

STEEL: une extension du Calcul des Événements pour la représentation de récits épidémiologiques

Hervé Chaudet

Laboratoire de Biomathématiques, Informatique Médicale
Faculté de Médecine, Université de la Méditerranée, Marseille.
herve.chaudet@medecine.univ-mrs.fr

Résumé : Cet article présente la façon dont nous avons abordé la représentation et le raisonnement portant sur les événements qui sont décrits dans des dépêches épidémiologiques. Nous proposons ainsi une extension spatiotemporelle du Calcul des Événements avec représentation et construction d'agrégats d'événements respectant leur distribution spatiotemporelle. Une étude pour preuve de concept, visant à comparer les performances d'une implémentation expérimentale de ce langage avec des experts humains au cours d'une tâche de questions-réponses sur un corpus de test, a montré un accord des experts avec les réponses du système.

Mots-clés : Représentation des événements, Agrégats d'événements, Représentation de dépêches épidémiologiques, Propagation d'épidémies.

1 Introduction

L'emploi de dépêches épidémiologiques sur Internet pour la diffusion large, rapide et précoce d'une information épidémiologique montre un succès croissant dans le cadre du suivi de phénomènes épidémiologiques (Woodall, 1997 ; Pavlin, Mostashari et al, 2003). Les possibilités descriptives de ces textes et leur capacité à rapporter des situations inattendues les rendent compétitifs face aux bases de données pour rapporter les épidémies de maladies émergentes et les comportements inhabituels de maladies, dont les menaces biologiques. C'est pourquoi nous développons un prototype de traitement automatique de dépêches épidémiologiques avec un double objectif final de question-réponse et de modélisation qualitative. Toutefois, si l'utilisation de courriels rend plus aisée la notification de phénomènes épidémiologiques, l'utilisation automatique de l'information ainsi transmise est particulièrement difficile, spécialement lorsque le système doit procéder à une tâche de question-réponse. Ceci demande en effet :

- La capacité de construire une représentation du contenu du récit axée sur l'information épidémiologique tout en gardant le lien avec le texte d'origine (Damianos, Day et al., 2002 ; Grishman, Huttunen et al., 2002).
- La capacité de raisonner sur ces représentations afin de reconstituer l'histoire de l'épidémie, qui correspond à l'évolution spatiotemporelle de ses caractéristiques.

L'expérience de BIOSFORM sur les bases de données médicales montre que ce raisonnement, très consommateur de connaissances, requiert une modélisation explicite des processus de reconstruction de l'histoire de l'épidémie, ainsi que l'utilisation d'abstractions temporelles et spatiales (Buckeridge, Graham et al., 2002). De plus, les systèmes d'extraction d'information cités précédemment nous ont appris que l'information dans les dépêches épidémiologiques était centrée sur les événements et que la reconstitution de la structure des scénarios des épidémies était particulièrement difficile du fait de la structure complexe des événements, des relations d'inclusion entre ceux-ci, de leur éparpillement dans l'ensemble des textes, et de l'aspect incertain de l'information.

Cet article présente la façon dont nous avons envisagé la représentation et le raisonnement portant sur les événements qui sont décrits dans les dépêches épidémiologiques envoyées par le système de bulletin électronique ProMED (<http://www.promedmail.org/>). Après avoir présenté les principales particularités de l'information issue de ces dépêches, nous décrivons STEEL (Spatio-Temporal Extended Event Language), l'extension du calcul des événements que nous avons développée pour la représentation de ces connaissances. Dans cette extension, nous avons remplacé le temps par la localisation spatio-temporelle et nous avons introduit la notion d'agrégat d'événements spatio-temporellement localisés comme fondement pour une abstraction spatio-temporelle. La dernière partie décrit succinctement une évaluation de l'adéquation de cette représentation pour notre domaine d'application en comparant les performances du système avec des experts humains dans une situation ad hoc de question-réponse sur un corpus de 35 dépêches épidémiologiques.

2 Le problème de l'exploitation des dépêches épidémiologiques

Les dépêches épidémiologiques sont de courts récits journalistiques qui visent à rapporter une succession d'événements connectés décrivant la progression d'une épidémie (la Figure 1 montre un exemple de dépêche). Le terme d'événement est utilisé ici pour tout ce qui correspond à des « actions, événements, déplacements, accomplissements et processus » (Schokkenbroek, 1999).

La structure de ces récits est complexe, entrelaçant et dispersant l'information en même temps que les événements sur l'ensemble des textes (Schokkenbroek, 1999). Trois raisons peuvent expliquer ceci :

- Une dépêche peut se rapporter à plus d'un événement épidémiologique, lieu ou temps.
- Sa rédaction est hautement influencée par un souci de brièveté. Ainsi l'histoire est compactée en un petit nombre de phrases, compliquant la structure du récit.
- Elle se présente souvent sous la forme d'une mise à jour de l'évolution des caractéristiques de l'épidémie depuis la dernière description.

Ainsi, en un seul récit, l'exemple rapporte 10 événements (8 associés aux malades) qui concernent un total de 6 lieux et 4 dates ou intervalles temporels, dont

la date de publication. Les événements « s'entrecroisent et se réfèrent de façon complexe, constituant des relations d'inclusion » (Huttunen, Yangarber et al., 2002). Ils décrivent fréquemment des agrégats de sous-événements de façon compacte, comme dans la phase « Of them, 45 have died so far ». Une représentation adéquate des événements se doit donc de prendre en compte ces relations, et en particulier la décomposition en sous-événements (Setzer & Gaizauskas, 2000).

Archive Number: 20021008.5493, Published Date: 08-OCT-2002

Bangladesh: Dengue Fever Continues to Spread

A dwindling supply of blood is exacerbating the dengue outbreak in Bangladesh. The total number of dengue-affected patients, according to the official accounts, stood at 4763 as of 16 Sep 2002. Of these, 45 have died so far. The Dengue Control Room sources said 509 persons affected with dengue virus were undergoing treatment at different hospitals across the country.

