Pizzuti et D. Talia. Eureka!: an interactive and visual knowledge discovery tool. *Journal of Visual Languages & Computing*, vol. 15(1), p. 1-35, 2004.

[Megherbi, 1999] B. Megherbi. *Scénarios types d'accidents de la circulation sur autoroute : élaboration, méthodes de reconnaissance et application pour le diagnostic et la prévention*. Thèse de Doctorat, Ecole nationale des ponts et chaussées, 1999.

[Miller, 1995] J. G. Miller. Living Systems, University Press of Colorado, 1995.

[Morin et Le Moigne, 1999] E. Morin et J. L. Le Moigne. L'intelligence de la complexité, L'Harmattan, Paris, 1999.

[Motta *et al.*, 1990] E. Motta, T. Rajan et M. Eisenstadt. Knowledge Acquisition as a Process of Model Refinement. *Knowledge acquisition*, vol. 2, 1990.

[Oliveira et Levkowitz, 2003] M. C. F. Oliveira et H. Levkowitz. From Visual Data Exploration to Visual Data Mining: A Survey. *IEEE Transactions on Visualization and Computer Graphics*, vol. 9(3), p. 378-394, 2003.

[Palmeri et Blalock, 2000] T. J. Palmeri et C. Blalock. The role of background knowledge in speeded perceptual categorization. *Cognition*, vol. 77, p. 45-57, 2000.

[Perron, 1997] T. Perron. *Méthode d'analyse de sécurité primaire automobile pour la spécification fonctionnelle et l'évaluation prévisionnelle d'efficacité de systèmes d'évitement d'accidents.* Thèse de Doctorat, Laboratoire Génie Industriel, Ecole Centrale Paris, 1997.

[Pyle, 1999] D. Pyle. Data Preparation For Data Mining, Morgan Kaufmann Publishers, 1999.

[Rasmussen, 1986] J. Rasmussen. Information Processing And Human-Machine Interaction, North-Holland, 1986.

[Sachinopoulou, 2001] A. Sachinopoulou. Multidimensional Visualization, Espoo, VTT Electronics, 2001.

[Schreiber *et al.*, 1994] A. T. Schreiber, B. Wielinga, D. H. R., H. Akkermans et W. Van de Velde. CommonKADS : A comprehensive methodology for KBS development. *IEEE Expert*, vol. 9(6), p. 28-37, 1994.

[Van Elslande et Alberton, 1997] P. Van Elslande et L. Alberton. *Scénarios-types de production de l'erreur humaine dans l'accident de la route, problématique et analyse qualitative,* Rapport de recherche N°218, INRETS, 1997.

[Vogel, 1989] C. Vogel. Knowledge Oriented Design (KOD): la mise en oeuvre, Editions Masson, 1989.

[Von Foerster, 1995] H. Von Foerster. The Cybernetics of Cybernetics (2nd edition), Future Systems Inc., Minneapolis, 1995.

[Wielinga *et al.*, 1994] B. J. Wielinga, Y. Hans Akkermans, H. Hassan, O. Olsson, K. Orsv¨arn, G. Schreiber, P. Terpstra, W. Van de Velde et S. Wells. *Expertise Model Definition Document,* ESPRIT Project P5248 KADSII, Document Id. KADSII/M2/UvA/026/5.0, University of Amsterdam, 1994.

[Zhang *et al.*, 2002] L. Zhang, C. Tang, Y. Song, A. Zhang et M. Ramanathan. *VizCluster and Its Application on Clustering Gene Expression Data*, Department of Computer Science and Engineering, State University of New York at Buffalo, 2002.

Summary

Knowledge Discovery in Databases (KDD) is an iterative process whose complexity depends on the nature of processed data, the nature of the extracted knowledge as well as the application domain. To handle this complexity, expert knowledge is required during the different KDD steps (i.e. data preparation, data mining and results interpretation). The aim of this paper is to combine Knowledge Engineering approach and KDD approach to develop a methodology, which allows a multi-view identification of domain knowledge, their formalization and their incorporation in the KDD process. This approach is applied in accidentology in order to extract knowledge from accident databases useful for the development of in-vehicle safety systems.