Meanwhile, it was reported that one person died from dengue hemorrhagic fever (DHF) in Magura district, while 4 others had been hospitalised in Jessore General Hospital in the last 24 hours. The Khulna City Corporation in the meantime has launched an anti-mosquito drive in the city. At least one person died in Chapai-Nawabganj district and 15 others have been hospitalised for DHF. All of them were admitted to the Rajshahi Medical College Hospital (RMCH) during the past 10 days. [...]

Fig. 1 - Exemple de dépêche épidémiologique extrait de ProMED-MAIL

Des interviews préliminaires d'experts en médecine tropicale et médecine des voyages, qui sont les principaux utilisateurs de ces dépêches dans notre hôpital, ont confirmé que l'utilisation de ces informations demandait la reconstruction des relations associant les événements, de leurs localisations temporelles et spatiales.

3 Le langage de représentation des connaissances : STEEL

STEEL est un langage typé en logique du premier ordre, basé sur le Calcul des Événements qui a été introduit par Kowalski et Sergot (Kowalski & Sergot, 1986) pour raisonner sur la survenue d'événements, sur les propriétés que ces événements déclenchent ou achèvent, et sur les intervalles maximaux de validité pendant lesquels ces propriétés existent sans discontinuer. Parmi les nombreuses extensions qui ont été proposées pour accroître son expressivité, la notion d'événements complexes a été explorée par Cervesato et Montanari (Cervesato & Montanari, 2000), et le problème de la localisation spatiale d'événements a été abordé par Galton (Galton, 1993) et par Bennett (Bennett, 2001). Notre représentation des connaissances continue ces travaux en introduisant une localisation conjointe spatio-temporelle et en utilisant les propriétés de cette localisation pour définir les règles de construction d'agrégats d'événements.

3.1 Ontologie de représentation

L'ontologie du langage STEEL est composée de 4 type : événements, fluents, expressions temporelles, et régions spatiales.

Un **événement** est la réalisation ou l'occurrence d'une action. Si les actions sont indépendantes du temps, définissant des activités pouvant être effectuées par des agents pour modifier l'état du monde, les événements sont des réalisations de ces actions et sont dépendants du temps. Les éléments du récit rapportant des événements peuvent être identifiés sur la base de leurs « verbes dynamiques » (Schokkenbroek, 1999) (e.g. « one person died from dengue », « it was reported that one person died »). Les événements peuvent être instantanés ou peuvent se dérouler sur un intervalle de temps, définissant ainsi la notion de durée d'événement.

Les **fluents** sont des expressions valuées qui décrivent des propriétés d'objets, et dont l'interprétation varie avec le temps (e.g. « dwindling supply of blood »). Quand l'objet acquiert cette propriété spécifique, le fluent est dit valide. Les états des fluents sont définis par les événements qui les initient ou les terminent et sinon sont stables.

Les **expressions temporelles** : à strictement parler le temps est un concept difficile à représenter (Combi & Shahar, 1997, Steedman, 1997). Notre objectif est de représenter le temps d'une façon pratique pour induire le moment de survenue et l'ordonnancement des événements. Dans notre ontologie, le temps est un ensemble ordonné $(T, <)$ où les éléments de T sont les time stamps de Shahar (Shahar, 1997) qui sont tirés des expressions temporelles rencontrées dans les récits et pouvant être placés sur un axe du temps (e.g. "16 September 2002").

Les **régions spatiales** : l'espace est bidimensionnel et correspond à l'ensemble $S = \mathcal{R} \times \mathcal{R}$ où $\mathcal{R}$ est l'ensemble des réels. Une région est un sous-ensemble de S , usuellement représenté par un nom (e.g. "Magura district"). Un point est une forme particulière de région (e.g. "Rajshahi Medical College Hospital"). Ce choix conserve un niveau de complexité en accord avec les objectifs du récit et avec la façon dont les relations spatiales sont exprimées en langage naturel. Ainsi que remarqué par Asher et Vieu (Asher & Vieu, 1995), la conception de l'espace topologique est étrangère à la façon dont l'espace est usuellement exprimé dans les récits, où il reste utilisable en l'absence d'un système de coordonnées.

Sur la base des deux derniers types, nous définissons la localisation spatiotemporelle d'un événement comme un couple $\langle t, l \rangle$ où t est une expression temporelle et l une région spatiale.

3.2 Description du langage

Les types de bases sont utilisés pour définir 3 groupes de prédicats de base.

3.2.1 Les événements et leurs influences sur les fluents

Ce groupe de prédicats modifiés est décrit dans le Tableau 1. Dans une première approche, nous avons évité de traiter les trajectoires spatiales lors de la survenue d'événements.

Les relations entre t_1 et t_2 dans le prédicat $\text{happens}(e, \langle t_1, l \rangle, \langle t_2, l \rangle)$ sont formalisées dans l'axiome suivant :

$$\text{happens}(e, \langle t_1, l \rangle, \langle t_2, l \rangle) \rightarrow t_1 \leq t_2 \quad (\text{A1})$$

Tableau 1 – Descriptions et significations des prédicats de base relatifs à la survenue des événements et à leur influence sur la valeur des fluents.

Prédicat	Signification
$t_1 < t_2$	Le time stamp t_1 précède le time stamp t_2 .
$t_1 = t_2$	Le time stamp t_1 est identique au time stamp t_2 .
$\text{happens}(e, \langle t_1, l \rangle, \langle t_2, l \rangle)$	L'événement e commence à la localisation spatiotemporelle $\langle t_1, l \rangle$ et se termine à $\langle t_2, l \rangle$ (note: $\text{happens}(e, \langle t, l \rangle) \equiv_{\text{def}} \text{happens}(e, \langle t, l \rangle, \langle t, l \rangle)$).
$\text{initiallyTrue}(f, l)$	Le fluent f est valide à partir de l'instant 0 à la localisation spatiale l .
$\text{initiallyFalse}(f, l)$	Le fluent f est non valide à partir de l'instant 0 à la localisation spatiale l .
$\text{initiates}(e, f, \langle t, l \rangle)$	Le fluent f est valide après l'occurrence de l'événement e à $\langle t, l \rangle$.
$\text{terminates}(e, f, \langle t, l \rangle)$	Le fluent f cesse d'être valide après l'occurrence de l'événement e à $\langle t, l \rangle$.
$\text{releases}(e, f, \langle t, l \rangle)$	Le fluent f n'est plus sujet à inertie après l'occurrence de l'événement e à $\langle t, l \rangle$. (sa valeur devient donc indéterminée)

3.2.2 Relations spatiotemporelles

Suivant la théorie méréotopologique de l'espace-temps de Hazarika et Cohn (Hazarika & Cohn, 2001), les relations spatio-temporelles entre objets peuvent être représentées par des relations binaires basées sur la notion de connexion.

Tableau 2 – Définition des relations méréologiques spatiotemporelles à partir de la primitive $\alpha\text{-connected}(x, y)$ (où $\alpha \in \{\text{st}, \text{sp}, \text{t}\}$)

Relation	Predicat	Définition
x is disconnected from y	$\alpha\text{-disconnected}(x, y)$	$\neg \alpha\text{-connected}(x, y)$
x is a part of y	$\alpha\text{-partof}(x, y)$	$\forall z. (\alpha\text{-connected}(z, x) \rightarrow \alpha\text{-connected}(z, y))$
x is a proper part of y	$\alpha\text{-properpart}(x, y)$	$\alpha\text{-partof}(x, y) \wedge \neg \alpha\text{-partof}(y, x)$
x is identical with y	$\alpha\text{-equal}(x, y)$	$\alpha\text{-partof}(x, y) \wedge \alpha\text{-partof}(y, x)$
x overlaps y	$\alpha\text{-overlap}(x, y)$	$\forall z. (\alpha\text{-partof}(z, x) \wedge \alpha\text{-partof}(z, y))$
x is discrete from y	$\alpha\text{-discrete}(x, y)$	$\neg \alpha\text{-overlap}(x, y)$
x partially overlaps y	$\alpha\text{-partoverlap}(x, y)$	$\alpha\text{-overlap}(x, y) \wedge \neg \alpha\text{-partof}(x, y) \wedge \neg \alpha\text{-partof}(y, x)$
x is externally connected to y	$\alpha\text{-externconnected}(x, y)$	$\alpha\text{-connected}(x, y) \wedge \neg \alpha\text{-overlap}(x, y)$
x is a tangential proper part of y	$\alpha\text{-tangproppart}(x, y)$	$\alpha\text{-properpart}(x, y) \wedge \exists z. (\alpha\text{-externconnected}(z, x) \wedge \alpha\text{-externconnected}(z, y))$
x is a non tangential proper part of y	$\alpha\text{-nontangproppart}(x, y)$	$\alpha\text{-properpart}(x, y) \wedge \neg \exists z. (\alpha\text{-externconnected}(z, x) \wedge \alpha\text{-externconnected}(z, y))$

Deux entités sont spatialement connectées (*sp-connected*) si elle partagent au moins un point spatial, éventuellement de façon non simultanée (p.e. le Zaïre et le Congo Démocratique). La connexion temporelle (*t-connected*) de deux intervalles de temps est définie de façon analogue. Finalement deux entités sont spatiotemporellement connectées (*st-connected*) si les fermetures de ces entités partagent au moins un point spatiotemporel. Cette primitive $\alpha\text{-connected}(x, y)$, où $\alpha \in \{\text{st}, \text{sp}, \text{t}\}$, permet de définir un ensemble de 10 autres relations qui constitue la base d'un langage qualitatif de représentation (Tableau 2). La Figure 2 présente graphiquement l'ensemble de ces relations, leur sémantique pour les cas spécifiques spatiaux (sp-) et temporels (t-), en rappelant leur dénomination par Allen (Allen, 1984) dans ce dernier cas, et montre leur treillis d'héritage. Les relations réciproques

ne sont pas représentées. Les six relations terminales du treillis sont conjointement exhaustives et disjointes deux à deux. L'adéquation cognitive de ces relations pour la représentation et le raisonnement spatiaux a été confirmée expérimentalement (Knauff, 1999).

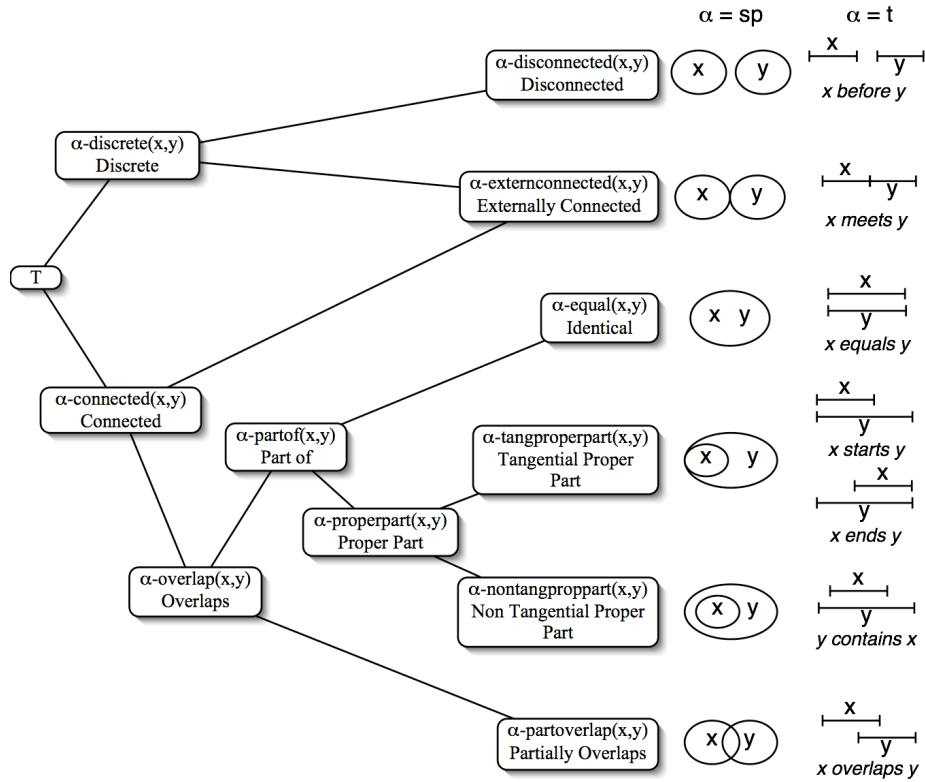


Fig. 2 – Treillis d'héritage des relations méréotopologiques et représentation graphique de leur sémantique pour l'espace et le temps.

La connexion spatiotemporelle de deux entités peut ainsi être facilement définie à partir des relations du treillis de la Figure 2 par :

$$\text{st-connected}(x,y) \rightarrow \text{t-connected}(x,y) \wedge \text{sp-connected}(x,y)$$

Ces relations permettent d'introduire un axiome de persistance spatiale établissant qu'un événement survient dans une région s'il survient dans une de ses parties (p.e. s'il existe un cas d'Ebola au Gabon, alors il existe un cas d'Ebola en Afrique) :

$$\text{happens}(e, \langle t_1, l_2 \rangle, \langle t_2, l_2 \rangle) \leftarrow \text{happens}(e, \langle t_1, l_1 \rangle, \langle t_2, l_1 \rangle) \wedge \text{sp-partof}(l_1, l_2) \text{ (A2)}$$

3.2.3 Macro-événements

Pouvoir représenter de façon compacte des assemblages complexes d'événements demande l'introduction de relations additionnelles entre événements, comme la

séquentialité, la simultanéité, l'itération, ou le délai entre événements. Cervesato et Montanari ont introduit la notion de macro-événements, qui sont des expressions définies par la grammaire suivante (Cervesato & Montanari, 2000) :

$$m ::= e \text{ (événement élémentaire)} \quad | \quad m_1 ;_d^D m_2 \text{ (séquence avec délai de } d \text{ à } D) \\ | \quad m_1 + m_2 \text{ (alternative)} \quad | \quad m_1 \parallel m_2 \text{ (parallélisme)} \\ | \quad m^n \text{ (itération } n \text{ fois)}$$

En gardant la même grammaire, nous considérons qu'un macro-événement est l'occurrence d'un agrégat structuré d'événements, et est un sous-concept direct d'*événement*. Chaque instance de macro-événement est définie par l'instance d'une structure de macro-événement (MES), qui est une formule obtenue en appliquant récursivement la grammaire.

Le MES résultant de cette grammaire est un arbre dont les feuilles sont des sous-concepts d'*événement*. Il peut être utilisé pour définir des sous-classes d'événements, et si m est un macro-événement, MES_m sa structure, alors $MEClass_m \doteq \text{MacroEvent} \sqcap MES_m$

Tableau 3 – Définition des macro-événements en logique du premier ordre.

Macro	Structure	Définition
$m_1 ;_d^D m_2$	$\text{sequevent}(m_1, m_2, d, D)$	$\text{happens}(m, \langle t_1, l \rangle, \langle t_2, l \rangle) \wedge \text{meventdef}(m, \text{sequevent}(m_1, m_2, d, D)) \Leftrightarrow \exists t_3, t_4 . (\text{happens}(m_1, \langle t_1, l \rangle, \langle t_3, l \rangle) \wedge \text{happens}(m_2, \langle t_4, l \rangle, \langle t_2, l \rangle) \wedge t_3 + d \leq t_4 \leq t_3 + D)$
$m_1 + m_2$	$\text{altevent}(m_1, m_2)$	$\text{happens}(m, \langle t_1, l \rangle, \langle t_2, l \rangle) \wedge \text{meventdef}(m, \text{altevent}(m_1, m_2)) \Leftrightarrow \text{happens}(m_1, \langle t_1, l \rangle, \langle t_2, l \rangle) \vee \text{happens}(m_2, \langle t_1, l \rangle, \langle t_2, l \rangle)$
$m_1 \parallel m_2$	$\text{parevent}(m_1, m_2)$	$\text{happens}(m, \langle t_1, l \rangle, \langle t_2, l \rangle) \wedge \text{meventdef}(m, \text{parevent}(m_1, m_2)) \Leftrightarrow \exists t_3, t_4, t_5, t_6 . (\text{happens}(m_1, \langle t_3, l \rangle, \langle t_4, l \rangle) \wedge \text{happens}(m_2, \langle t_5, l \rangle, \langle t_6, l \rangle) \wedge t_1 = \min(t_3, t_5) \wedge t_2 = \max(t_4, t_6))$
m^n	$\text{iterevent}(m, n)$	$\text{happens}(m, \langle t_1, l \rangle, \langle t_2, l \rangle) \wedge \text{meventdef}(m, \text{iterevent}(E, n)) \Leftrightarrow \exists t_3, t_4 . (\text{happens}(m_1, \langle t_1, l \rangle, \langle t_3, l \rangle) \wedge \text{happens}(m_2, \langle t_4, l \rangle, \langle t_2, l \rangle) \wedge \text{meventdef}(m_1, \text{iterevent}(E, n-1)) \wedge E(m_2) \wedge t_3 \leq t_4)$

Le Tableau 3 montre comment ces macro-événements ont été exprimés en logique du premier ordre. Ces règles montrent le cadre logique permettant de construire la représentation d'événements complexes avec des limites spatio-temporelles cohérentes. La relation entre un macro-événement m et sa MES est entretenue par un prédicat $\text{meventdef}(m, MES_m)$.

Ces macro-événements peuvent se substituer pleinement aux événements dans toutes les propriétés de ces derniers (en particulier le Tableau 1).

3.3 Boîte à outil de raisonnement

Reconstruire l'histoire de la propagation de l'épidémie à partir des connexions spatiotemporelles d'événements et de fluents décrites par les observations partielles rapportées dans les dépêches demande de capturer la notion de continuité spatiotemporelle pour déterminer les intervalles maximaux de validité (MVI) des fluents et les structures des macro-événements décrits sur plusieurs textes (évolution du nombre de décès, de l'incidence, ...).

3.3.1 Persistance des fluents

L'influence des événements sur les valeurs des fluents ainsi que la persistance de ces dernières dans le temps et l'espace est exprimée par un ensemble de 6 axiomes décrivant la sémantique de 3 prédicats de base (Tableau 4) :

$$\text{clipped}(\langle t_1, l_1 \rangle, f, \langle t_4, l_1 \rangle) \Leftrightarrow \exists e, t_2, t_3, t_5, l_2 . \text{happens}(e, \langle t_2, l_2 \rangle, \langle t_3, l_2 \rangle) \wedge (\text{terminates}(e, f, \langle t_5, l_1 \rangle) \vee \text{releases}(e, f, \langle t_5, l_1 \rangle)) \wedge t_2 < t_5 < t_3 \wedge t_1 < t_5 < t_4 \wedge \text{sp-partof}(l_1, l_2) \quad (\text{A3})$$

$$\text{declipped}(\langle t_1, l_1 \rangle, f, \langle t_4, l_1 \rangle) \Leftrightarrow \exists e, t_2, t_3, t_5, l_2 . \text{happens}(e, \langle t_2, l_2 \rangle, \langle t_3, l_2 \rangle) \wedge (\text{initiates}(e, f, \langle t_5, l_1 \rangle) \vee \text{releases}(e, f, \langle t_5, l_1 \rangle)) \wedge t_2 < t_5 < t_3 \wedge t_1 < t_5 < t_4 \wedge \text{sp-partof}(l_1, l_2) \quad (\text{A4})$$

$$\text{holdsAt}(f, \langle t, l \rangle) \leftarrow \text{initiallyTrue}(f, l) \wedge \neg \text{clipped}(\langle 0, l \rangle, f, \langle t, l \rangle) \quad (\text{A5})$$

$$\text{holdsAt}(f, \langle t, l \rangle) \leftarrow \exists e, t_1, t_2, t_3, l_2 . \text{happens}(e, \langle t_1, l_2 \rangle, \langle t_2, l_2 \rangle) \wedge \text{initiates}(e, f, \langle t_3, l_1 \rangle) \wedge t_1 < t_3 < t_2 \wedge t_3 < t \wedge \neg \text{clipped}(\langle t_3, l_1 \rangle, f, \langle t, l_1 \rangle) \wedge \text{sp-partof}(l_1, l_2) \quad (\text{A6})$$

$$\neg \text{holdsAt}(f, \langle t, l \rangle) \leftarrow \text{initiallyFalse}(f, l) \wedge \neg \text{declipped}(\langle 0, l \rangle, f, \langle t, l \rangle) \quad (\text{A7})$$

$$\neg \text{holdsAt}(f, \langle t, l \rangle) \leftarrow \exists e, t_1, t_2, t_3, l_2 . \text{happens}(e, \langle t_1, l_2 \rangle, \langle t_2, l_2 \rangle) \wedge \text{terminates}(e, f, \langle t_3, l_1 \rangle) \wedge t_1 < t_3 < t_2 \wedge t_3 < t \wedge \neg \text{declipped}(\langle t_3, l_1 \rangle, f, \langle t, l_1 \rangle) \wedge \text{sp-partof}(l_1, l_2) \quad (\text{A8})$$

Tableau 4 – Prédicats de base décrivant la persistance des fluents.

Predicat	Signification
$\text{clipped}(\langle t_1, l \rangle, f, \langle t_2, l \rangle)$	Le fluent f est mis à faux entre t_1 et t_2 dans la région l .
$\text{declipped}(\langle t_1, l \rangle, f, \langle t_2, l \rangle)$	Le fluent f est mis à vrai entre t_1 et t_2 dans la région l .
$\text{holdsAt}(f, \langle t, l \rangle)$	Le fluent f est valide à $\langle t, l \rangle$ (la propriété est vraie).

3.3.2 Agrégation d'événements et continuité spatiotemporelle

Pour permettre la construction d'agrégats d'événements spatio-temporellement localisés, nous avons introduit un constructeur, noté $\bigwedge_h$, qui assemble les événements de deux prédicats *happens* pour construire le macro-événement résultant sur la base de leurs relations spatiotemporelles et ontologiques.

Soit e_1 et e_2 deux instances respectives de $E_1 \sqsubseteq \text{Événement}$ et $E_2 \sqsubseteq \text{Événement}$, $\text{happens}(e_1, \langle t_1, l_1 \rangle, \langle t_1', l_1 \rangle)$ et $\text{happens}(e_2, \langle t_2, l_2 \rangle, \langle t_2', l_2 \rangle)$ leurs représentations dans la base de connaissance. L'intervalle durant lequel chaque événement survient est respectivement défini par $[t_1, t_1'] = d_1$ et $[t_2, t_2'] = d_2$.

Si les événements sont de même nature, c'est à dire $E_1 = E_2$ avec des instances disjointes ($e_1 \neq e_2$ or $\text{MEClass}_{e_1} \cap \text{MEClass}_{e_2} = \emptyset$) ou $E_2 = (\text{Macroevent} \sqcap \text{iterevent}(E_1, *))$, alors le constructeur en génère une itération. Dans tous les autres cas, la structure du macro-événement résultant fait appel à la séquentialité ou le parallélisme suivant les relations spatiotemporelles des événements de départ (Tableau 5).

Tableau 5 – Résultats de $\wedge_h(\text{happens}(e_1, \langle t_1, l_1 \rangle, \langle t_1', l_1' \rangle), \text{happens}(e_2, \langle t_2, l_2 \rangle, \langle t_2', l_2' \rangle))$.

Si: $E_1=E_2=E$ or $E_2=\text{Macroevent} \sqcap \text{iterevent}(E_1=E, *)$		
<i>sp relations</i>	<i>Resultats</i>	
sp-partof(l', l) with ($l'=l_1 \wedge l=l_2$) or ($l'=l_2 \wedge l=l_1$)	$\text{happens}(m, \langle t, l \rangle, \langle t', l' \rangle) \wedge \text{meventdef}(m, \text{iterevent}(E, *)) \wedge t = \min(t_1, t_2) \wedge t' = \max(t_1', t_2')$	
sp-partoverlap(l_1, l_2) $\vee$ sp-discrete(l_1, l_2)	$\text{happens}(m, \langle t, l \rangle, \langle t', l' \rangle) \wedge \text{meventdef}(m, \text{iterevent}(E, *)) \wedge t = \min(t_1, t_2) \wedge t' = \max(t_1', t_2') \wedge l = (l_1 \cup l_2)$	
Sinon		
sp relations : cas où sp-equal(l_1, l_2)		
<i>t relations</i>	<i>Resultats</i>	
t-discrete(d_1, d_2)	$\text{happens}(m, \langle t_1, l \rangle, \langle t_2', l' \rangle) \wedge \text{meventdef}(m, \text{sequevent}(e_1, e_2, d, d)) \wedge d = t_2 - t_1' \wedge l_1 = l_2 = l$	
equals(d_1, d_2)	$\text{happens}(m, \langle t, l \rangle, \langle t', l' \rangle) \wedge \text{meventdef}(m, \text{parevent}(e_1, e_2)) \wedge t_1 = t_2 = t \wedge t_1' = t_2' = t' \wedge l_1 = l_2 = l$	
t-properpart(d_1, d_2)	$\text{happens}(m, \langle t_2, l \rangle, \langle t_2', l' \rangle) \wedge \text{meventdef}(m, \text{parevent}(e_1, e_2)) \wedge l_1 = l_2 = l$	
overlaps(d_1, d_2)	$\text{happens}(m, \langle t, l \rangle, \langle t', l' \rangle) \wedge \text{meventdef}(m, \text{parevent}(e_1, e_2)) \wedge t = \min(t_1, t_2) \wedge t' = \max(t_1', t_2') \wedge l_1 = l_2 = l$	
sp relations : cas où sp-discrete(l_1, l_2) $\vee$ sp-partoverlap(l_1, l_2)		
<i>t relations</i>	<i>Resultats</i>	
t-discrete(d_1, d_2)	$\text{happens}(m, \langle t_1, l \rangle, \langle t_2', l' \rangle) \wedge \text{meventdef}(m, \text{sequevent}(e_1, e_2, d, d)) \wedge d = t_2 - t_1' \wedge l = (l_1 \cup l_2)$	
equals(d_1, d_2)	$\text{happens}(m, \langle t, l \rangle, \langle t', l' \rangle) \wedge \text{meventdef}(m, \text{parevent}(e_1, e_2)) \wedge t_1 = t_2 = t \wedge t_1' = t_2' = t' \wedge l = (l_1 \cup l_2)$	
t-properpart(d_1, d_2)	$\text{happens}(m, \langle t_2, l \rangle, \langle t_2', l' \rangle) \wedge \text{meventdef}(m, \text{parevent}(e_1, e_2)) \wedge l = (l_1 \cup l_2)$	
overlaps(d_1, d_2)	$\text{happens}(m, \langle t, l \rangle, \langle t', l' \rangle) \wedge \text{meventdef}(m, \text{parevent}(e_1, e_2)) \wedge t = \min(t_1, t_2) \wedge t' = \max(t_1', t_2') \wedge l = (l_1 \cup l_2)$	
sp relations : cas où sp-properpart(l_1, l_2)		
<i>t relations</i>	<i>Resultats</i>	
t-discrete(d_1, d_2)	$\text{happens}(m, \langle t_1, l \rangle, \langle t_2', l' \rangle) \wedge \text{meventdef}(m, \text{sequevent}(e_1, e_2, d, d)) \wedge d = t_2 - t_1'$	
equals(d_1, d_2)	$\text{happens}(m, \langle t, l \rangle, \langle t', l' \rangle) \wedge \text{meventdef}(m, \text{parevent}(e_1, e_2)) \wedge t_1 = t_2 = t \wedge t_1' = t_2' = t'$	
t-properpart(d_1, d_2)	$\text{happens}(m, \langle t_2, l \rangle, \langle t_2', l' \rangle) \wedge \text{meventdef}(m, \text{parevent}(e_1, e_2))$	
overlaps(d_1, d_2)	$\text{happens}(m, \langle t, l \rangle, \langle t', l' \rangle) \wedge \text{meventdef}(m, \text{parevent}(e_1, e_2)) \wedge t = \min(t_1, t_2) \wedge t' = \max(t_1', t_2')$	

3.3.3 Occurrence de macro-événement et Intervalle maximal de validité de propriétés

Un macro-événement m , de structure MES_m , est survenu durant un intervalle $[\langle t, l \rangle, \langle t', l' \rangle]$, ce qui s'écrit $\text{meo}(m, t, t', l)$, ssi $\exists t_1, t_2, l_1. \text{meventdef}(m, \text{MES}_m) \wedge \text{happens}(m, \langle t_1, l_1 \rangle, \langle t_2, l_1 \rangle) \wedge t \leq t_1 \leq t_2 \leq t' \wedge \text{sp-partof}(l_1, l)$. L'occurrence du macro-événement peut ne pas être explicitement décrite, et peut être déterminée en appliquant récursivement les définitions des résultats de $\bigwedge_h$.

Le MVI de p , écrit $\text{mvi}(p, t, t', l)$, est l'intervalle spatiotemporel maximal $[\langle t, l \rangle, \langle t', l' \rangle]$ durant lequel p est vrai sans interruption, ce qui se formule :

$$\begin{aligned} \text{mvi}(p, t, t', l) \leftrightarrow & t < t' \wedge (\text{initiallyTrue}(p, l) \vee (\text{happens}(e_1, \langle t_1, l_1 \rangle, \langle t_2, l_1 \rangle) \wedge \\ & \text{initiates}(e_1, p, \langle t, l \rangle) \wedge t_1 < t < t_2 \wedge \text{sp-partof}(l_1, l)) \wedge \\ & \text{happens}(e_2, \langle t_3, l_2 \rangle, \langle t_4, l_2 \rangle) \wedge (\text{terminates}(e_2, p, \langle t', l' \rangle) \vee \\ & \text{releases}(e_2, p, \langle t', l' \rangle)) \wedge t_3 < t' < t_4 \wedge \text{sp-partof}(l, l_2) \wedge \\ & \neg \text{clipped}(\langle t, l \rangle, p, \langle t', l' \rangle)) \end{aligned}$$

STEEL est capable de déterminer ou de tester un MVI ou la survenue d'un macro-événement et de procéder à des combinaisons booléennes de ces vérifications.

4 Évaluation

Pour preuve de concept, nous avons étudié l'adéquation de STEEL pour représenter le contenu de dépêches épidémiologiques en comparant les performances d'une implémentation expérimentale de ce langage en Prolog avec celles d'experts humains dans une situation d'interrogation.

STEEL a été implémenté en SWI-Prolog (University of Amsterdam, <http://www.swi-prolog.org/>). Les axiomes et les définitions de la section précédente ont été transcrits sous forme de règles prolog. L'ensemble représente un module d'environ 200 règles. La base de connaissance de test a été construite à partir d'un corpus de 35 dépêches ProMED qui décrivaient une épidémie de fièvre Ebola au Gabon entre décembre 2001 et mai 2002. La taille de ce corpus était de 8105 mots en 213 phrases. De ce corpus, 224 événements, dont 148 macro-événements, ont été extraits manuellement à l'aide d'un logiciel d'annotation construit ad hoc sur la base des conclusions de (Setzer & Gaizauskas, 2000), qui sollicitait l'annotateur pour inférer les localisations spatiotemporelles chaque fois que celles-ci n'étaient pas directement spécifiées dans le texte. Les réponses de l'annotateur étaient alors soit une localisation, soit une expression relationnelle de localisation (p.e. inclusion, précédence,...). La Figure 3 montre un extrait de la base de connaissance, qui est d'environ 2000 règles.

```
ist(happens(reports(proMED,system,proMedMail20010001),[[2001,12,05],_]),system).
...
ist(happens(reports(_WHO,[proMedMail20010001|1]),[[2001,12,04],_]),proMedMail20010001).
ist(happens(event1,[time1,'Ogooué Ivindo Province'],[time2,'Ogooué Ivindo Province']),
[proMedMail20010001|1]).
ist(agent(event1,isPossibly('viral haemorrhagic fever',_)),[proMedMail20010001|1]).
ist(sp-partof(northEastPart('Gabon'),'Ogooué Ivindo Province'),[proMedMail20010001|1]).
...
instance(event1,macroevent).
...
meventdef(event1,itervent(death,7)).
```

Fig. 3 – Extrait de la base de connaissance de test.

Pour l'expérimentation, nous avons construit à partir du corpus un ensemble de 18 questions faisant appel à la boîte à outil de raisonnement. Elles ont été adressées, sous leur forme logique, au système qui a trouvé une réponse dans tous les cas. Le temps de CPU requis pour une réponse a été de 273 ± 98 ms sur un PowerPC G4 sous Darwin/MacOSX à 1.2 Ghz avec 512 Ko de L2 cache et un bus à 167 Mhz.

Nous avons donné dans un livret le même corpus à 3 experts en médecine tropicale qui sont par ailleurs des utilisateurs habituels de ProMED, et leur avons demandé de répondre aux mêmes questions formulées en langage naturel. Dans une

deuxième phase, nous leur avons communiqué les réponses du système sans indiquer leur origine, et leur avons demandé si celles-ci étaient acceptables ou non.

Chaque expert a pris entre 75 et 90 minutes pour terminer l'épreuve. Le nombre de réponses spontanées différentes entre les experts et entre les experts et le système a été constamment de 3, sauf un accord total entre le système et un expert. Un test d'homogénéité montre que les experts ne font aucune distinction entre le système et les autres experts (test de Fisher, $p=0,676$). En fusionnant toutes les réponses des experts, le nombre maximal de discordances avec le système est de 4, ce qui est en faveur d'un agrément experts-système sur les réponses spontanées (test binomial sur les accords, $P_{H0}(\text{nb d'accords} \geq 14)=0,015442$). Cette conclusion est confirmée par le fait que toutes les réponses du système ont été considérées comme exactes par les 3 experts lors du second questionnaire.

5 Conclusion et perspectives

L'adéquation du Calcul des Événements à la représentation des récits, et les caractéristiques structurelles des dépêches épidémiologiques, notamment leur centrage sur la description d'événements, nous a fait choisir ce formalisme plutôt que d'autres, en particulier ceux qui sont basés sur les actions (Kakas, Miller et al., 2001). Dans sa version native ou ses extensions, ce calcul ne pouvait cependant pas représenter directement la double localisation spatio-temporelle des événements, caractéristique importante des dépêches, ni l'agrégation des événements en tenant compte de la localisation de leur survenue. L'extension proposée ici permet d'obtenir une représentation très proche du récit, centrée sur les événements et conservant l'organisation de ces derniers en réseaux. Elle permet aussi de regrouper le contenu de plusieurs dépêches en une description synthétique d'une épidémie. De plus elle permet de considérer une épidémie comme un événement complexe (mécanisme d'abstraction) correspondant à l'agrégation des événements qui la compose. Le problème est ici plus de faire la synthèse des informations que de reconnaître des scénarios comme dans le cas des chroniques (Sabouret & Sansonnet, 2000).

Toutefois, si cette représentation paraît adéquate pour les objectifs affichés, il reste à vérifier la mise en cohésion, l'adéquation et la cohérence de ce langage avec un module d'analyse du langage naturel qui devra traiter les documents d'origine.

Références

- ALLEN J. F. (1984). Towards a General Theory of Action and Time. *Artificial Intelligence*, 23, 123-154.
- ASHER N. & VIEU L. (1995). Toward a geometry of Common Sense: A Semantic and Complete Axiomatization of Mereotopology. In: *Proceedings of IJCAI'95*, Montréal, Québec, Canada; 20-25 August 1995. p. 846-852.
- BENNETT B. (2001). Space, Time, Matter and Things. In: *Proceedings of the international conference on Formal Ontology in Information Systems (FOIS'2001)*, Ogunquit, Maine, USA. New York, NY, USA: ACM Press; 2001. p. 105-116.

- BUCKERIDGE DL, GRAHAM J, O'CONNOR MJ, CHOY MK, TU SW, MUSEN MA. (2002). Knowledge-based bioterrorism surveillance. *Proc AMIA Symp. 2002*, 76-80.
- CERVESATO I, MONTANARI A. (2000). A Calculus of Macro-Events: Progress Report. In A Trudel, SD Goodwin (eds), *7th International Workshop on Temporal Representation and Reasoning (TIME'00)*, Cape Breton, Nova Scotia, Canada; 7-9 July 2000. IEEE Computer Society Press; 2000. p. 47-58.
- COMBI C, SHAHAR Y. (1997). Temporal reasoning and temporal data maintenance in medicine: issues and challenges. *Computers in Biology and Medicine*, 27(5), 353-368.
- DAMIANOS L, DAY D, HIRSCHMAN L ET AL. (2002). Real users, real data, real problems: the MiTAP system for monitoring bio events. In: *Proceedings of BTR2002: Unified Science & Technology for Reducing Biological Threats & Countering Terrorism*, The University of New Mexico, Albuquerque, New Mexico; March 2002.
- GALTON A. (1993). Toward an integrated logic of space, time, and motion. In: *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI'93)*, Chambéry, France; 28 August-3 September 1993. p. 1550-1555.
- GRISHMAN R, HUTTUNEN S, YANGARBER R. (2002). Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, 35, 236-246.
- HAZARIKA SM, COHN AG. (2001). Qualitative Spatio-Temporal Continuity. In: D Montello, editors. *Proceedings of COSIT'01*, Morro Bay, California, USA; September 2001. Volume 2205 of Lecture Notes in Computer Sciences. Berlin: Springer-Verlag; 2001. p. 92-107.
- HUTTUNEN S, YANGARBER R, GRISHMAN R. (2002). Diversity of Scenarios in Information Extraction. In: *Proceedings of 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, Spain; May 2002.
- KAKAS AC, MILLER R, TONI F (2001). $\mathcal{E}$ -RES : Reasoning about Actions, Events and Observations. In : T. Eiter, W. Faber, M. Truszczynski (Eds), 6th Int. Conf. LPNMR 2001, Vienna, Austria, September 17-19, 2001, p. 254-266.
- KNAUFF M. (1999). The cognitive adequacy of Allen's interval calculus for qualitative spatial representation and reasoning. *Spatial Cognition and Computation*, 1, 261-290.
- KOWALSKI R, SERGOT M. (1986). A logic-based calculus of events. *New Generation Computing*, 4, 67-95.
- PAVLIN JA, MOSTASHARI F ET AL (2003). Innovative surveillance methods for rapid detection of disease outbreaks and bioterrorism: results of an interagency workshop on health indicator surveillance. *Am J Public Health*. 2003 Aug;93(8):1230-5.
- SABOURET N, SANSONNET JP (2000). Un modèle de raisonnement sur les systèmes évolutifs fondé sur l'intentionnalisation de chroniques. *Proc. RFIA 2000*, 2 :305-314.
- SCHOKKENBROEK C. (1999). News Stories - Structure, time and evaluation. *Time and Society*, 8(1), 59-98.
- SETZER A, GAIZAUSKAS R. (2000). Annotating Events and Temporal Information in Newswire Texts. In: *Proceedings of the Second International Conference On Language Resources And Evaluation (LREC-2000)*, Athens, Greece; 31 May- 2 June 2000. p. 1287-1293.
- SHAHAR Y. (1997). A framework for knowledge-based temporal abstraction. *Artificial Intelligence*, 90, 79-133.
- STEEDMAN M. (1997). Temporality. In: J Van Benthem and A ter Meulen, editors, *Handbook of Logic and Language*. Amsterdam: Elsevier. p. 895-938.
- WOODALL J. (1997). Official versus unofficial outbreak reporting through the Internet. *International Journal of Medical Informatics*, 47, 31-34.